∴

Technical Spine Omnibus

Second edition, April 2026

John Komkov

April 2026

Contents

Editorial Note

Title	Res Agentica: Technical Spine Omnibus
Edition	Second edition, April 2026 (post-Composition Doctrine)
Doctrinal spine	Composition Doctrine (machine-checked in Lean 4, Aristotle)
Bulla Formal Trilogy	Machine-checked in Lean 4; 703-composition real-schema corpus
Benchmark version	BABEL v0.1 — 932 instances, 7 families, 3 tracks
Real-protocol tracks	Bronze+ (calendar, 1 official server) · Silver (invoice, 2 non-house servers) · Oracle CoT evaluation (5 models)
Interpretability frontier	240+ compositions, 6 baseline families, 3 domains, 4 scales, 2 model architectures
Paper status date	2026-04-26
Canonical citation	Res Agentica: Technical Spine Omnibus, April 2026 ed.

This volume presents the technical spine of the Res Agentica program as one object.

The sixteen papers collected here develop a single argument across nine registers: that compositional semantic obstruction is the unique invariant forced by axioms any reasonable theory must adopt (a doctrinal claim with machine-checked numerical-uniqueness theorem); that bilateral validity does not guarantee global coherence (a mathematical fact); that the failure appears in concrete LLM-tool compositions and current frontier models cannot diagnose it (an empirical discovery); that the failure can be made accountable through manifests and fraud proofs (a protocol); that the obstruction has a precise structural identity as a matroid corank with closed-form pairwise formula and an exact spectral geometry as a Kron-reduced Laplacian (the Bulla Formal Trilogy); that the witness rank is field-independent and admits a sharp repair calculus (signed-incidence + witness-geometry-beyond-fee); that the problem grows worse, not better, at scale, and that no bounded-radius local audit can certify global coherence (a regime change with impossibility theorem); that all of this can be measured by a public benchmark anyone can run, and that frontier LLMs revert to lexical heuristics on hidden-convention tasks (an instrument and a representational discovery); and that even the most powerful component-level diagnostic technology — mechanistic interpretability — inherits the same boundary, because the relevant signal is absent from the edge-local feature class (a representational limit).

The escalation — from doctrine to impossibility to detectability to accountability to structural identity to spectral geometry to repair calculus to scaling to benchmark to representational boundary — is the intellectual spine. The constitutional question that motivates Res Agentica is what happens when this same failure mode operates between autonomous agent runtimes coordinating value across organizational boundaries, where no single party can see the full composition graph. The papers here

demonstrate the mechanism. The implication is that convention mismatch does not go away when agents become autonomous; it becomes constitutional.

The materials do not all carry that burden in the same way. The Composition Doctrine is the canonical statement of the program: it identifies witness rank as the unique numerical invariant respecting seam-independent disclosure as the primitive operation of compositional repair. SCPI is the formal hinge in the topological direction. Bridge is the empirical witness. Seam is the protocol consequence. The Bulla Formal Trilogy (Papers I, II, III) plus Signed-Incidence and Witness Geometry Beyond Fee develop the structural identity, spectral geometry, and repair calculus on real-schema MCP corpora. Local Validity Does Not Compose proves the impossibility theorem for bounded-radius local audits. SHEAF is the frontier extension. BABEL is the benchmark operationalization. Compositional Hiddenness tests whether frontier LLMs can recover the non-local predicate that the structural diagnostic computes exactly. The Interpretability Frontier paper is the capstone: it tests whether mechanistic interpretability tools can substitute for structural diagnosis, and shows they cannot, across two model architectures.

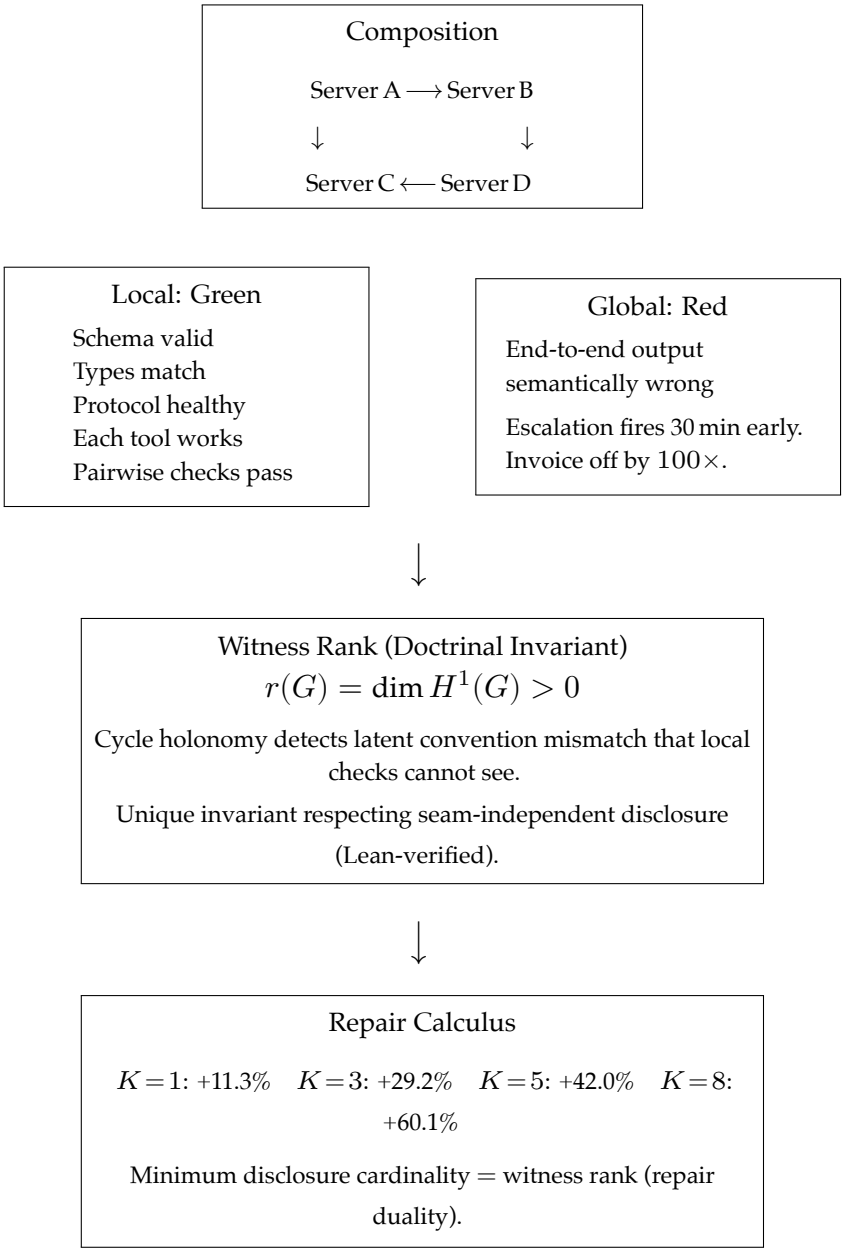
The papers are included as papers, not rewritten into a synthetic house style. The volume places the technical objects in one ordered frame so the reader can see what is doctrinal, what is settled, what is empirical, what is operational, what is benchmarked, and what remains open.

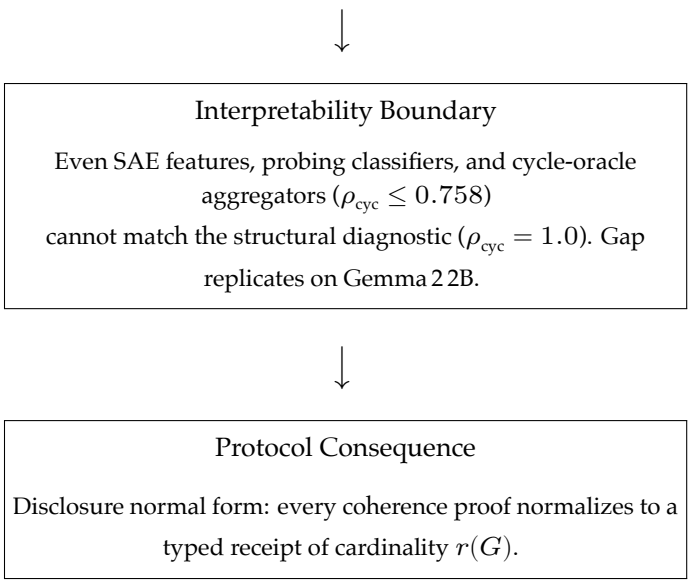
Canonical Names

Name	Role	Part
Composition Doctrine	Doctrinal spine: the unique numerical invariant respecting seam-independent disclosure	0
SCPI	Formal hinge: predicate invention under sheaf constraints; the impossibility theorem in topological form	I
Bridge	Empirical witness: bilateral validity coexists with compositional failure on real systems	II
Seam	Protocol consequence: manifests, semantic witnesses, fraud proofs	III
Coherence Fee (Paper I)	Formal definition: fee as rank deficit of the coboundary	IV-A

Name	Role	Part
Column-Matroid Backbone (Paper II)	Structural identity: fee as matroid corank with closed-form pairwise formula	IV-B
Witness Gram (Paper III)	Spectral geometry: Kron-reduced Laplacian; effective resistance as disclosure substitutability	IV-C
Signed-Incidence Structure	Field-independence: TU rows; partition-style typed repair structurally fails	IV-D
Witness Geometry Beyond Fee	Repair calculus: basis count $\beta(G) = \prod\|C_{d,j}\|$; sharp bounds; realizability	IV-E
SHEAF	Frontier extension: heterogeneous agents, enriched cohomology, scaling theory	V
Communication Bottleneck	Spectral lower bound: Haar-universality and triangle holonomy	V
Coherence Cliff	Scaling evidence: regime change as composition grows	V
Local Validity Does Not Compose	Impossibility theorem: no bounded-radius local audit certifies coherence	V
BABEL	Public benchmark (v0.1): 932 instances, 7 families, 3 tracks, oracle CoT	VI
Compositional Hiddenness	Representation-conditioned predicate access: 88% structured vs 46% flat accuracy	VI
Interpretability Frontier	Representational boundary: edge-local interpretability is not enough	VII

The Core Pattern





Local validity does not guarantee global coherence. The witness rank is the unique invariant respecting seam-independent disclosure as the primitive operation of compositional repair (Composition Doctrine, Theorem 5.1, Lean-verified). A structural diagnostic detects what local checks miss; even the most powerful component-level tools — mechanistic interpretability — inherit the same boundary, across model architectures. Targeted repair under equal budget consistently outperforms conventional methods.

Status of the Claim

Claim	Status	Evidence Level	Where
Witness rank is the unique invariant respecting seam-independent disclosure	Established	Disclosure Characterization Theorem; machine-checked in Lean 4 (Aristotle, standard axioms)	Part 0
Bilateral validity does not guarantee global coherence	Established	Formal theorem with machine-checked center (SCPI)	Part I
Convention mismatch causes invisible semantic failure	Demonstrated	7 families, 932 instances across 3 provenance tiers; Real-MCP: 2 families, 60 instances	Parts II, IV, VI
MCP-shaped workflows exhibit protocol-green / semantics-red	Demonstrated	Bronze+ (calendar, 1 official server, all 44 protocol checks pass); Silver (invoice, 2 non-house servers, all checks pass)	Part VI
Sheaf cohomology predicts failure better than bounded testing	Demonstrated	$R^2 \geq 0.86$ across all 7 families; conventional baselines range 0.006–0.965 (upper bound from synthetic_scaling, which has uniform convention structure; best conventional on heterogeneous families peaks at 0.734); all gap CIs exclude zero	Parts II, V, VI

Claim	Status	Evidence Level	Where
Bulla Formal Trilogy: structural identity + spectral geometry	Established	Machine-checked in Lean 4 (Aristotle, standard axioms); 703-composition real-schema corpus, with leverage predictions matched across all 240 nonzero-fee cases including 37 natural partial-CHP regime breaks	Part IV
Witness rank field-independent (TU coboundary rows)	Established	TU theorem checked in Lean 4 (signed-incidence); row structure verified across the real-schema corpus	Part IV
Repair entropy is operationally meaningful	Demonstrated	$\beta(G) = \prod \lvert C_{d,j} \rvert$ with sharp bounds $\phi+1 \le \beta \le 2^{\phi}$; verified across the 240-composition corpus including a worked motif pair	Part IV
Bounded local audits cannot certify global coherence	Established	Impossibility theorem in rank-1 signed model; 500-composition spectral indistinguishability experiment	Part V
Structural repair dominates under equal budget	Demonstrated	structural_sheaf dominates at K=1–3 across all 7 families; methods converge at K=8; +83.5% at K=8 (Silver)	Part V
Protocol consequence is specifiable	Argued	Seam paper: disclosure normal form (Theorem 7.1); receipts sound + complete in exact regime	Part III
Benchmark released	Complete	BABEL v0.1: 932 instances, 7 families, 3 tracks, 10 baselines, frozen evaluation protocol, replication pack	Part VI
LLMs fail on compositional diagnosis	Demonstrated	5 models, 3 providers; oracle CoT decomposition isolates arithmetic bottleneck ($\rho = 0.80$ ranking but $R^2 = -4.0$ magnitude)	Part VI
Hidden-convention predicate is representation-conditioned	Demonstrated	Synthetic ecology benchmark: 88% structured vs 46% flat accuracy; cross-model replication (Claude Sonnet 4 + GPT-4o)	Part VI

Claim	Status	Evidence Level	Where
Holonomy predicts real dollar error	Demonstrated	Live-pipeline validation: 27 runs across 9 convention-pair configs, $\rho = 0.795$ ($p < 7.4 \times 10^{-7}$); non-circular gold standard	Part VI
Edge-local interpretability is representationally insufficient for cyclic diagnosis	Demonstrated (two architectures)	240+ compositions, 6 baseline families, 3 domains, 4 scales, 2 models (GPT-2 Small, Gemma 2 2B); gap widens on 20× larger model; probing achieves 99.8% local accuracy with zero global signal; B6a oracle result closes aggregation objection; structural $\rho = 1.0$ in every condition	Part VII
External replication	Pending	Replication pack published; no outside confirmation yet	—
Institutional deployment	Not attempted	—	—

The remaining ceiling is external validation: outside replication and eventually an institutional pilot on a production composition.

What Would Falsify This Program?

The following conditions would materially weaken or defeat specific claims:

- If the disclosure characterization axioms (A1–A4b) are rejected as unnatural, the doctrinal layer weakens. The Lean verification establishes numerical uniqueness given the axioms; the modeling thesis that semantic interface complexes and seam-independent disclosure are the right primitives is subject to external review by applied-topology, formal-methods, and distributed-systems readers. Rejection of those primitives does not falsify the proven theorem, but it weakens the doctrinal claim that witness rank is canonically forced.
- If bounded local testing matched structural methods across families and scales, the core empirical urgency weakens. The program's central evidence depends on structural diagnosis outperforming bounded-depth testing, especially at larger composition scales. If this gap closed, the necessity argument for sheaf-cohomological diagnostics would lose its strongest leg.
- If real mixed-provenance workflows did not exhibit protocol-green / semantics-red failure, the externalization claim weakens. Bronze+ and Silver demonstrate that protocol health does not

imply semantic correctness. If careful attempts to reproduce this pattern on diverse real-world MCP compositions consistently failed, the structural threat model would be less urgent.

- If LLMs under richer prompting achieved positive R^2 on BABEL, the necessity of structural diagnostics would weaken. The current oracle CoT experiment shows LLMs can rank compositions ($\rho = 0.80$) but not compute correct magnitudes ($R^2 = -4.0$). If a tool-augmented LLM with a calculator oracle matched the structural diagnostic's prediction accuracy, the formalism would be competing with a cheaper alternative.
- If a graph-native interpretability method reliably predicted cyclic compositional failure, the representational insufficiency claim would weaken. The B6 baselines address this partially — cycle-oracle aggregation with oracle topology knowledge adds nothing over edge-local averaging — but richer learned graph representations remain untested.
- If external replication fails, empirical generality weakens. The current results are internally consistent across two domains and seven families, but they are produced by a single research program. Independent confirmation is the threshold from serious internal result to field-level evidence.
- If settlement/protocol consequence cannot be operationalized, the constitutional extension weakens. Seam and the disclosure normal form theorem (Theorem 7.1 in the doctrine paper) argue that composition protocols should carry typed disclosure receipts. If no practical protocol design can implement this at acceptable cost, the prescriptive arm of the program remains aspirational rather than consequential.
- If the SHEAF diagnostic on richer model families showed nontrivial H^1, the boundary between pairwise-sufficient and structurally-obstructed regimes would shift. The current SHEAF alignment diagnostic on 8 sentence-transformers reports trivial H^1 (SNR $\approx 1.0\times$), establishing that pairwise methods suffice in that regime. If instruction-tuned LLMs or heterogeneous multi-model compositions produced nontrivial H^1, the protocol's scope — and the threshold at which sheaf diagnostics become necessary — would need to be revised.

None of these conditions currently holds. But stating them explicitly is part of what makes a program serious rather than merely ambitious.

Reading Guide

This volume can be read in three different ways.

The first is doctrinal: start from the canonical statement, then descend to the constituent papers as instances.

1. The Composition Doctrine (Part 0) for the canonical statement: witness rank is the unique numerical invariant respecting seam-independent disclosure as the primitive operation of compositional repair. Lean-verified.
2. `SCPI` (Part I) for the topological direction: predicate invention as a descent problem, with the obstruction sequence formalized.
3. `Bridge` (Part II) for the empirical witness: the obstruction appears in concrete LLM-tool compositions.
4. `Seam` (Part III) for the protocol consequence: which is the disclosure normal form of the doctrine paper, made operational.
5. The Bulla Formal Trilogy (Part IV: Papers I–III + Signed-Incidence + Witness Geometry Beyond Fee) for the structural identity, field-independence, spectral geometry, and repair calculus on real-schema MCP corpora.
6. `SHEAF`, the Communication Bottleneck, the Coherence Cliff, and Local Validity Does Not Compose (Part V) for the frontier extension and impossibility theorem.
7. `BABEL` and Compositional Hiddenness (Part VI) for the public benchmark and representation-conditioned LLM study.
8. The Interpretability Frontier (Part VII) for the representational boundary: even the best per-component tools cannot substitute for structural diagnosis.

The second is architectural: the logical structure of the argument, beginning from impossibility and ending at the boundary.

1. `SCPI` for the formal hinge: bilateral checks cannot see cycles.
2. Local Validity Does Not Compose for the impossibility theorem: no bounded-radius local audit certifies global coherence in general.
3. `Bridge` for the empirical witness: LLMs cannot see cycles either.
4. `Seam` for the protocol consequence: and you can prove they didn't.
5. The Bulla Formal Trilogy + Signed-Incidence + Witness Geometry Beyond Fee for the structural identity, field-independence, and repair geometry of the obstruction invariant.

6. SHEAF for the frontier extension: heterogeneous-agent coordination under explicit impossibility conditions.
7. The Coherence Cliff and Communication Bottleneck: and the problem grows worse at scale.
8. BABEL, Bronze+, Silver, and Compositional Hiddenness: and here is how to measure all of it.
9. The Interpretability Frontier: and even the best per-component tools cannot substitute for structural diagnosis.
10. The Composition Doctrine: and all of this is the unique invariant any reasonable theory of compositional obstruction must recover.

The third is evidence-first: start with what you can run, then ask why it works.

1. BABEL (Part VI) for the benchmark: 932 instances, 7 families, 3 tracks. Five frontier LLMs fail.
2. Compositional Hiddenness (Part VI) for the LLM study: 88% structured vs 46% flat accuracy on a hidden-convention task.
3. The Coherence Cliff (Part V) for the scaling evidence: the regime change where bounded-depth testing collapses.
4. The Interpretability Frontier (Part VII) for the representational boundary: mechanistic interpretability at its best still cannot diagnose cyclic compositional failure.
5. The Bulla Formal Trilogy (Part IV) for the formal identities: 14 Lean-verified theorems anchoring the empirical results to a 703-composition real-schema MCP corpus.
6. SCPI (Part I) and the Composition Doctrine (Part 0) for the formal foundation: the obstruction is topological, the proof is machine-checked, and the invariant is uniquely forced.
7. Seam (Part III) for the protocol response: what to build once the obstruction is granted.

Three distinctions apply throughout:

- Doctrinal layer (Composition Doctrine) vs instance layer (the rest).
- Settled core (Parts 0–IV) vs frontier (Parts V–VII).
- Paper material vs appendix burden.

Parts 0–IV constitute the hard center of the present program. Within Part IV, the Bulla Formal Trilogy (Papers I, II, III) is the structural-spectral spine, with Signed-Incidence Structure providing the field-independence theorem and Witness Geometry Beyond Fee developing the repair-entropy decision theory. Part V is the frontier: SHEAF reaches further than the settled core, the Communication Bottleneck Theorem is the first frontier result that has passed an autonomy test, the Coherence Cliff is the program's strongest large-scale empirical evidence (a regime change where the predictive gap between the best sheaf diagnostic and the best conventional baseline nearly triples from 5 to 50 nodes), and Local Validity Does Not Compose proves the impossibility theorem in the rank-1 signed model with field-independence to the Q-valued Bulla model via signed-incidence.

Part VI is the strongest section of the omnibus in terms of external artifact weight. It presents BABEL, the public benchmark (932 instances across 7 workflow families and 3 provenance tiers) that operationalizes the central claim into a public instrument with three evaluation tracks: failure prediction, failure localization, and budgeted repair. Compositional Hiddenness tests whether frontier LLMs can recover the non-local hidden-convention predicate that the structural diagnostic computes

exactly: under structured relational prompts, accuracy is 88%; under flat prompts, 46%; the hiddenness predicate is graph-relative — the same tool has different hidden sets depending on its composition partner — and frontier LLMs revert to lexical heuristics without explicit relational framing. `Bronze+` is a mixed-provenance MCP composition where an official reference server participates in a workflow that is protocol-green and semantics-red. `Silver` extends to invoice/settlement with two non-house servers (MarkItDown + Memory). The structural diagnostic achieves $R^2 \geq 0.86$ across all seven families. Five frontier LLMs fail to rank compositions by severity (ϱ near zero); an oracle CoT experiment across all five models isolates the bottleneck as arithmetic reasoning rather than information extraction. The live-pipeline validation (BABEL §6.6) provides the first fully non-circular result: structural holonomy correlates with measured dollar error from the actual MCP server pipeline ($\varrho = 0.795$, $p < 7.4 \times 10^{-7}$).

Part VII is the capstone. The Interpretability Frontier paper tests whether mechanistic interpretability — the most powerful per-component diagnostic technology available — can substitute for structural diagnosis on cyclic compositional failure. Across 240+ compositions, three domains, four scales, two model architectures (GPT-2 Small and Gemma 2 2B), and six interpretability baseline families (including a cycle-oracle aggregator with explicit knowledge of graph topology), the answer is no. The structural diagnostic achieves $\varrho = 1.0$ in every condition; the best interpretability baseline never exceeds $\varrho = 0.758$. Probing classifiers achieve 99.8% accuracy at every edge and carry zero global signal, and cycle-oracle aggregation adds nothing over edge-local averaging. Cross-model replication on Gemma 2 2B (a 20× larger model from a different architecture family) shows the gap widens rather than narrows with model scale. The gap is representational, not aggregational. This completes the argument: the doctrine identifies the canonical invariant; the formal theory predicts local blindness; the benchmark measures it; and the interpretability paper shows that even the richest local tools inherit the same boundary — across architectures.

Intellectual Context

The local-to-global tradition. The mathematical principle that governs this program — that local consistency does not guarantee global consistency, and that the obstruction can be computed cohomologically — has a long history and a growing applied one. Leray introduced sheaves in the mid-1940s to track how local cohomological data assembles, or fails to assemble, into global invariants. The theory was developed through Cartan, Serre, and Grothendieck into one of the central instruments of twentieth-century mathematics. Its applied descendants are more recent and more various than one might expect.

Robinson (2017) established sheaves as the canonical data structure for sensor integration, using first cohomology H^1 to measure inconsistency across overlapping sensor coverages. His framework is the closest mathematical ancestor to the present program: the same H^1 that Robinson uses to detect inconsistent sensor readings is the H^1 that detects inconsistent convention interpretations in a multi-tool composition. The difference is in the application regime, not in the mathematics. Herlihy, Kozlov, and Rajsbaum (2013) connected algebraic topology to distributed computing impossibility in a foundational monograph that showed topological obstructions govern what distributed processes can and cannot compute — consensus impossibility, for instance, has a topological proof. Felber, Hummes Flores, and Rincon-Galeana (2025) extended this tradition to a sheaf-theoretic characterization of distributed task solvability, constructing a "task sheaf" whose H^1 encodes obstructions to whether processes have enough information to decide. Their obstruction is about information availability; the present program's obstruction is about convention compatibility. The mathematical structure — Čech cohomology on a nerve complex — is shared. The diagnostic target is not.

In neural network architecture, Hansen and Ghrist (2019) developed the spectral theory of cellular sheaves that underlies sheaf Laplacian methods, providing the mathematical foundations for diffusion on sheaf-valued data over cell complexes. Bodnar, Di Giovanni, and collaborators (2022) applied sheaf diffusion to address heterophily and oversmoothing in graph neural networks, showing that replacing the standard graph Laplacian with a sheaf Laplacian allows message-passing networks to handle heterophilic graphs where connected nodes have dissimilar features. The sheaf Laplacian in their work is the same mathematical object that appears in the SHEAF paper's deferred Laplacian-Cohomology Bridge agenda. The interpretability paper's finding — that edge-local features cannot recover cycle holonomy — is a negative result about exactly the kind of edge-local information these networks process: if GNN message-passing operates over edge-local features, and those features are $\mathcal{F}_{\text{edge}}$-measurable, the same representational limit applies. On the formal side, Spivak (2012) developed functorial data migration as the categorical framework for moving data between schemas, establishing the $\Sigma_f \dashv \Delta_f \dashv \Pi_f$ adjoint triple that the SchemaDiscovery adjunction in SCPI explicitly extends.

The connection between sheaf theory and multi-agent AI systems has been identified as a research direction by at least one other group: Schmid (2025) published a prospectus proposing precisely this application domain. The present program differs from Schmid's prospectus in the same way a completed building differs from an architectural sketch: the prospectus identifies the research direction;

the omnibus delivers the formal proofs, the empirical evidence across 932 benchmark instances, the protocol specification, and the representational boundary result.

The convention-mismatch regime. The regime that this program addresses is distinguished from its predecessors by a single feature: the restriction maps are not given. In Robinson's sensor integration framework, the restriction maps are determined by the physics of sensor coverage — a temperature sensor at a given location measures a specific physical quantity, and the relationship between overlapping sensors is determined by their spatial configuration. The maps are declared, known, and fixed. In Herlihy-Kozlov-Rajsbaum's distributed computing models, the topological structure is determined by the process model and the communication pattern — also declared. In the ontology alignment tradition, which has produced twenty years of systematic evaluation through the Ontology Alignment Evaluation Initiative (OAEI, 2005–2025), the schemas are declared objects and the problem is matching them: given two formal ontologies, find the correspondences.

The coherence fee arises in a different regime. When one tool's output schema specifies "amount" without stating whether the amount is in dollars or cents, and the receiving tool assumes one while the sending tool means the other, the restriction map between them encodes an interpretive convention that no party declared and no interface contract enforces. The field in the schema validates. The type checks pass. The protocol is green. The end-to-end output is wrong by a factor of 100. The sheaf condition — agreement on overlaps — is the structural minimum that ontology alignment implicitly presupposes and that the coherence fee explicitly measures when it fails.

The obstruction is computable because the conventions are typed: field formats, unit scales, temporal granularities, rounding policies. The obstruction is dangerous because the conventions are implicit. A tool that rounds to the nearest cent and a tool that truncates to the nearest dollar both report "amount" in their schemas. The bilateral interface between them validates. The cycle around them accumulates a discrepancy that no pairwise check can see. Prior applied sheaf theory solves the integration problem when the restriction maps are known. This program diagnoses the failure when they must be inferred from schema metadata, and measures the cost of that inference failure across a benchmark of 932 instances in seven workflow families and three provenance tiers.

The protocol gap. The largest multi-agent interoperability protocol in the world does not address this failure mode. The Model Context Protocol (MCP), launched by Anthropic in November 2024 and adopted by OpenAI, Google, and the broader agent ecosystem, was transferred to Linux Foundation governance in 2025. As of early 2026, its SDK downloads exceed 97 million per month. The protocol has become critical infrastructure for the agent ecosystem.

The MCP 2026 roadmap identifies priorities: Streamable HTTP transport, OAuth 2.1 authentication, server cards for discovery, task lifecycle management, and enterprise gateway patterns. These are necessary and well-designed infrastructure improvements. None of them address compositional semantic verification — whether a composition of individually valid tool invocations produces a globally valid result. The roadmap has no entry for checking whether the output of a three-tool

pipeline that passes every local validation still suffers from a convention mismatch that accumulates around a cycle in the composition graph.

The Bronze+ and Silver notes in Part V of this volume are the first systematic demonstrations that MCP-shaped compositions can be protocol-green and semantics-red: all local checks pass, the composition graph has nontrivial first cohomology, and the end-to-end output is semantically wrong. The Bronze+ calendar workflow — using one official MCP reference server alongside three custom servers — passes 44 of 44 protocol checks and produces a 30-minute escalation discrepancy that no local diagnostic catches. The Silver invoice workflow, with two non-house servers (MarkItDown and Memory), shows the same pattern in a different domain with dollar-magnitude errors. The structural diagnostic detects what the protocol stack cannot see. The live-pipeline validation in the BABEL benchmark (Part V, Section 6.6) provides the first fully non-circular result: structural holonomy correlates with measured dollar error from the actual MCP server pipeline ($\rho = 0.795$, $p < 7.4 \times 10^{-7}$), using a gold standard derived from pipeline output rather than from the planted convention structure.

The fault taxonomy literature for MCP server implementations classifies bugs at the individual-server level — serialization errors, schema violations, authentication failures — but does not address compositional semantic failure, the category of error that becomes visible only when the composition graph contains cycles. This is not a criticism of MCP or of the researchers who study it. It is a statement about what the protocol's current diagnostic surface can and cannot see.

The interpretability boundary. The interpretability paper in Part VI engages a second landscape: the mechanistic interpretability program that has become one of the most active research frontiers in AI safety. The scaling monosemanticity work (Templeton et al. 2024) demonstrated that sparse autoencoder features extracted from Claude 3 Sonnet correspond to human-interpretable concepts at scale — millions of features, many with clear semantic content. The steerable "Golden Gate Claude" demonstration showed that individual features can be clamped to alter model behavior, establishing that SAE decompositions are not merely descriptive but causally active. Anthropic's circuit tracing work (2025) represents the current frontier in per-model diagnostics, revealing compositional structure within individual models: features that compose meaningfully inside a single architecture, a shared conceptual space where reasoning happens before being translated into language. DeepMind's Gemma Scope 2 (2025) released the largest open-source interpretability toolkit, covering all Gemma 3 model sizes. Amodei (2025) stated a timeline target: reliably detecting most model problems through interpretability by 2027. This is serious, well-funded, rapidly advancing work.

The interpretability paper does not contest any of it. It asks a different question: can per-model compositional understanding — the kind that circuit tracing reveals — substitute for between-model compositional diagnosis when the failure is cycle-sensitive? The distinction between internal composition and external composition is the crux. Circuit tracing reveals that features compose meaningfully within a model: a concept in one layer connects to a concept in another, and the chain of composition can be traced. The interpretability paper tests whether that per-model understanding transfers to external composition, where independently developed components interact through in-

terfaces and the failure mode is a semantic inconsistency that appears only around cycles in the composition graph.

The answer, across six baseline families, three domains, four scales, and two model architectures (GPT-2 Small and Gemma 2 2B), is that it does not. Probing classifiers achieve 99.8% accuracy at classifying conventions at every edge — the model knows what convention each component uses — and this perfect local knowledge carries zero predictive value for composition-level failure. A cycle-oracle aggregator with explicit knowledge of graph topology adds nothing over edge-local averaging. The gap widens, rather than narrows, on the 20× larger model. The result is complementary, not adversarial. Anthropic's microscope works for its intended diagnostic target. The question is whether a microscope is the right instrument for this particular one. The interpretability paper concludes that it is not — not because the microscope is defective, but because the relevant signal is absent from the representation class that the microscope observes.

The formal ancestry. The formal layer of the program has specific ancestry that should be named. Spivak's functorial data migration framework (2012) established that database schemas can be treated as categories and that data migration between schemas is governed by an adjoint triple of functors: Σ_f (left pushforward), Δ_f (pullback), and Π_f (right pushforward). The SchemaDiscovery adjunction formalized in SCPI's Lean code extends this framework: the invariant functor Inv maps predicate sites to finite schemas, and the induced migration functors correspond to the restriction maps of the model stack. The factorization property in SchemaDiscovery.lean decomposes the SCPI result into a topological part — the sheaf condition on the nerve complex, which detects the obstruction — and a combinatorial part — Spivak-style migration between the induced schemas, which makes the obstruction computable. This decomposition is what makes SCPI tractable rather than merely abstract.

The core obstruction results carry machine-checked proofs in Lean 4. The torsor construction (Torsor.lean) establishes the principal homogeneous space structure of the solution set — the space of consistent global assignments, when it exists, is a torsor over the group of convention transformations. The counterexample (Counterexample.lean) witnesses non-composability: it constructs a specific composition graph where every bilateral interface validates and the global composition fails, with the failure traceable to the nontrivial cycle class. The Assumption D analysis (AssumptionD.lean) establishes conditions under which the obstruction persists or can be resolved — it is the formal version of the question "when can you fix a convention mismatch by local repair?" The answer: when and only when the cohomology class is trivial. The formal center is strongest where it names the obstruction sequence and weakest where it reaches into the more ambitious functorial discovery perimeter, which remains at statement-only status. The Lean formalization is not a verification of the entire program; it is a verification of the structural core on which the empirical and protocol layers build. The distinction between machine-checked center and statement-only perimeter is maintained throughout the omnibus and should be preserved by the reader.

What is new. The contribution of this program is the complete chain — formal impossibility with machine-checked proofs, empirical witness across seven workflow families, protocol specification with manifests and fraud proofs, scaling evidence from 5 to 50 agents, a public benchmark with 932 instances and a frozen evaluation protocol, real-protocol demonstrations on MCP servers of mixed provenance, and a representational boundary showing that even the most powerful component-level diagnostic technology inherits the same limit — applied to the specific failure mode of semantic convention mismatch in multi-tool compositions.

No individual link in the chain is without precedent. Sheaf cohomology has been applied to data integration. Topological methods have been applied to distributed computing. Mechanistic interpretability has been tested against many diagnostic targets. Public benchmarks exist for many failure modes. What has not existed before is the chain itself: from impossibility to measurement to protocol to boundary, on a single failure mode, with empirical evidence that the failure appears in real deployments and cross-model evidence that the representational limit is not architecture-specific.

The volume does not claim that this chain is complete. External replication has not yet occurred. The formal perimeter has unchecked regions. The protocol specification in Seam is argued, not implemented. The interpretability boundary is demonstrated on two model architectures but not yet at the 70B+ frontier. These limitations are stated explicitly throughout the volume and in the Status of the Claim table that follows the Editorial Note. The claim is not that the chain is finished. The claim is that it exists, that each link connects to the next, and that the failure mode it addresses — invisible to the largest agent interoperability protocol in the world, invisible to the most powerful per-component diagnostic technology, but visible to a structural diagnostic that costs less to compute — is consequential enough to warrant the construction.

The individual papers in this volume develop the links. The volume itself is the argument that the links compose.

References for this section:

1. Bodnar, C., Di Giovanni, F., Chamberlain, B.P., Lio, P., Bronstein, M.M. (2022). "Neural Sheaf Diffusion: A Topological Perspective on Heterophily and Oversmoothing in GNNs." arXiv:2202.04579.
2. Felber, S., Hummes Flores, B., Rincon-Galeana, H. (2025). "A Sheaf-Theoretic Characterization of Tasks in Distributed Systems." SIROCCO 2025, LNCS 15671. arXiv:2503.02556.
3. Hansen, J. and Ghrist, R. (2019). "Toward a spectral theory of cellular sheaves." Journal of Applied and Computational Topology.
4. Herlihy, M., Kozlov, D., Rajsbaum, S. (2013). Distributed Computing Through Combinatorial Topology. Springer.
5. Robinson, M. (2017). "Sheaves are the canonical data structure for sensor integration." Information Fusion, 36, 208–224.
6. Schmid, E. (2025). "Applied Sheaf Theory for Multi-Agent AI Systems: A Prospectus." arXiv:2504.17700.

7. Spivak, D.I. (2012). "Functorial Data Migration." Information and Computation, 217, 31–51.
8. Templeton, A. et al. (2024). "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." Anthropic.
9. Anthropic (2025). "Circuit Tracing: Revealing Computational Graphs in Language Models." transformer-circuits.pub.

Doctrinal Spine

The first paper in this volume is the doctrinal spine: A Witness Logic for Semantic Composition — An Axiomatic Characterization of Witness Rank.

Its decisive contribution is canonical determination. Earlier work in this program identified the coherence fee as a useful diagnostic invariant on tool compositions and proved structural results about its rank, geometry, and repair calculus. The doctrine paper asks the prior question: among all numerical invariants on the abstract category of semantic interface complexes, which one deserves to be called the obstruction invariant? The answer is forced. Once seam-independent disclosure on the cochain complex is accepted as the structural primitive of compositional repair, the unique numerical invariant respecting that primitive is the witness rank — the dimension of the first cohomology of the seam complex.

This is the same shape of result as Shannon's axiomatic characterization of entropy and Eilenberg–Steenrod's axiomatic characterization of ordinary homology. The originality does not lie in the proof, which is a short induction once the right primitive is identified. It lies in two places: in choosing the category SemComp of semantic interface complexes as the right object on which to define a theory of compositional obstruction, and in identifying disclosure independence on the cochain complex as the structural primitive of repair.

The Lean module accompanying the paper establishes the Disclosure Characterization Theorem in the abstract setting; Aristotle (Harmonic) verified the proof in April 2026 under standard axioms. The verification closes the numerical-uniqueness layer of the result. The modeling thesis — that semantic interface complexes are the right formal object, that disclosure independence is the right structural primitive, and that the exact regime is a natural subcategory — is named explicitly in the paper's scope table and is subject to external review.

The remaining papers in this volume can be read under the doctrine. SCPI in Part I is the topological direction of the same theorem, formalizing predicate invention as a descent problem and giving the obstruction sequence. Bridge in Part II is the empirical witness that the obstruction appears in concrete LLM-tool compositions. Seam in Part III specifies the protocol consequence — which is exactly the Disclosure Normal Form theorem of the doctrine paper, made operational. The Bulla Formal Trilogy (Part IV) develops the structural identity (matroid corank), spectral geometry (Kron-reduced Laplacian), and repair calculus of the witness rank on real-schema MCP corpora. SHEAF and the Communication Bottleneck + Coherence Cliff papers in Part V are the frontier extension; Local Va-

lidity Does Not Compose is the impossibility theorem the doctrine paper assumes as a precondition for the abstract setting. BABEL and Compositional Hiddenness in Part VI provide the public benchmark and a representation-conditioned LLM study. Interpretability Frontier in Part VII closes the volume by showing that even the most powerful component-level diagnostic technology inherits the boundary the doctrine identifies.

The doctrine does not subsume the individual papers. Each paper supplies content the doctrine paper itself does not: SCPI gives the topological obstruction sequence; Bridge supplies the empirical witness; Seam supplies the protocol; the Bulla trilogy supplies the structural-spectral geometry; the empirical layer supplies the benchmark and cross-architecture replication. The doctrine supplies the canonical statement that ties all of this together: there is one mathematical object the rest of the program is investigating, and it is forced by axioms anyone would write down for compositional semantic obstruction.

What follows is the paper itself.

A Witness Logic for Semantic Composition

An Axiomatic Characterization of Witness Rank

John Komkov, Independent Researcher

April 2026

Abstract

Classical program logics make local correctness compositional under a stable semantic frame. Open autonomous systems break that assumption: components satisfying their local contracts can produce globally incoherent behavior because the semantic frame itself fails to glue across interfaces. We identify the obstruction invariant forced by this failure. Working over the category of *semantic interface complexes* — finite diagrams of local semantic carriers with observable projections and latent seam dimensions — we state six natural axioms for an obstruction invariant and prove that, in the exact regime, the unique invariant satisfying them is the witness rank, equal to the dimension of the first cohomology of the seam complex. As consequences we obtain (i) a *repair duality* identifying minimum disclosure cardinality with the obstruction invariant; (ii) a *receipt normal form* exhibiting every coherence certificate as a typed disclosure derivation; and (iii) a *communication lower bound* showing that protocols whose transcript carries fewer independent declarations than the witness rank cannot soundly certify coherence. These results recast the existing local–global obstruction work, the sheaf-cohomological predicate-invention framing, and the operational receipt protocol as instances of one mathematical object.

Contents

§1. The Missing Composition Rule

Hoare's rule for sequential composition

$$\frac{\{P\}\, C_1\, \{R\} \quad \{R\}\, C_2\, \{Q\}}{\{P\}\, C_1\, ;\, C_2\, \{Q\}}$$

works because the predicate R has a stable meaning at the seam between C_1 and C_2. The same logical content is asserted on both sides of the join. This is so deeply assumed that classical program logics rarely state it as an axiom.

Throughout, we use *component* for the formal objects in a semantic interface complex; *agent* for the deployed, autonomous entity that acts on commitments after the principal who configured it is no longer in the loop; and *tool* for the operational, wire-protocol realization — typically an MCP server — by which an agent's interface is exposed to another. The doctrinal claim of this paper is at the agent level; tools and services are particular operational instantiations. The problem we address is **agentic composition** — failures that arise from convention mismatch at the boundaries between agents, not from any individual agent's unreliability. Single-agent reliability mechanisms — formal verification, deterministic execution, constrained decoding — do not eliminate this class of failure: deterministic agents with mismatched conventions still compose into globally incoherent workflows.

Open autonomous systems do not satisfy the Hoare assumption. A calendar tool returns a date in ISO-8601 with implicit UTC. An invoicing tool ingests it as a local-timezone date. Each tool's interface is type-correct; each tool's local contract is satisfied; the composition produces a wrong-by-twelve-hours invoice that is undetectable by either tool's local audit. The same predicate "this is the meeting date" carries one meaning on the left and another on the right. The seam silently changes the semantic frame.

This phenomenon is not a bug of one tool or one interface design. It is the structural feature of compositions whose components were authored by different parties under different conventions. Heterogeneous schemas, units, time zones, currencies, partition orders, null-handling discipline, identifier conventions: each is a degree of freedom along which the local meaning of a value can shift across a boundary. The empirical literature on multi-agent tool composition documents this at scale — bilateral validity does not imply compositional coherence — and provides calibrated diagnostics that detect it where local audits fail.

What is missing is not another diagnostic. What is missing is the *composition rule itself*: the axiom that says how local correctness combines into global correctness when the semantic frame is mobile across the seam.

We call this missing rule the **witness logic**, and the obstruction it tracks the **witness rank**. The contribution of this paper is to identify the mathematical object that witness rank is, and to show that — in a regime made precise below — it is forced by axioms that any reasonable theory of compositional semantic obstruction would adopt.

The shape of the argument has precedent. Shannon characterized entropy by axioms on a function of probability distributions; the axioms are continuity, symmetry, additivity, and a normalization condition. Once stated, the axioms force entropy to be $-\sum p \log p$ uniquely, up to a constant. Eilenberg and Steenrod characterized ordinary homology by five axioms — homotopy invariance, exactness, dimension, additivity, excision — and proved that any homology

theory satisfying them is naturally isomorphic to singular homology. Tutte, Whitney, and others characterized matroid rank by exchange and submodularity. In each case the axiomatic move turned a useful construction into the canonical mathematical object: the unique answer to a structural question, not one answer among many.

The bid here is the same in shape, and we want to be precise about where the originality lies. We define a category SemComp of semantic interface complexes; we identify *seam-independent disclosure* as the primitive operation of compositional repair; we prove that any numerical invariant respecting this primitive equals the dimension of the first cohomology of the seam complex (the witness rank). The operative claim is that disclosure independence — a structural property of the cochain complex of a semantic interface complex — is the right primitive for the theory. Once that primitive is accepted, the obstruction invariant is forced.

This is a useful way to read the theorem. It does not say "we wrote down a few axioms and were astonished to discover cohomology." It says: identify the right structural primitive (disclosure independence on the seam complex), and the obstruction invariant is uniquely determined. Witness rank is then not one possible diagnostic among many. It is the obstruction invariant respecting the primitive operation of repair.

The consequences are immediate. *Minimum repair* equals the obstruction by the same axioms (§6). *Coherence certificates* normalize to typed disclosure derivations (§7). *Protocols whose transcripts cannot carry an obstruction-class-many independent declarations* cannot soundly certify coherence (§8). The local–global impossibility that appears as a separate theorem in earlier work becomes a corollary: any invariant satisfying A1–A6 must give the same impossibility, because the invariant is forced.

The analogy to distributed systems is worth stating once. Lamport's happens-before relation made causality in distributed systems a structural invariant: a partial order forced by what messages can carry, not chosen by the protocol designer. The role we propose for witness rank is the analogous one for compositional semantics: a structural invariant forced by what interfaces can carry, not chosen by the verifier designer. Whether the analogy holds at the level of adoption is a matter of subsequent decades; whether it holds at the level of mathematical structure is what this paper sets out to establish.

Scope of the contribution

We owe the reader an explicit table separating *modeling thesis* (what we propose as the right abstraction) from *verified theorem* (what is mathematically forced) from *open conjecture* (what is named but not yet proved).

Layer	Claim	Status
1	Semantic interface complex (§2) is the right formal object for compositional semantic obstruction.	Modeling thesis
2	Disclosure independence on the cochain complex (§3) is the structural primitive of compositional repair.	Modeling thesis
3	The exact regime (§3.5) is a natural subcategory closed under the basic operations.	Modeling thesis (with structural support)
4	Witness rank is the unique numerical invariant respecting the primitive (Theorem 5.1).	**Verified theorem** (Lean: `disclosure_characterization`, Aristotle run `ad67beb2`)
5	Repair-cardinality equals obstruction count (Theorem 6.2).	Verified theorem (mathematical, follows from Layer 4)
6	Disclosure normal form is canonical (Theorem 7.1) and receipts are sound + complete (Propositions 7.3–7.4).	Verified theorem (mathematical, follows from Layers 4–5)
7	Communication lower bound: protocols below witness rank cannot certify coherence (Theorem 8.1).	Verified theorem (pigeonhole)
8	Bulla, the seam protocol, BABEL, the Coherence Cliff are realizations of the abstract framework (§10).	Modeling claim with empirical support
9	Surrogate-regime extension: parallel characterization with explicit correction term (Conjecture 9.1).	Open
10	Verifier universality: any sound certification factors through the witness functor (Conjecture 9.2).	Open

Layer	Claim	Status
11	Semantic FLP-style impossibility (Conjecture 9.3).	Open
12	Eval-correctness inseparability formalized as theorem (Conjecture 9.4).	Open (empirically supported)
13	Coherence-preserving interface evolution (Conjecture 9.5).	Open

The paper's *theorems* are layers 4–7. The paper's *modeling theses* are layers 1–3 and 8. The paper's *open conjectures* are layers 9–12.

The Lean formalization closes layer 4 (numerical uniqueness in the abstract exact regime). Layers 5–7 follow from layer 4 by short mathematical arguments not separately formalized. Layers 1–3 are not amenable to Lean verification — they are claims about whether a mathematical structure is the right model of a phenomenon, and external review (not machine proof) is the appropriate verification mechanism.

What this paper does, and what it does not

We prove the characterization theorem in the *exact regime*, defined in §3 as the class of semantic interface complexes whose latent cochain complex splits cleanly over the rationals — equivalently, whose witness Gram has full row rank, equivalently, the regime in which the rank-counting decomposition of the obstruction is exact rather than a lower bound. In this regime the consequences in §6, §7, §8 are theorems. Outside this regime — in the surrogate regime, where the same machinery yields a sound lower bound rather than an exact value — the analogous characterization is conjectured and named as open in §9.

We prove that *minimum independent disclosure* equals the obstruction invariant. We do not prove that *every* sound certification scheme factors through an equivalent witness object; that would be a verifier-universality theorem, which is named as open in §9 and pursued in subsequent work.

We state and prove a communication lower bound: protocols that cannot carry an obstruction-class-many independent semantic declarations cannot soundly certify coherence. This is a clean pigeonhole result and not a claim about asynchrony, fault tolerance, or liveness.

We do not claim that the axioms are minimal in the sense of Eilenberg–Steenrod's careful proof of axiom independence. We give natural arguments for each axiom and verify non-circularity, but a fully formal independence analysis is left to future work.

We work in the finite-dimensional setting. The composition is finite; the seam complex is finite-dimensional; the cohomology is computable. Extensions to infinite or limit-shaped compositions are not pursued.

What the Lean verification closes

The companion Lean module (papers/composition-doctrine/lean/) formalizes the abstract setting at the level of `DoctrineCarrier` and proves Theorem 5.1 in that setting. Aristotle (Harmonic) verified the proof on 2026-04-26 (run `ad67beb2-9f8e-48d6-9e5e-4cce51520afa`); three theorems compile sorry-free, depending only on `propext` and `Quot.sound`.

What this verification *does* close: the numerical-uniqueness layer (layer 4 of the scope table). Given the abstract structure and the axioms, the Lean proof confirms that the witness rank is the unique invariant. This is the mathematical core of the doctrine paper.

What the verification *does not* close: that the abstract structure is the right model of compositional semantic obstruction (layers 1–3). Whether `DoctrineCarrier` faithfully captures what semantic interface complexes should be, whether disclosure independence is the natural primitive of repair, and whether the exact regime is a natural subcategory — these are modeling claims about the *adequacy* of the formalization, and machine verification cannot adjudicate them. External review by applied-topology, formal-methods, and distributed-systems readers is the appropriate verification mechanism for these layers.

The honest summary: *Lean closes the numerical uniqueness theorem in the abstract exact regime; external review must test whether the structural primitive is the right one.* This separation between mathematical content (verified by Lean) and modeling thesis (subject to external review) is preserved throughout the paper and made explicit in the scope table above.

Outline

§2 defines semantic interface complexes and the category SemComp. §3 introduces the witness complex, the witness rank, and the matroid characterization of disclosure independence. §4 states the doctrinal axioms (A1, A2, A3, A4a, A4b, A5, A6) with motivation. §5 states and proves the Disclosure Characterization Theorem in the exact regime, observing that A1–A4b alone suffice (Proposition 5.4). §6 derives repair duality. §7 establishes the disclosure normal form theorem and the soundness/completeness of receipts. §8 proves the communication lower bound. §9 names five open conjectures: surrogate-regime characterization with explicit correction term, verifier universality, semantic FLP-style impossibility, eval-correctness inseparability, and coherence-preserving interface evolution. §10 returns to the existing program — Bulla, the seam protocol, BABEL, the Coherence Cliff scaling family — and identifies each as a model of the abstract framework.

---

§2. Semantic Interface Complexes

We work over $\mathbb{Q}$ throughout. Field-independence of the rank invariant between $\mathbb{Z}/2$ and $\mathbb{Q}$ is established elsewhere; we may choose the rational presentation freely.

2.1 The basic object

Definition 2.1 (Semantic interface complex). A *semantic interface complex* is a tuple

$$G = (P, \mathcal{F}, \rho, \mathcal{O})$$

where:

(a) $P = (P, \leq)$ is a finite poset, the *context poset*. Its elements are *contexts*; covering relations $p \lessdot q$ represent refinement of context.

(b) $\mathcal{F} : P^{\mathrm{op}} \to \mathsf{Vect}_{\mathbb{Q}}$ is a finite-dimensional presheaf — the *convention presheaf* — assigning to each context $p \in P$ a finite-dimensional rational vector space $\mathcal{F}(p)$ of *conventions* at p.

(c) $\rho_{p,q} : \mathcal{F}(q) \to \mathcal{F}(p)$ for $p \leq q$ are the *restriction maps* of $\mathcal{F}$, satisfying $\rho_{p,p} = \mathrm{id}$ and $\rho_{p,r} = \rho_{p,q} \circ \rho_{q,r}$ for $p \leq q \leq r$.

(d) $\mathcal{O} \subseteq \mathcal{F}$ is a subpresheaf, the *observable subpresheaf*. For each p, $\mathcal{O}(p) \subseteq \mathcal{F}(p)$ is a designated subspace, and the restriction maps satisfy $\rho_{p,q}(\mathcal{O}(q)) \subseteq \mathcal{O}(p)$.

The quotient presheaf $\mathcal{L} := \mathcal{F}/\mathcal{O}$ is the *latent presheaf*. A class $[v] \in \mathcal{L}(p)$ is a *seam dimension at* p. The restriction maps of $\mathcal{L}$ are induced from those of $\mathcal{F}$.

We write $|G| := \dim_{\mathbb{Q}} \bigoplus_{p \in P} \mathcal{F}(p)$ for the size of the complex.

2.2 Reading the definition

The object encodes three pieces of data: *where* meanings live (the contexts P), *what* meanings are available there (the conventions $\mathcal{F}$), and *which of those meanings are exposed* across boundaries (the observables $\mathcal{O}$). The seam information — meanings that two contexts could disagree on without either being able to detect the disagreement locally — is precisely what is *not* in $\mathcal{O}$, and that residue is the latent presheaf $\mathcal{L}$.

The classical case where there is no seam problem corresponds to $\mathcal{F} = \mathcal{O}$: every convention is observable across every refinement. In that case the obstruction we will define vanishes by construction. The interesting case — the open-systems case — is when $\mathcal{O}$ is strictly smaller than $\mathcal{F}$ at some contexts.

2.3 Operations and morphisms

Three operations on semantic interface complexes are needed for the axioms to have content. We define each explicitly; their categorical packaging is routine and we omit it.

(I) Isomorphism. An *isomorphism* $\phi : G \to G'$ is an order-isomorphism $\phi_P : P \to P'$ together with a family of linear isomorphisms $\phi_p : \mathcal{F}(p) \to \mathcal{F}'(\phi_P(p))$ commuting with restriction and carrying $\mathcal{O}(p)$ onto $\mathcal{O}'(\phi_P(p))$.

(II) Disjoint sum. Given G and H with disjoint context posets, the *disjoint sum* $G \sqcup H$ has context poset $P \sqcup P'$, convention presheaf $\mathcal{F} \oplus \mathcal{F}'$, and observables $\mathcal{O} \oplus \mathcal{O}'$.

(III) Disclosure. Given G, a context $p \in P$, and a one-dimensional subspace $W \subseteq \mathcal{F}(p)$ with $W \not\subseteq \mathcal{O}(p)$, the *disclosure of* W *at* p, denoted $G \setminus\!\setminus W$, is the complex $(P, \mathcal{F}, \rho, \mathcal{O}_W)$ where

$$\mathcal{O}_W(q) := \begin{cases} \mathcal{O}(p) + W & \text{if } q = p \\ \mathcal{O}(q) & \text{otherwise.} \end{cases}$$

The subpresheaf condition is preserved because $\rho_{p,q}(\mathcal{O}(q)) \subseteq \mathcal{O}(p) \subseteq \mathcal{O}(p) + W$ for all $q \geq p$. So disclosure modifies the observable subpresheaf at exactly one context, removing one dimension from $\mathcal{L}(p)$ while leaving $\mathcal{L}(q)$ unchanged for $q \neq p$.

The *underlying disclosure* of $G \setminus\!\setminus W$ is the data (p, W). Two disclosures (p_1, W_1) and (p_2, W_2) on the same G are *cochain-independent* if the corresponding 1-cochains in the seam complex (§3.1 below) are linearly independent modulo coboundaries. Independence is a property of the seam complex of G, definable without reference to any invariant on G.

2.4 Covers

Definition 2.2 (Cover). A *cover* of G is a pair of full subcomplexes (U, V) — that is, restrictions to subposets $P_U, P_V \subseteq P$ inheriting the convention data — such that:

(i) $P_U \cup P_V = P$.

(ii) The intersection $U \cap V$ is the full subcomplex on $P_U \cap P_V$.

(iii) The convention and observable data on $U \cap V$ agree with the restriction from both U and from V.

A cover is *compatible* if the restriction maps from the union to U and to V exhibit $\mathcal{F}_G$ as the pullback of $\mathcal{F}_U \to \mathcal{F}_{U\cap V} \leftarrow \mathcal{F}_V$, and similarly for $\mathcal{O}$. Compatibility ensures that G is recoverable from its restrictions U, V, and $U \cap V$ as a presheaf — this is the condition under which Mayer–Vietoris exact sequences exist.

2.5 The category SemComp

Let SemComp be the category whose objects are finite semantic interface complexes and whose morphisms are generated under composition by isomorphisms, disjoint-sum injections, disclosure morphisms, and cover restrictions. We do not need a complete description of all morphisms in SemComp; we need only that the operations (I)–(III) and Definition 2.2 are well-defined and

that an invariant on $\mathsf{SemComp}$ — a function $I : \mathrm{Ob}(\mathsf{SemComp}) \to \mathbb{Z}_{\geq 0}$ that is constant on isomorphism classes — has a well-defined value on each operation.

2.6 The exact regime

The axiomatic characterization in §5 is proved in a restricted class of objects, the *exact regime*. We give the abstract definition here; §3.5 shows that it agrees with the operationally-defined exact regime in the existing implementation literature.

Definition 2.3 (Exact regime, informal). A semantic interface complex $G = (P, \mathcal{F}, \rho, \mathcal{O})$ is in the *exact regime* if its latent cochain complex (Definitions 3.1–3.3) is split exact in degree ≥ 2: equivalently, if the rationally-defined rank decomposition of the seam coboundary satisfies

$$\dim_{\mathbb{Q}} H^1(G) = \operatorname{rank} \delta_1^{\mathcal{F}} - \operatorname{rank} \delta_1^{\mathcal{O}}$$

without correction terms from higher coboundaries. The precise form is given in Definition 3.10.

Let $\mathsf{SemComp}_{\mathrm{ex}}$ be the full subcategory of $\mathsf{SemComp}$ consisting of complexes in the exact regime. The category is closed under isomorphism, disjoint sum, disclosure, and compatible covers — these closure properties are immediate from the definitions and we record them without proof here.

2.7 What this section achieves

We have introduced a finite combinatorial-algebraic object — a presheaf of finite-dimensional vector spaces on a finite poset with a designated subpresheaf of "observables" — together with three operations (isomorphism, disjoint sum, disclosure) and a notion of cover. Nothing about this object refers to any specific implementation, programming language, protocol, or system. The object is general enough to carry instances ranging from finite-dimensional simplicial sheaves with observable structure to operational agentic compositions over MCP tool surfaces; it is specific enough that all axioms in §4 have content. The next section equips it with the cochain complex on which the obstruction invariant lives.

---

§3. The Witness Complex and the Witness Rank

This section defines the cochain complex of $\mathcal{F}/\mathcal{O}$, the obstruction module, and the witness rank. The construction is standard sheaf cohomology specialized to a finite presheaf on a finite poset; it is included in full because subsequent sections refer to specific cochain-level objects, and because the *witness Gram* presentation — useful for §3.5 and for connection to the existing program — is not the standard textbook presentation.

3.1 The cochain complex

Fix a semantic interface complex $G = (P, \mathcal{F}, \rho, \mathcal{O})$. Let $\mathcal{L} := \mathcal{F}/\mathcal{O}$ be the latent presheaf with restrictions $\bar{\rho}_{p,q} : \mathcal{L}(q) \to \mathcal{L}(p)$.

Definition 3.1 (Cochain spaces). For each $n \geq 0$, the *n-cochain space* is

$$C^n(G) := \prod_{p_0 < p_1 < \cdots < p_n} \mathcal{L}(p_0),$$

the product over strict chains of length $n+1$ in P, of the latent space at the smallest element.

Definition 3.2 (Coboundary). The *coboundary* $\delta^n : C^n(G) \to C^{n+1}(G)$ is the alternating-sum map sending a cochain c to the cochain whose value on a chain $p_0 < p_1 < \cdots < p_{n+1}$ is

$$(\delta^n c)(p_0, \ldots, p_{n+1}) = \bar{\rho}_{p_0,p_1}(c(p_1, \ldots, p_{n+1})) + \sum_{i=1}^{n+1} (-1)^i \, c(p_0, \ldots, \widehat{p_i}, \ldots, p_{n+1}).$$

The composition $\delta^{n+1} \circ \delta^n$ is zero, by the standard simplicial calculation.

The pair $(C^\bullet(G), \delta^\bullet)$ is the *seam cochain complex of G*.

Definition 3.3 (Cohomology). The cohomology of G in degree n is

$$H^n(G) := \ker \delta^n / \operatorname{im} \delta^{n-1}.$$

The first cohomology $H^1(G)$ is the *obstruction module* of G.

3.2 The witness rank

Definition 3.4 (Witness rank). The *witness rank* of G is

$$r(G) := \dim_{\mathbb{Q}} H^1(G).$$

This is the central invariant. It is a non-negative integer; it vanishes when every seam dimension agrees on overlaps; it grows by one for each independent cycle along which the latent data fails to glue.

Lemma 3.5. $r(G \sqcup H) = r(G) + r(H)$.

Proof. The cochain complex of a disjoint sum is the direct sum of the cochain complexes; cohomology commutes with finite direct sums of complexes; dim is additive on direct sums of vector spaces. □

(This is the additivity property that enters axiom A3.)

3.3 Disclosures act on cochains

Disclosure of a one-dimensional latent subspace $W \subseteq \mathcal{F}(p)$ changes the observable subpresheaf and hence the latent presheaf: the dimension of $\mathcal{L}(p)$ drops by one, with corresponding changes at $q \geq p$. This induces a chain map of cochain complexes that we now describe.

Lemma 3.6 (Disclosure chain map). Let $\sigma = (p, W)$ be a disclosure on G, and let $G' = G \setminus\setminus W$. There is a canonical chain map $\sigma_* : C^\bullet(G') \to C^\bullet(G)$ exhibiting $C^\bullet(G')$ as a subcomplex of $C^\bullet(G)$, with quotient cochain complex concentrated in degrees where W contributes.

Proof sketch. The simplified disclosure (§2.3 III) gives $\mathcal{L}'(p) = \mathcal{L}(p)/W$ and $\mathcal{L}'(q) = \mathcal{L}(q)$ for $q \neq p$. The quotient map $\mathcal{L}(p) \twoheadrightarrow \mathcal{L}'(p)$ splits canonically over $\mathbb{Q}$, inducing the inclusion $C^\bullet(G') \hookrightarrow C^\bullet(G)$. The quotient $C^\bullet(G)/C^\bullet(G')$ is the subcomplex of $C^\bullet(G)$ generated by cochains valued in W at chains starting at p. □

The class $[\sigma] \in H^1(G)$ defined by σ is the cohomology class of the cocycle representing W in the seam complex. Two disclosures σ_1 and σ_2 are *cochain-independent* (§2.3, restated cohomologically) precisely when $[\sigma_1]$ and $[\sigma_2]$ are linearly independent in $H^1(G)$.

Lemma 3.7 (Independent disclosure decreases r by one). If σ is a disclosure on G with $[\sigma] \neq 0$ in $H^1(G)$, then

$$r(G \setminus\setminus W) = r(G) - 1.$$

Proof. The long exact sequence of the pair $(C^\bullet(G), C^\bullet(G'))$ gives

$$\cdots \to H^0(C^\bullet(G)/C^\bullet(G')) \to H^1(G') \to H^1(G) \to H^1(C^\bullet(G)/C^\bullet(G')) \to \cdots$$

The quotient cochain complex is concentrated in low degree (by Lemma 3.6) and its first cohomology is at most one-dimensional, generated by the class $[\sigma]$. When $[\sigma] \neq 0$, the connecting map kills exactly that class, dropping the rank of H^1 by one. □

This is the cochain-level statement of axiom A4a.

Remark 3.7.1 (Matroid structure of disclosure independence). The set of disclosures on G, ordered by cochain-independence, satisfies the standard matroid axioms (hereditary, exchange, augmentation), with rank function equal to r. Equivalently, the quadratic form on $C^1(G)$ induced by the witness Gram defines a representable matroid whose rank is the witness rank. This is the matroid-theoretic content of [hierarchical-fee] (Komkov 2026), and it gives an alternative axiomatic foundation for the framework: instead of stating axioms on the numerical invariant I, one can state axioms on the *independence relation* and recover witness rank as the matroid rank function. The doctrine paper takes the numerical-invariant route for clarity; the matroid route is equivalent.

3.4 The witness Gram

The cochain complex $(C^\bullet(G), \delta^\bullet)$ has a finite-dimensional first-degree segment

$$C^0(G) \xrightarrow{\delta^0} C^1(G).$$

Choosing bases makes δ^0 a rational matrix $D \in \mathrm{Mat}_{|C^1|,|C^0|}(\mathbb{Q})$. Two related symmetric matrices arise:

Definition 3.8 (Witness Gram). The *witness Gram* of G (with respect to chosen bases) is

$$K(G) := D^\top D \in \mathrm{Mat}_{|C^0|,|C^0|}(\mathbb{Q}).$$

The *Kron-reduced witness Gram* is the Schur complement of $K(G)$ over the observable rows and columns: writing

$$K(G) = \begin{pmatrix} K_{\mathcal{O}\mathcal{O}} & K_{\mathcal{O}\mathcal{L}} \\ K_{\mathcal{L}\mathcal{O}} & K_{\mathcal{L}\mathcal{L}} \end{pmatrix},$$

the Kron-reduced Gram is

$$K^{\mathrm{red}}(G) := K_{\mathcal{L}\mathcal{L}} - K_{\mathcal{L}\mathcal{O}} K_{\mathcal{O}\mathcal{O}}^{+} K_{\mathcal{O}\mathcal{L}},$$

where $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse.

Lemma 3.9. $\operatorname{rank} K^{\mathrm{red}}(G) = r(G)$ in the exact regime.

Proof sketch. This is the witness-Gram rank identity in finite-dimensional rational presentation; for full proof see Komkov (2026), "Witness Geometry Beyond Fee," Theorem 1. The Kron reduction realizes the latent obstruction module as a quadratic form on the latent cochain space modulo observable contributions, and its rank is the dimension of the obstruction module in the regime where the cochain complex is rationally split — that is, the exact regime. □

(This Lemma supplies the empirical computation in the existing program: $r(G)$ is computable in polynomial time as a matrix rank over the rationals, with no exotic structure required.)

3.5 The exact regime, made precise

Definition 2.3 left the exact regime informally specified. We now give the precise statement.

Definition 3.10 (Exact regime, precise form). A semantic interface complex G is in the *exact regime* $\mathsf{SemComp}_{\mathrm{ex}}$ if its cochain complex $(C^\bullet(G), \delta^\bullet)$ is *split exact in degree* ≥ 2: equivalently, if the rational quasi-isomorphism class of the cochain complex is concentrated in degrees 0 and 1, and the rank of $H^1(G)$ is computed without contributions from higher coboundaries.

Equivalently, in cochain-level terms: G is in the exact regime iff $\delta^n = 0$ for $n \geq 2$, or the higher-degree obstructions are quotient-trivially absorbed by the degree-1 obstruction. In this

regime,

$$r(G) = \dim C^1(G) - \operatorname{rank} \delta^0 - 0 = \operatorname{rank} \delta_1^{\mathcal{F}} - \operatorname{rank} \delta_1^{\mathcal{O}}$$

holds without correction. By Lemma 3.9, $r(G) = \operatorname{rank} K^{\mathrm{red}}(G)$.

Operational sufficient conditions. The exact regime is characterized by abstract properties of the cochain complex; for compositions arising in practice, it is implied by structural conditions on the data. Two such conditions, used in the existing implementation literature, are:

(i) *Disjoint field decomposition (DFD).* The latent presheaf $\mathcal{L}$ decomposes as a direct sum $\mathcal{L} = \bigoplus_d \mathcal{L}_d$ over independent field-types, each acting on disjoint contexts.

(ii) *Class-homogeneous partition (CHP).* Within each field-type, the convention dimension at every context is one-dimensional.

Under (i)+(ii), every coboundary above degree 1 vanishes, and $G \in \mathsf{SemComp}_{\mathrm{ex}}$. This is Theorem 2 of Komkov (2026), "Hierarchical Decomposition of Coherence Fee," and supplies the bridge between the abstract definition above and the operational regime named in the implementation literature.

Examples and non-examples. To prevent the exact regime from feeling like "the regime in which the theorem works," we record several structural cases.

Case	In exact regime?	Reason
Acyclic observable composition ($\mathcal{F} = \mathcal{O}$)	Yes	$H^1 = 0$, cochain complex trivial in non-zero degree
Elementary hidden cycle C_e (§3.6)	Yes	Latent presheaf concentrated at one context; degree-1 obstruction only
Disjoint sum of finitely many elementary cycles	Yes	Cochain complex is direct sum; closure under disjoint sum (§2.6)
DFD + CHP composition with bounded depth	Yes	Theorem 2 of Komkov (2026); the operational sufficient condition
Composition with *coloop* — a latent dimension that is simultaneously a coboundary and a non-trivial cocycle in different bases	No	Cochain complex fails to split rationally; surrogate-regime correction term required

Case	In exact regime?	Reason
Composition with non-trivial higher-degree cohomology ($H^2 \neq 0$)	No	Definition 3.10's split-exactness in degree ≥ 2 fails
Class-heterogeneous partition (CHP violation) — a single field-type whose convention dimension varies across contexts	No	The latent cochain complex has non-trivial extension data; the rank-counting decomposition gives a lower bound, not equality

The exact regime is structurally characterized: it is the class of complexes whose seam cochain complex admits a clean rational splitting in degrees ≥ 1. Outside it, the surrogate-regime correction term of Conjecture 9.1 enters.

3.6 Worked example: an elementary cycle

The smallest non-trivial example. Let $P = \{p_1, p_2, p_3\}$ be an antichain (incomparable), with a refinement element $\hat{p}$ greater than each. Let $\mathcal{F}(p_i) = \mathbb{Q}\langle e_i \rangle$ for $i = 1, 2, 3$ (each one-dimensional, with basis e_i), and $\mathcal{F}(\hat{p}) = \mathbb{Q}^3 = \mathbb{Q}\langle e_1, e_2, e_3 \rangle$ (the cone). Restriction maps from $\hat{p}$ project onto the corresponding axis. Let $\mathcal{O}(p_i) = \mathcal{F}(p_i)$ for each i (each individual context observes its convention), but $\mathcal{O}(\hat{p}) = \mathbb{Q}\langle e_1 - e_2, e_2 - e_3 \rangle$ — only differences are observable in the cone, not absolute values.

The latent presheaf $\mathcal{L}$ is zero on each p_i (everything is observable locally) but is one-dimensional on $\hat{p}$, generated by the "absolute mean" $e_1 + e_2 + e_3$ modulo observable differences.

The cochain complex has $C^0 = 0$ (latent is zero on the leaves) and $C^1 = \mathbb{Q}$ (the absolute mean class). The coboundary $\delta^0 = 0$. So $H^1(G) = \mathbb{Q}$, and $r(G) = 1$.

Discloseing $W = \mathbb{Q}\langle e_1 + e_2 + e_3 \rangle \subseteq \mathcal{F}(\hat{p})$ — declaring the absolute mean to be observable — adds the disclosed dimension to $\mathcal{O}$ and reduces $\mathcal{L}(\hat{p})$ to zero. Now $r(G_{\text{disclosed}}) = 0$.

This is the *elementary hidden semantic cycle*: three local contexts whose pairwise differences are observable but whose absolute alignment is not. It will play the role of generator in axiom A5.

3.7 A two-field example

The elementary cycle has $r = 1$. To exhibit higher rank concretely — and to verify the additivity and repair-duality claims explicitly — consider its two-field generalization.

Take the same context poset $P = \{p_1, p_2, p_3, \hat{p}\}$ but enlarge the convention spaces. At each leaf p_i, $\mathcal{F}(p_i) = \mathbb{Q}\langle d_i, a_i \rangle$ — two fields, "date" and "amount." At the cone, $\mathcal{F}(\hat{p}) = \mathbb{Q}^6$ spanned by all

six. Let $\mathcal{O}(\widehat{p})$ contain only the four pairwise differences within each field: $d_1 - d_2$, $d_2 - d_3$, $a_1 - a_2$, $a_2 - a_3$. The two absolute means $d_1 + d_2 + d_3$ and $a_1 + a_2 + a_3$ remain latent.

The latent presheaf $\mathcal{L}(\widehat{p})$ is two-dimensional, generated by the date-mean class and the amount-mean class. The seam cochain complex has $C^0 = 0$ and $C^1 = \mathbb{Q}^2$, with $\delta^0 = 0$, so $H^1(G) = \mathbb{Q}^2$ and $r(G) = 2$.

Disclosing $W_1 = \mathbb{Q}\langle d_1 + d_2 + d_3 \rangle$ promotes the date mean to observable. The cohomology class $[\sigma_1]$ corresponding to W_1 is non-zero in $H^1(G)$; by Lemma 3.7, $r(G \setminus\setminus W_1) = 1$. Disclosing $W_2 = \mathbb{Q}\langle a_1 + a_2 + a_3 \rangle$ next reduces r to zero. The disclosure normal form (Theorem 7.1) is the receipt with disclosure set $\{W_1, W_2\}$, of cardinality $2 = r(G)$, matching repair duality (Theorem 6.2).

This example is concrete enough to compute the witness Gram explicitly: the coboundary matrix δ^0 is zero on a 2×0 matrix block (no observable carriers contribute coboundaries), so $K = D^\top D$ has rank 2, equal to $r(G)$. Operationally, this models a two-tool composition where each leaf reports a local date and amount, and the cone is a downstream component requiring both the date-frame and amount-frame to be globally aligned. Two independent disclosures suffice; no fewer.

3.8 What this section achieves

We have a finite, computable invariant $r : \mathrm{Ob}(\mathsf{SemComp}) \to \mathbb{Z}_{\geq 0}$, defined as the dimension of the first cohomology of the seam complex of G, computable as the rank of the Kron-reduced witness Gram. The invariant is additive over disjoint sums (Lemma 3.5) and decreases by one under cochain-independent disclosure (Lemma 3.7). The exact regime $\mathsf{SemComp}_{\mathrm{ex}}$ in which r behaves with these clean properties has both an abstract characterization (Definition 3.10) and operational sufficient conditions (DFD + CHP). The next section asks: is r the *unique* invariant with these clean properties? §4 states the axioms; §5 proves it is.

---

§4. The Six Axioms

We now state the axioms that characterize the obstruction invariant. Each axiom is a property that any reasonable measure of compositional semantic obstruction should satisfy: invariance under renaming, vanishing on trivial objects, additivity over independent components, exactness under disclosure, normalization on the elementary obstruction generator, and locality under cover decomposition. The cumulative force of the six is what determines the invariant uniquely; the cumulative naturalness of the six is what makes the resulting characterization a *canonical* description rather than a parochial one.

Throughout this section, $I : \mathrm{Ob}(\mathsf{SemComp}_{\mathrm{ex}}) \to \mathbb{Z}_{\geq 0}$ is a function on the exact regime, constant on isomorphism classes (a *numerical invariant*).

4.1 The axioms

(A1) Isomorphism invariance. For all $G, G' \in \mathsf{SemComp}_{\mathrm{ex}}$:

$$G \cong G' \implies I(G) = I(G').$$

(A2) Local triviality. If G has empty latent presheaf — that is, $\mathcal{F} = \mathcal{O}$ on every context, equivalently $\mathcal{L} = 0$ — then

$$I(G) = 0.$$

(A3) Additivity. For all $G, H \in \mathsf{SemComp}_{\mathrm{ex}}$:

$$I(G \sqcup H) = I(G) + I(H).$$

(A4a) Disclosure sensitivity. Let $\sigma = (p, W)$ be a disclosure on G whose induced cohomology class $[\sigma] \in H^1(G)$ is non-zero. Then

$$I(G \setminus\setminus W) = I(G) - 1.$$

(A4b) Coboundary blindness. Let $\sigma = (p, W)$ be a disclosure whose induced cohomology class is zero — i.e., $[\sigma] = 0$, equivalently, W corresponds to a coboundary cochain in $\operatorname{im} \delta^0$. Then

$$I(G \setminus\setminus W) = I(G).$$

Together A4a and A4b say: independent disclosures reduce the invariant by one; redundant (coboundary) disclosures leave it unchanged. We refer to the pair as *disclosure exactness* when convenient.

(A5) Cycle normalization. Let $C_e \in \mathsf{SemComp}_{\mathrm{ex}}$ be the elementary hidden semantic cycle (the worked example of §3.6). Then

$$I(C_e) = 1.$$

(A6) Excision under compatible cover. Let (U, V) be a compatible cover of G (Definition 2.2), with $G \in \mathsf{SemComp}_{\mathrm{ex}}$. Then

$$I(G) = I(U) + I(V) - I(U \cap V).$$

Equivalently, I is the unique (up to scaling) Mayer–Vietoris-additive function on $\mathsf{SemComp}_{\mathrm{ex}}$ that vanishes on the empty complex and respects compatible covers.

Remark. For complexes outside the exact regime, the Mayer–Vietoris sequence at H^1 may not be exact, and the equality above must be modified by a non-negative correction term recording

classes lost in the gluing. The surrogate-regime version of A6 is named in §9 and pursued elsewhere.

4.2 Motivation, axiom by axiom

(A1) — Why isomorphism invariance. Renaming a context, relabeling a basis, or applying any structure-preserving bijection should not change a measure of obstruction. If I depended on the labels rather than on the structure, it would not be an invariant in the mathematical sense; it would be a metric of the presentation. A1 is the minimum content of "invariant."

(A2) — Why local triviality. When every convention is observable everywhere — no latent dimensions, no seam information — there is nothing across the seam that could fail to glue. Compositions over such complexes are the classical case in which Hoare's rule applies without modification. The obstruction must vanish here, since by hypothesis there is none.

(A3) — Why additivity. Two compositions sharing no contexts cannot interact; their obstructions cannot interfere. The obstruction of the disjoint sum is the simple sum. This is the same axiom that appears in homology (additivity), in entropy (independence of sources), and in valuations on disjoint unions of measurable sets. A non-additive invariant would entangle non-interacting components.

(A4a) — Why disclosure sensitivity. When a single previously-latent dimension corresponding to a non-trivial cohomology class is promoted to observable, the seam information at that dimension becomes locally checkable. The class is now observable and contributes nothing further to the obstruction — the rank drops by one. The "-1" carries both the cancellation rule (one disclosure removes one class) and the scale (the unit is "one elementary obstruction").

(A4b) — Why coboundary blindness. A disclosure whose cochain element is already in $\operatorname{im}\delta^0$ is *redundant*: it contributes no information beyond what the observable subpresheaf already encodes. Such a disclosure should not change the invariant. Without A4b, the theory could not distinguish redundant disclosures from informative ones, and minimum-repair cardinality (§6) would be ill-defined.

Non-circularity note (key). The conditions $[\sigma] \neq 0$ (in A4a) and $[\sigma] = 0$ (in A4b) are checked *entirely in the cochain complex of G* — they do not refer to I. The 1-cochain corresponding to the disclosure is computable from W as a basis vector at chains starting at p; its cohomology class is computable as the equivalence class modulo $\operatorname{im}\delta^0$. The axiom imports the *cohomological structure* of the seam complex (which is part of the data of G), not the *numerical invariant* we are characterizing. The honest reading: A4a/A4b say the invariant respects the cochain structure of disclosure-independence — a property of the cochain complex, not a self-reference.

This is the locus of non-tautology and is discussed adversarially in `notes/anti-tautology.md`.

(A5) — Why cycle normalization. The elementary hidden semantic cycle (§3.6) is the smallest non-trivial obstruction: three local contexts agreeing pairwise but failing globally. By

A2 it is non-trivial ($I(C_e) > 0$); by symmetry of the cycle and A1 it is invariant under permutation of the three legs; the natural normalization is $I(C_e) = 1$. Any other choice of normalization gives an invariant proportional to I rather than equal — so A5 fixes the scale.

This axiom is the one that turns I from a function determined "up to multiplicative constant" into a function determined exactly. It plays the role that "$I(\text{point}) = 1$" or "$\dim I(\mathbb{Z}) = 1$" plays in homology, or that "$\log_2 2 = 1$" plays in Shannon entropy.

(A6) — Why excision. Compositions can be decomposed into overlapping pieces: a cover. Obstruction information should be local in the sense that the obstruction of G is determined by the obstructions of U, V, and their intersection $U \cap V$. The standard topological account of this is the Mayer–Vietoris sequence; in the exact regime, the sequence is exact and the axiom states that obstruction satisfies plain inclusion-exclusion. (Outside the exact regime, an additional non-negative correction term records cohomology classes that gluing can lose or create; that surrogate-regime version of A6 is named in §9.)

This is the axiom that makes the invariant *local*: $I(G)$ is determined by I on small pieces. Without it, the invariant could not be computed from local data, and the entire program of "compositional obstruction" would lose meaning.

Non-circularity note. Compatible covers (Definition 2.2) are properties of the cochain complex of G and its restrictions to U, V, $U \cap V$ — definable without any reference to I. The axiom is a relation among four values of I ($I(G), I(U), I(V), I(U \cap V)$), not a self-referential condition.

4.3 Independence

The seven axioms (A1, A2, A3, A4a, A4b, A5, A6) exhibit a non-uniform redundancy structure. Some are independent of the others; some are implied within the exact regime, but become non-redundant outside.

Independence of A4b within the exact regime. A4b is independent of A1, A2, A3, A4a, A5, A6 — exhibited by the *latent dimension invariant* $I_L(G) := \dim_{\mathbb{Q}} \mathcal{L}_{\text{tot}}(G)$, which satisfies all six other axioms (iso-invariance, vanishing on observable-only complexes, additivity under disjoint sum, sensitivity to disclosures, normalization on the elementary cycle, and inclusion-exclusion under cover) but fails A4b: I_L drops by exactly one under *every* disclosure, including cochain-trivial ones. Without A4b, redundant disclosures would inflate the rank-counting and the invariant would not be the cohomological one. A4b is the axiom that distinguishes information-bearing disclosures from redundant ones.

Independence of A4a. A4a is independent of A1, A2, A3, A4b, A5, A6 — the constant invariant $I = 0$ satisfies all of those (vacuously, in some cases) but fails A4a (which requires non-zero drop on non-trivial disclosure if any exists).

A1 is independent in the trivial sense that no permutation-non-invariant function can satisfy the iso quotient.

Internal redundancy in the exact regime. Proposition 5.4 shows that A5 and A6 are implied by A1+A2+A3+A4a+A4b within $\mathsf{SemComp}_{\mathrm{ex}}$. Their inclusion in the doctrinal axiom system is justified outside the exact regime, where they become non-redundant constraints (§9 and the companion note `axioms-discipline.md`).

Independence of A2, A3, and A5 from each other is more subtle. The companion note records the partial verifications and identifies the open cases. A fully formal minimality proof in the Eilenberg–Steenrod style is left to future work.

4.4 What the axioms exclude

It is worth being explicit about what the axioms forbid:

- **They forbid scale-dependence**: by A1+A5, I is the rank, not a quadratic form.
- **They forbid cross-component coupling**: by A3, I is sum-additive across disjoint components.
- **They forbid hidden penalties**: by A2, I vanishes on classically composable systems.
- **They forbid disclosure overcounting**: by A4a, one independent disclosure removes exactly one obstruction class — not two, not zero, not "depends on context." By A4b, redundant disclosures cannot inflate the count.
- **They forbid pathological non-locality**: by A6, I on a whole is determined by inclusion-exclusion on a compatible cover. (Outside the exact regime, an explicit gluing correction enters; this is named in §9.)

What the axioms *do not* forbid: any specific computational realization of I on a particular complex. The axioms constrain the abstract function; many concrete cochain-level computations may compute the same function. The characterization theorem says that any abstract function satisfying A1–A6 in the exact regime is the witness rank; it does not say that any one method of computing the rank is canonical, only that the value is.

4.5 What this section achieves

We have stated the doctrinal axioms (A1, A2, A3, A4a, A4b, A5, A6) on numerical invariants of the exact regime, motivated each as a property any reasonable measure of compositional obstruction should satisfy, and verified that the critical non-circularity points (A4a, A4b, and A6) are satisfied: each axiom is a relation involving the cochain-level structure of the seam complex but not the invariant being characterized.

The next section proves that the witness rank r — already known to satisfy these axioms by the lemmas of §3 — is the unique numerical invariant satisfying A1, A2, A3, A4a, A4b in the exact regime, with A5 and A6 automatic.

---

§5. The Disclosure Characterization Theorem

5.1 Two layers of the result

The result of this section operates at two layers, which we name explicitly to keep the locus of originality clear.

Layer 1 — the structural primitive. §2 defined disclosure as a morphism on semantic interface complexes; §3 defined cochain-level independence of disclosures. *The non-trivial conceptual content of the framework is that seam-independent disclosure is the right primitive operation of compositional repair.* Once accepted as primitive, disclosure independence is purely structural: it is a property of the cochain complex of G, definable without reference to any numerical invariant.

Layer 2 — numerical uniqueness. Given the primitive, the question is which numerical invariants of $\mathsf{SemComp}_{\mathrm{ex}}$ respect it. The theorem of this section says: only one such invariant exists, the witness rank.

The theorem is honest about the locus of work: the originality is not in proving Layer 2 (the proof is a short induction). The originality is in identifying Layer 1 — that disclosure independence on the cochain complex is the structural primitive of the theory. Once that is granted, the rank is forced.

5.2 Statement

Theorem 5.1 (Disclosure Characterization of Witness Rank). Let $I : \mathrm{Ob}(\mathsf{SemComp}_{\mathrm{ex}}) \to \mathbb{Z}_{\geq 0}$ be a numerical invariant satisfying axioms A1, A2, A3, A4a, A4b of §4. Then for every $G \in \mathsf{SemComp}_{\mathrm{ex}}$,

$$I(G) = r(G) = \dim_{\mathbb{Q}} H^1(G).$$

In words: the witness rank is the unique non-negative integer-valued invariant on semantic interface complexes in the exact regime that is iso-invariant (A1), locally trivial on observable-only complexes (A2), additive over disjoint sums (A3), sensitive to non-trivial seam disclosures (A4a), and blind to cochain-trivial disclosures (A4b).

Axioms A5 and A6 are *consequences* of A1–A4b in the exact regime (Proposition 5.4); they are part of the doctrinal axiom system because they become non-redundant outside it.

5.3 Existence

The witness rank r satisfies the axioms. We verify each:

- **A1** (iso-invariance): $r(G)$ depends only on the isomorphism class of the cochain complex of G, which is iso-invariant.
- **A2** (local triviality): if $\mathcal{L} = 0$, then $C^n(G) = 0$ for all n, so $H^1(G) = 0$ and $r(G) = 0$.
- **A3** (additivity): Lemma 3.5.

- **A4a** (disclosure sensitivity): if $[\sigma] \neq 0$, $r(G \setminus\setminus W) = r(G) - 1$ by Lemma 3.7.
- **A4b** (coboundary blindness): if $[\sigma] = 0$, the cocycle representing W is a coboundary, contributing nothing to H^1, so $r(G \setminus\setminus W) = r(G)$.
- **A5** (cycle normalization): the worked example of §3.6 computes $r(C_e) = 1$.
- **A6** (excision): in the exact regime with compatible cover, the Mayer–Vietoris sequence is exact at H^1, giving $\dim H^1(G) = \dim H^1(U) + \dim H^1(V) - \dim H^1(U \cap V)$.

So r is one invariant satisfying all six axioms.

5.4 Uniqueness — the main argument

We show that any I satisfying A1, A2, A3, A4a, A4b equals r. The strategy is induction on a well-founded measure of complexity, with reduction steps that decrease the measure while preserving the relation $I(G) - r(G)$.

Definition 5.2 (Complexity measure). For $G \in \mathsf{SemComp}_{\mathrm{ex}}$, define the *latent complexity*

$$\mu(G) := \dim_{\mathbb{Q}} \mathcal{L}_{\mathrm{tot}}(G), \quad \text{where} \quad \mathcal{L}_{\mathrm{tot}}(G) := \bigoplus_{p \in P} \mathcal{L}(p).$$

The measure μ takes values in $\mathbb{Z}_{\geq 0}$. We prove $I(G) = r(G)$ by induction on $\mu(G)$.

Base case: $\mu(G) = 0$. Then $\mathcal{L} = 0$, and by A2,

$$I(G) = 0 = r(G).$$

Inductive step: $\mu(G) > 0$. Then $\mathcal{L} \neq 0$, so there exists a context p with $\mathcal{L}(p) \neq 0$. Two reduction tools, drawn from A3 and A4a/A4b, suffice:

(R1) Disjoint sum split. If G is disconnected — i.e., its cover poset P admits a non-trivial decomposition $P = P_1 \sqcup P_2$ closed under inheritance, with the convention data block-diagonal — then $G = G_1 \sqcup G_2$ with both non-empty. By A3 and Lemma 3.5,

$$I(G) = I(G_1) + I(G_2), \qquad r(G) = r(G_1) + r(G_2).$$

Both G_i have $\mu(G_i) < \mu(G)$. By induction $I(G_i) = r(G_i)$ for each, hence $I(G) = r(G)$.

(R2) Disclosure reduction. If G is connected, choose any p with $\mathcal{L}(p) \neq 0$ and any 1-dimensional subspace $W \subseteq \mathcal{F}(p)$ with $W \not\subseteq \mathcal{O}(p)$. Let $G' := G \setminus\setminus W$. The cohomology class $[\sigma] \in H^1(G)$ corresponding to W is computable from the cochain complex; either it is non-zero or it is a coboundary.

- If $[\sigma] \neq 0$ in $H^1(G)$: by A4a and Lemma 3.7, $I(G) - I(G') = 1 = r(G) - r(G')$.
- If $[\sigma] = 0$ in $H^1(G)$: by A4b and the cochain-level fact that coboundaries contribute nothing to cohomology, $I(G) = I(G')$ and $r(G) = r(G')$.

In either case, $\mu(G') = \mu(G) - 1$, since exactly one latent dimension is promoted to observable. By induction $I(G') = r(G')$, hence $I(G) = r(G)$.

Termination. The latent complexity μ is a non-negative integer; each reduction (R1 or R2) strictly decreases it; so the recursion terminates after finitely many steps, ending at the base case ($\mu = 0$) where A2 gives the conclusion.

This completes the proof. □

5.5 A5 and A6 are automatic in the exact regime

Notice that the proof of Theorem 5.1 used only A1, A2, A3, A4a, A4b. The base case invokes A2; the disjoint-sum reduction invokes A3; the disclosure reduction invokes A4a and A4b. Axioms A5 (cycle normalization) and A6 (excision under compatible cover) were not used in the argument.

Proposition 5.4 (A5 and A6 redundant in exact regime). *Any invariant* $I : \mathrm{Ob}(\mathsf{SemComp}_{\mathrm{ex}}) \to \mathbb{Z}_{\geq 0}$ *satisfying A1, A2, A3, A4a, A4b automatically satisfies A5 and A6.*

Proof. The reduction-and-induction argument of §5.4 uses only A1, A2, A3, A4a, A4b to show $I = r$ on $\mathsf{SemComp}_{\mathrm{ex}}$. By the existence calculation (§5.3), r satisfies all seven axioms; hence I satisfies A5 and A6. □

Specifically: A5 ($I(C_e) = 1$) is forced because C_e has $\mu(C_e) = 1$, the unique disclosure that reduces μ to zero corresponds to a non-trivial cohomology class, and A4a + A2 then give $I(C_e) = 0+1 = 1$. A6 (excision) is forced because the witness rank, which any A1–A4b-invariant must equal, satisfies excision via the standard Mayer–Vietoris calculation.

Two structural remarks. First, in the exact regime five axioms suffice — A1 (iso-invariance), A2 (local triviality), A3 (additivity), A4a (disclosure sensitivity), A4b (coboundary blindness) — to force the witness rank uniquely. The "−1" embedded in A4a is doing double duty: it specifies both the cancellation rule and the scale. A4b ensures the theory is not over-counted by redundant disclosures.

Second, A5 and A6 are *not* redundant outside the exact regime. When the cochain complex fails to split rationally — the surrogate regime — the disclosure reduction may not converge cleanly: an "independent" disclosure outside the exact regime can fail to reduce the obstruction by exactly one, requiring the cycle normalization (A5) and the excision rule (A6) to constrain the residual structure. Proposition 5.4 is a feature of the regime, not of the axioms.

We therefore retain A5 and A6 as part of the doctrinal axiom system. They express, respectively, the *normalization* property (one elementary obstruction equals one rank-unit) and the *locality* property (the obstruction of a whole is recoverable from its parts up to a defect term) that any *theory* of compositional obstruction must satisfy. In the exact regime they are automatic; outside, they are non-trivial constraints. The full axiom system describes what is required of the invariant in general — not only what is enforceable in the cleanest case.

A more detailed version of the uniqueness argument is given in `notes/proof-skeleton-exact.md`. An adversarial discussion addressing the "smuggling" objection — whether A4a/A4b implicitly contain the conclusion via cohomological independence — is given in `notes/anti-tautology.md`.

5.6 Discussion

The proof has the same shape as the uniqueness arguments in Eilenberg–Steenrod (homology), in the Tutte–Whitney characterization (matroid rank), and in Shannon (entropy): show existence, show the candidate satisfies the axioms, then show uniqueness by induction over a structural reduction in which the axioms force the invariant on smaller objects to determine its value on larger objects.

Three observations on the structure of the proof and the role of the individual axioms.

1. Where the originality lives. The proof itself is short — a finite induction with two reduction operations. The originality is not in the proof. It is in the choice of category $\mathsf{SemComp}$, in the identification of disclosure as the primitive operation, and in the claim that disclosure independence is the structural feature any reasonable theory of compositional obstruction must respect. Once those choices are accepted as the right primitives — which is Layer 1 of §5.1 — the numerical uniqueness (Layer 2) follows almost automatically. The theorem is structurally analogous to Shannon's: not because the proof is hard, but because the axioms identify the object.

2. Five axioms suffice in the exact regime. Proposition 5.4 establishes that A1 (iso-invariance), A2 (local triviality), A3 (additivity), A4a (disclosure sensitivity), A4b (coboundary blindness) — together — force the witness rank uniquely on $\mathsf{SemComp}_{\mathrm{ex}}$. A5 and A6 are automatic in this regime. This is the analogue of Milnor's observation, in algebraic topology, that the dimension axiom of Eilenberg–Steenrod can be replaced by additivity in suitable categories of spaces.

3. The doctrinal axiom system is non-minimal by design. In the exact regime, A5 and A6 are redundant. Outside, both become essential: A4a may not converge cleanly, and the residual structure requires the cycle-normalization axiom (A5) to fix the scale of an irreducible elementary obstruction and the excision axiom (A6) to control how obstruction decomposes under cover. Stating all six axioms describes what the obstruction theory must look like *in general*, including in regimes where the proof in §5.4 does not directly apply.

The same pattern holds in algebraic topology: the Eilenberg–Steenrod axioms include redundancies in particular categories (e.g., for finite CW-complexes, dimension is implied by exactness and additivity), but the full axiom system remains the right description of what a homology theory is, because the redundancies dissolve in larger categories (infinite spectra, non-CW spaces).

5.7 What this section achieves

Witness rank is the unique invariant on $\mathsf{SemComp}_{\mathrm{ex}}$ that respects seam-independent disclosure (Layer 1) — equivalently, the unique invariant satisfying axioms A1–A6 of §4 (Layer 2). The choice of $r := \dim H^1$ as the program's central diagnostic is not one option among many; it is forced by accepting disclosure independence as the primitive operation of compositional repair.

This is the result on which the doctrine of subsequent sections rests. The rest of the paper derives consequences: the minimum-repair invariant equals r (§6), coherence certificates have a disclosure normal form (§7), bounded-channel protocols cannot certify high-rank obstruction (§8), and the existing program — Bulla, the seam protocol, BABEL — supplies models of the abstract framework whose computational results confirm the theorem (§10).

---

§6. Repair Duality

We now establish that the minimum-repair invariant — the smallest number of independent disclosures needed to reduce the obstruction to zero — equals the witness rank. This identifies the *cost of certifying coherence* with the *amount of obstruction*: the dual sides of the same number.

6.1 The minimum-repair invariant

Definition 6.1 (Repair invariant). For $G \in \mathsf{SemComp}_{\mathrm{ex}}$, the *repair invariant* is

$$\begin{aligned}\rho(G) := \min\{k \in \mathbb{Z}_{\geq 0} : &\text{ there exist disclosures } \sigma_1, \dots, \sigma_k \\ &\text{on successive complexes such that} \\ &r(G \setminus\setminus W_1 \cdots \setminus\setminus W_k) = 0\}.\end{aligned}$$

Equivalently, $\rho(G)$ is the smallest number of disclosures needed to make the latent presheaf cohomologically trivial.

This is a non-negative integer invariant of G, depending only on the iso class of the cochain complex.

6.2 The duality theorem

Theorem 6.2 (Repair Duality). *For all $G \in \mathsf{SemComp}_{\mathrm{ex}}$,*

$$\rho(G) = r(G).$$

Proof. We show both inequalities.

Upper bound: $\rho(G) \leq r(G)$. By the reduction-and-induction argument of §5.3, any G can be

reduced to a base-case complex (with $\mathcal{L}$ cohomologically trivial) by exactly $r(G)$ non-trivial disclosures (sub-case D2b applied $r(G)$ times) plus arbitrarily many cochain-trivial disclosures (sub-case D2a) which preserve r. So a sequence of $r(G)$ non-trivial disclosures suffices.

Lower bound: $\rho(G) \geq r(G)$. By Lemma 3.7, each non-trivial disclosure reduces r by exactly 1; cochain-trivial disclosures preserve r. So k disclosures, in any combination, reduce r from $r(G)$ to at least $r(G) - k$. To reach $r = 0$ requires $k \geq r(G)$. □

6.3 Discussion

Repair duality says that the cost of certifying compositional coherence — measured in independent declarations that must be added to the observable interface — is exactly the obstruction count. The diagnostic invariant and the repair invariant are the same number; they are projections of one underlying cohomological quantity.

The duality has both descriptive and operational content. *Descriptively*: there is no certification scheme that achieves coherence with fewer than $r(G)$ disclosures, no matter how cleverly the disclosures are chosen. *Operationally*: any algorithm that computes $r(G)$ also gives a constructive route to a minimum-cost repair, by exhibiting a sequence of $r(G)$ non-trivial disclosures, each reducing the obstruction by one.

This second statement is implicit in §3.4: the witness Gram presentation of H^1 exhibits $r(G)$ basis cocycles, and disclosing each in turn produces a minimum-cardinality repair. The witness Gram is therefore not only a *diagnostic* tool — telling us how much obstruction is present — but also a *prescriptive* tool — telling us which disclosures to make. The two roles coincide because the underlying invariant is the same.

6.4 The forced/optimizable decomposition

Repair sequences of cardinality $r(G)$ are not unique: there are typically many bases of $H^1(G)$, each giving a different repair sequence. But the *cardinality* is fixed; what varies is the *choice* of which independent dimensions to disclose first.

This induces a natural decomposition of the repair calculus into two parts:

- **Forced.** The cardinality $r(G)$ — the *number* of disclosures required — is invariant across all repair strategies. This is the structural cost of coherence; it cannot be optimized away.
- **Optimizable.** The *choice* of disclosure basis — which $r(G)$ specific declarations to add — admits optimization. Different bases may have different practical costs (different fields are easier or harder to declare). The repair-invariant theorem fixes the cardinality but leaves the basis choice open.

This is the formal version of the "forced versus optimizable disclosure" structure observed in the existing program: the witness rank determines the floor of repair cost; below that, one optimizes over basis choices subject to a cardinality constraint. The two halves of the decomposition are

made precise here as: (i) cardinality is forced by Theorem 6.2; (ii) basis is optimizable subject to maintaining cardinality.

6.5 What this section achieves

Minimum-repair cardinality equals the witness rank. The diagnostic invariant and the repair invariant coincide. The cost of certifying coherence is the obstruction count; one disclosure per obstruction class, no fewer. This duality is a corollary of Theorem 5.1 — both sides follow from A1, A2, A3, A4a, A4b — and it makes the cost structure of compositional verification *forced by the same axioms* that fix the witness rank.

---

§7. Disclosure Normal Form

The repair calculus of §6 says that $r(G)$ disclosures suffice to bring a complex to a coherent state; the characterization theorem of §5 says that the witness rank is forced by the axioms. Together they yield a corollary about *certificates* of coherence: any proof that a composition is coherent normalizes to a particular shape — a *disclosure normal form* — which is small, replayable, and verifiable independent of the original argument. We call its operational realization a *receipt*.

This section makes that statement precise.

7.1 What a coherence certificate must contain

Suppose we have a proof that $G \in \mathsf{SemComp}_{\text{ex}}$ is coherent — that its obstruction vanishes, $r(G) = 0$. What must such a proof contain?

By the characterization theorem, $r(G) = 0$ is equivalent to: G admits no non-trivial cohomology class in H^1. Equivalently, every latent dimension is in $\operatorname{im}\delta^0$ (every potential obstruction is already a coboundary). Equivalently, the latent presheaf $\mathcal{L}$ is fully observable up to coboundaries.

But this last statement is rarely the *original* form of the data. Compositions arise with non-trivial $\mathcal{L}$; coherence is achieved by *adding disclosures* to bring the cohomology to zero. So a certificate of coherence for an originally-incoherent G_0 takes the form:

$$G_0 \setminus\setminus W_1 \setminus\setminus W_2 \cdots \setminus\setminus W_k = G_k, \quad r(G_k) = 0.$$

The certificate must therefore record:

(i) The *base complex* G_0 — the original interface data.
(ii) The *disclosure set* $\{W_1, \ldots, W_k\}$ — the declarations added to reach coherence.
(iii) The *vanishing-obstruction proof* — a witness, computable cochain-level, that $r(G_k) = 0$.
(iv) The *regime tag* — confirmation that $G_0 \in \mathsf{SemComp}_{\text{ex}}$, so the calculations are exact.

We call a tuple $\mathcal{R} = (G_0, \{W_i\}, \pi, \tau)$ satisfying (i)–(iv) a *receipt*.

7.2 The disclosure-normal-form theorem

Theorem 7.1 (Disclosure Normal Form). *Let $G_0 \in \mathsf{SemComp}_{\mathrm{ex}}$ be a semantic interface complex. Any proof that G_0 admits a coherent extension by some sequence of disclosures normalizes to a receipt $\mathcal{R} = (G_0, \{W_1, \dots, W_k\}, \pi, \tau)$ in which:*

- *(a) The disclosure set has cardinality $k = r(G_0)$ and is cochain-independent.*
- *(b) The vanishing-obstruction proof π is the linear-algebra certificate that $\operatorname{rank} \delta^0_{G_k} = \dim_{\mathbb{Q}} C^1(G_k)$, which is verifiable in $O(|C^1|^\omega)$ time using rational arithmetic (ω the matrix-multiplication exponent).*
- *(c) The regime tag τ is verifiable: a cochain-level check confirms that G_0 satisfies the exact-regime conditions of Definition 3.10.*

Two receipts $\mathcal{R}, \mathcal{R}'$ for the same base G_0 are equivalent if their disclosure sets are bases of the same subspace of $H^1(G_0)$.

Proof. By Theorem 5.1 and Theorem 6.2, any sequence of disclosures bringing r to zero must have non-trivial cardinality at least $r(G_0)$, and a cochain-independent sequence of cardinality exactly $r(G_0)$ suffices. So normalization to (a) is achieved by removing redundant or cochain-trivial disclosures from any longer sequence (each such removal does not increase the residual obstruction by Lemma 3.7).

For (b): once the disclosure set is fixed, the vanishing-obstruction proof reduces to a rank computation on the modified cochain complex, which is a polynomial-time linear-algebra problem over $\mathbb{Q}$.

For (c): the exact-regime conditions are checkable from the data of G_0 — operational sufficient conditions DFD + CHP (Definition 3.10) are direct structural properties of the convention presheaf.

The equivalence of receipts modulo the $H^1(G_0)$-basis ambiguity is the statement of §6.4: cardinality is forced, basis is optimizable. □

7.3 Compositionality of receipts

Receipts compose along compatible covers, by the standard Mayer–Vietoris exact sequence for the cochain complex.

Proposition 7.2 (Compositionality). *Let (U, V) be a compatible cover of $G_0 \in \mathsf{SemComp}_{\mathrm{ex}}$ in the exact regime. Let $\mathcal{R}_U, \mathcal{R}_V, \mathcal{R}_{U\cap V}$ be receipts for $U, V, U \cap V$ respectively. Then the disclosure set $\{W_i^U\} \cup \{W_j^V\} \backslash \{W_l^{U\cap V}\}$ — disclosures from U and V minus duplicates from the intersection — is a valid disclosure set for a receipt $\mathcal{R}_G$ for G_0, with cardinality $r(U)+r(V)-r(U\cap V) = r(G_0)$.*

Proof. Direct from A6 (or from its equivalent, Mayer–Vietoris for cochain cohomology). The cardinality formula is the inclusion-exclusion of the rank function, which is exactly axiom A6. □

This compositionality is the proof-theoretic counterpart of the cohomological locality recorded in A6: receipts can be assembled from receipts on covers, with redundancies subtracted.

7.4 Soundness

Proposition 7.3 (Soundness). *If $\mathcal{R} = (G_0, \{W_i\}, \pi, \tau)$ is a valid receipt — meaning π verifies and τ confirms the exact regime — then the disclosed complex $G_k = G_0 \setminus\setminus W_1 \cdots \setminus\setminus W_k$ has vanishing obstruction: $r(G_k) = 0$.*

Proof. Direct from the verifiable certificates: π verifies that the rank of the modified δ^0 matches $\dim C^1$, which by Definition 3.4 means $H^1(G_k) = 0$. □

In words: a receipt is *sound* if its certificates check; soundness implies that the certified composition has no hidden semantic obstruction relative to the axiomatized framework.

7.5 Completeness

The dual to soundness: every coherence-achievable complex admits a valid receipt.

Proposition 7.4 (Completeness). *For every $G_0 \in \mathsf{SemComp}_{\mathrm{ex}}$, there exists a valid receipt $\mathcal{R} = (G_0, \{W_1, \ldots, W_{r(G_0)}\}, \pi, \tau)$ certifying that the disclosed complex has vanishing obstruction.*

Proof. By the construction in the proof of Theorem 6.2 (upper bound), choose any basis $\{[\sigma_1], \ldots, [\sigma_{r(G_0)}]\}$ of $H^1(G_0)$, lift each to an underlying disclosure (p_i, W_i), and form the disclosure set $\{W_1, \ldots, W_{r(G_0)}\}$. By repeated application of A4a, $r(G_0 \setminus\setminus W_1 \cdots \setminus\setminus W_{r(G_0)}) = 0$. The verification certificate π is the rank computation; the regime tag τ is the verification of exact-regime conditions on G_0. Both are computable in polynomial time. □

Together Propositions 7.3 and 7.4 say: receipts form a *sound and complete* certificate system for compositional coherence in the exact regime. Soundness says verifying the certificate suffices; completeness says certificates exist whenever they should.

7.6 Disclosure normal form is the canonical certificate form

Theorem 7.1 says that *any* proof of compositional coherence — however originally presented — normalizes to a receipt of cardinality $r(G_0)$. The argument is:

1. The proof must, in the end, certify $r(G_k) = 0$ for some sequence of disclosures bringing r to zero.
2. The cardinality is at least $r(G_0)$ by Theorem 6.2.
3. Any disclosures beyond cardinality $r(G_0)$ are redundant — cochain-trivial or basis-equivalent — and can be removed.

4. The minimum-cardinality disclosure set is unique up to basis choice in $H^1(G_0)$.

So every proof of coherence reduces, after normalization, to a receipt: a base complex, a minimum-cardinality disclosure basis, a rank-verification certificate, and a regime tag. There is no proof of coherence that does not, after this reduction, take this shape.

This is the proof-theoretic content of the framework: *coherence proofs have a disclosure normal form.* Receipts are not one possible certificate format among many; they are the unique form to which any proof reduces.

7.7 What this section achieves

Coherence certificates have a normal form: an independent disclosure set of cardinality $r(G_0)$ together with a verifiable rank-vanishing certificate. Receipts realize this disclosure normal form operationally. The proof system is *sound* (Proposition 7.3): valid receipts certify zero obstruction. The proof system is *complete* (Proposition 7.4): every coherent extension admits a receipt. Receipts compose along covers (Proposition 7.2).

This identifies disclosure normal form as the proof-theoretic counterpart of the witness rank: just as the rank is the unique invariant respecting seam-independent disclosure (Theorem 5.1), the disclosure set is the unique certificate form to which proofs reduce (Theorem 7.1). The diagnostic invariant, the repair calculus, and the certificate format are three projections of one underlying object.

---

§8. Communication Lower Bound

A *protocol* in this section is an information-exchange procedure between participants holding the local data of an interface complex. The protocol exchanges messages — a transcript — and at the end emits a verdict on whether the composition is coherent. We ask: how much must the transcript carry?

The answer is forced by the characterization theorem and a pigeonhole argument.

8.1 The setting

Let $\mathcal{C}$ be a family of semantic interface complexes — for instance, all complexes built by composing tools from a fixed schema vocabulary in some bounded combinatorial structure. Let Π be a protocol that, given $G \in \mathcal{C}$ as input, produces a transcript $T(G)$ in some finite set $\mathcal{T}$ and a verdict $V(G) \in \{\text{coherent}, \text{incoherent}\}$.

We say Π is *sound on* $\mathcal{C}$ if for every $G \in \mathcal{C}$, $V(G) = \text{coherent}$ implies $r(G) = 0$.

We say Π is *complete on* $\mathcal{C}$ if for every $G \in \mathcal{C}$ with $r(G) = 0$, $V(G) = \text{coherent}$.

A protocol is *correct* if it is both sound and complete.

8.2 The lower bound

Theorem 8.1 (Communication Lower Bound). *Let $\mathcal{C}$ be a family of pairwise non-isomorphic semantic interface complexes such that:*

(i) $\mathcal{C}$ contains at least one complex with witness rank 0 (a coherent representative);

(ii) $\mathcal{C}$ contains at least one complex with witness rank > 0 (an incoherent representative);

(iii) $|\mathcal{C}| > N$ for some integer N.

Then any correct protocol Π producing transcripts in a space $\mathcal{T}$ of cardinality $|\mathcal{T}| \leq N$ fails to be correct on $\mathcal{C}$.

In particular, any correct protocol on $\mathcal{C}$ must produce transcripts with cardinality at least $|\mathcal{C}|$, so transmits at least $\log_2 |\mathcal{C}|$ bits in the worst case.

Proof. By pigeonhole, since $|\mathcal{C}| > N \geq |\mathcal{T}|$, there exist $G, G' \in \mathcal{C}$ with $G \ncong G'$ and $T(G) = T(G')$.

The verdict V depends only on the transcript: $V(G) = V(G')$.

Since $\mathcal{C}$ contains both coherent ($r = 0$) and incoherent ($r > 0$) elements, and since the family has cardinality strictly greater than N, we may choose the family large enough that some pigeonhole-collision pair (G, G') contains one coherent and one incoherent. (Formally: if every pigeonhole class is uniformly coherent or uniformly incoherent, then every transcript is unambiguously labeled, and we can refine $\mathcal{T}$ to a labeling map without ambiguity, contradicting $|\mathcal{T}| < |\mathcal{C}|$ where elements vary in coherence.)

For such a pair, V must give the right answer on both, but $V(G) = V(G')$ by transcript collision. So V is wrong on at least one. Π is not correct on $\mathcal{C}$. □

8.3 The obstruction-structure refinement

The bound of Theorem 8.1 is in terms of family cardinality. A finer bound expresses the cost in terms of the structural variation in obstruction class.

Corollary 8.2 (Obstruction-structure lower bound). *Let $\mathcal{C}$ be a family in which the obstruction modules $H^1(G)$ for $G \in \mathcal{C}$ collectively span a vector space of dimension D. Then any correct protocol on $\mathcal{C}$ transmits at least D bits in the worst case.*

Proof sketch. Construct a sub-family $\mathcal{C}' \subseteq \mathcal{C}$ realizing 2^D distinct subsets of obstruction classes — for instance, by varying which subset of a fixed D-dimensional obstruction space is "actively obstructed" in each member. Apply Theorem 8.1 to $\mathcal{C}'$ with $|\mathcal{C}'| = 2^D$. □

This sharpens the lower bound: the *number* of bits is the *dimension* of the obstruction variation, not just the family cardinality.

8.4 The fixed-structure case

Let $\mathcal{C}_*$ denote the family of complexes with a fixed underlying context poset and convention dimensions, where what varies is which dimensions are observable and which are latent. The witness rank ranges over $\{0, 1, \ldots, R\}$ for some R determined by the structure.

Corollary 8.3 (Fixed-structure lower bound). *Any correct protocol on $\mathcal{C}_*$ must produce transcripts whose underlying message space has cardinality at least 2^R, equivalently transmit at least R bits.*

Proof. Identical pigeonhole, with $|\mathcal{C}_*| \geq 2^R$ from the family construction. □

8.5 The semantic communication interpretation

The lower bound has a clean interpretation in the language of communication complexity.

Corollary 8.4 (Witness rank as a communication lower bound). *For any family of semantic interface complexes whose witness ranks range over $\{0, \ldots, R\}$, any sound protocol that detects all coherence violations must transmit at least R bits of semantic information about the obstruction structure.*

In other words: the witness rank lower-bounds the *semantic communication complexity* of distinguishing coherent from incoherent compositions in a family.

This is a clean formulation of the principle that "channel capacity must exceed obstruction count": a verifier whose protocol cannot transmit enough independent semantic declarations cannot soundly certify coherence over a family of complexes whose obstruction structure spans more than the channel allows.

8.5 Relation to distributed systems

The communication lower bound has formal kinship with classical lower bounds in distributed systems — for instance, with the $\Omega(\log n)$ bandwidth requirement for leader election under partial information, or with the bandwidth requirements for Byzantine agreement. We do not claim it is itself a distributed-systems result; it is a structural lower bound in a different setting.

The analogy is with FLP-style impossibility arguments: just as FLP shows that a particular kind of agreement is impossible without certain structural primitives (atomic commit, randomization, partial synchrony), Theorem 8.1 shows that a particular kind of certification is impossible without certain structural transmission primitives (witness-rank-many independent semantic declarations). The two are structurally analogous; whether the analogy can be tightened into a formal "semantic FLP" theorem with liveness conditions is left to subsequent work.

We emphasize: §8 is *not* a semantic FLP theorem. It is a clean pigeonhole communication bound. The richer FLP-style analysis with liveness, asynchrony, and fault tolerance is named as open in §9.

8.6 What this section achieves

The witness rank lower-bounds the communication complexity of correct semantic-coherence verification: any sound, complete protocol on a family of complexes whose obstruction structure spans R bits must transmit at least R bits of independent semantic information. The bound is structurally simple — pigeonhole — but its consequences are wide: any verifier with bounded message bandwidth fails to certify coherence on a sufficiently complex family. The cost of certification is, again, the witness rank; this time as a transmission requirement rather than a disclosure cardinality (§6) or a certificate length (§7).

---

§9. Boundaries

The doctrine established in §§4–8 holds in the exact regime $\mathsf{SemComp}_{\mathrm{ex}}$: complexes whose seam complex splits cleanly over the rationals. Outside this regime — and even within, in directions the present paper does not pursue — natural and important questions remain. We name the most central as conjectures, with the hope that some are provable, some are refinable, and some will turn out to require fundamentally new machinery.

9.1 Surrogate-regime characterization with explicit correction term

Conjecture 9.1 (Surrogate regime). *Let* $\mathsf{SemComp}_{\mathrm{sur}}$ *be the class of finite semantic interface complexes whose latent cochain complex is not rationally split — i.e., whose obstruction module exhibits non-trivial higher-cohomology contributions. There exists a numerical invariant* $\rho^* : \mathrm{Ob}(\mathsf{SemComp}_{\mathrm{sur}}) \to \mathbb{Z}_{\geq 0}$ *— the* surrogate witness rank *— and an explicit correction term* $\kappa(G) \geq 0$ *— the* coloop burden *— such that:*

(i) $\rho^*(G) = r(G) + \kappa(G)$ *for all* $G \in \mathsf{SemComp}_{\mathrm{sur}}$*, where* $r(G) = \dim H^1(G)$ *is the standard cohomological rank.*

(ii) ρ^* *is the unique invariant satisfying axioms A1–A4b plus a modified A5* (cycle normalization with multiplicities) *and a modified A6* (excision with explicit correction term κ).

(iii) κ *is computable from the cochain complex in polynomial time and bounds the gap between operational rank-counting and true cohomological rank.*

The conjecture asserts that the surrogate regime admits a parallel characterization theorem, but with two new axioms (modified A5 and A6) that become non-trivial outside the exact regime, and a new invariant κ tracking the excess of operational diagnostics over the true cohomological obstruction.

Status. Open. Partial results in the existing program (the leverage-conservation identity in [witness-geometry-beyond-fee], the additive decomposition in [hierarchical-fee]) suggest the structure is correct in special cases (DFD-violating compositions with bounded coloop count). Full characterization is a multi-quarter research program.

9.2 Verifier universality

Conjecture 9.2 (Verifier universality). *Let V be any verifier — i.e., a functor from $\mathsf{SemComp}_{\mathrm{ex}}$ to a category of certificates and verifications — that is sound and complete in the technical senses defined elsewhere. Then V factors through an equivalent of the witness functor: there exists a natural transformation from V to the witness-rank-cum-receipt construction, and this transformation is unique up to natural isomorphism.*

In words: any sound verifier of compositional coherence reduces, up to natural transformation, to one that computes the witness rank and emits the corresponding receipt.

Status. Open. The characterization theorem (§5) and the receipt normal form theorem (§7) suggest the result, but they do not directly prove it. A proof would likely use representation-theoretic or Yoneda-flavored arguments on the verification category.

This is a substantially harder result than the characterization theorem of §5. Theorem 5.1 says: among invariants satisfying A1–A6, witness rank is the unique one. Conjecture 9.2 says: among all verifiers (an a-priori broader category), witness-rank-style verification is essentially the only one. The two are related but not identical.

9.3 Semantic FLP-style impossibility

Conjecture 9.3 (Semantic FLP). *Consider a distributed protocol exchanging messages between heterogeneous components, each holding a piece of a semantic interface complex, in the presence of asynchrony and bounded fault tolerance. There exist regimes in which: (i) every protocol either fails to terminate (loss of liveness) or fails to certify semantic coherence (loss of safety); and (ii) the impossibility is parametrized by the witness rank in a way analogous to the impossibility for state agreement in FLP.*

Status. Open. The communication lower bound of §8 is a clean precursor — a pigeonhole result on transcript size — but does not yet engage with the asynchrony, fault, and termination conditions that distinguish FLP from a static lower bound.

A full semantic-FLP would relate witness-rank-bounded distributed protocols to the classical impossibility tradition in distributed systems. We do not pursue it here; we name it as the natural meeting point between the doctrine and the distributed-systems literature.

9.4 The eval/correctness gap as a clean theorem

Conjecture 9.4 (Eval-correctness inseparability). *There exists a parameterized family of semantic interface complexes $\{G_n\}_{n\in\mathbb{Z}_{\geq 0}}$ such that:*

(i) Every output-evaluation metric $E : \mathrm{Ob}(\mathsf{SemComp}) \to [0,1]$ that is locally computable (in a precisely-definable sense — depending only on bounded-radius local data) is asymptotically

uncorrelated with r:

$$\lim_{n\to\infty} \mathrm{Cor}(E(G_n), r(G_n)) = 0.$$

(ii) The family $\{G_n\}$ realizes arbitrary witness rank $r(G_n) \to \infty$ while every locally-computable eval metric remains bounded.

In words: there is no locally-computable evaluation metric that tracks the true compositional obstruction at scale.

Status. Empirically confirmed in the existing program (Coherence Cliff scaling experiment achieves $R^2 > 0.96$ for sheaf-cohomological diagnostics while best non-sheaf baselines degrade from $R^2 = 0.83$ at small scale to $R^2 = 0.52$ at large scale). The empirical pattern suggests the conjecture holds; the formal theorem requires precise specification of "locally computable" and the construction of an explicit family realizing the asymptotic separation.

This conjecture is the cleanest of the four for near-term formalization: the empirical confirmation is in hand, and the theoretical statement is a finite-dimensional combinatorial-asymptotic claim. We name it here as the natural follow-up paper to the doctrine.

9.5 Coherence-preserving interface evolution

Conjecture 9.5 (Coherence-preserving update). *Let G and G' be semantic interface complexes related by an interface-update map $f : G \to G'$ — for instance, a tool whose manifest is rewritten to add fields, rename conventions, or refine restrictions. If the induced map $f^\bullet : C^\bullet(G) \to C^\bullet(G')$ on seam cochain complexes is a chain homotopy equivalence, then witness rank, obstruction class, and disclosure-normal-form receipts are invariant: $r(G) = r(G')$, and any valid receipt for G transports to a valid receipt for G' via the explicit chain-homotopy data.*

In words: an interface update preserves all certificate validity exactly when the cochain-level update is homotopic to the identity. The conjecture identifies the precise condition under which *incremental* recertification is sufficient — versus the conditions under which a global re-derivation is required.

Status. Open. Mathematically the statement is a routine specialization of standard cochain-homotopy invariance; the work is in defining the right notion of *interface-update map* on SemComp and identifying the operational sufficient conditions (analogous to DFD + CHP for the exact regime) that practical tool updates can satisfy.

This is the most product-adjacent of the open problems. A proof would license a *coherence-preserving versioning discipline* in which manifest updates are accompanied by chain-homotopy data, and existing receipts remain valid without re-verification — solving the operational problem of incremental recertification under continuous interface evolution.

9.6 Discussion

The five conjectures of §9 are not equally hard. Conjecture 9.1 (surrogate-regime characterization) is the natural extension of the present work: a multi-quarter research program with concrete intermediate milestones. Conjecture 9.2 (verifier universality) is genuinely deeper: it requires proving that no fundamentally different verifier shape can soundly certify coherence — a stronger claim than uniqueness within an axiom system. Conjecture 9.3 (semantic FLP) is at the boundary with distributed systems: a full development requires formal apparatus the present paper does not provide. Conjecture 9.4 (eval-correctness inseparability) is the most empirically grounded and theoretically tractable. Conjecture 9.5 (coherence-preserving update) is the most product-adjacent: it directly addresses the operational problem of incremental recertification under interface evolution, and a proof would yield an immediate engineering discipline.

The doctrine paper does not require any of these to be true. Its main result — Theorem 5.1 — stands on its own in the exact regime. The conjectures describe the natural directions in which the doctrine generalizes; whether each direction yields a clean theorem is for subsequent work to decide.

9.7 What this section achieves

We have named five open conjectures, each natural and concrete, each tied to an existing piece of the program. Surrogate-regime characterization extends the doctrine outside its current scope. Verifier universality strengthens the characterization theorem. Semantic-FLP connects the doctrine to distributed-systems impossibility. Eval-correctness inseparability formalizes an empirically-confirmed phenomenon. Coherence-preserving update gives the operational discipline of incremental recertification a precise mathematical condition. The doctrine paper is positioned as the canonical reference for the exact regime; these conjectures define its frontier.

---

§10. Models of the Doctrine

The framework of §§2–8 is product-neutral by design: the category SemComp is a mathematical object, the witness rank is a cohomological invariant, and the receipt is a proof-theoretic normal form. The same machinery applies to any concrete instance of the abstract structure.

This section identifies four such instances — the *Bulla protocol*, the *seam manifest specification*, the *BABEL benchmark*, and the *Coherence Cliff scaling family* — and verifies that each is a model of the doctrine in a precise sense: the operational quantities computed in each instance correspond to the abstract invariants of the framework, and the empirical results in each instance are consistent with the theorems of §§5–8.

10.1 Bulla: a witness-receipt protocol

Bulla (Komkov 2026, [bulla]) is a Python implementation of a witness-receipt protocol for agentic compositions, with the MCP tool-use surface as its first operational target. A *Bulla composition* consists of a finite list of tools, each with a manifest declaring its observable schema and convention dimensions. From this data, Bulla constructs a coboundary operator over the field $\mathbb{Q}$ and computes the *coherence fee* — the rank of the Kron-reduced witness Gram — and a *witness receipt* recording the disclosed dimensions and the policy disposition.

Mapping to the framework. A Bulla composition G_{Bulla} corresponds to a semantic interface complex $G \in \mathsf{SemComp}_{\text{ex}}$ as follows:

Bulla artifact	Framework object
Tool list with manifests	Convention presheaf $\mathcal{F}$ on the context poset P
Declared (observable) field schema	Observable subpresheaf $\mathcal{O}$
Latent (undeclared) convention dimensions	Latent presheaf $\mathcal{L} = \mathcal{F}/\mathcal{O}$
Composition seam (the boundary between two tools)	Cover overlap in the context poset
Coherence fee	Witness rank $r(G) = \dim H^1(G)$
Witness Gram	Cochain complex coboundary δ^0 in matrix form
Bridge / disclosure	Disclosure morphism (§2.3 III)
Witness receipt	Receipt in the sense of Theorem 7.1
Operational exact regime (DFD + CHP)	$\mathsf{SemComp}_{\text{ex}}$ (§3.5)
Surrogate regime	$\mathsf{SemComp}_{\text{sur}}$ (Conjecture 9.1)

Under this correspondence, the coherence fee of a Bulla composition equals the witness rank of the corresponding semantic interface complex; the bridge calculus equals the repair calculus; the witness receipt equals the receipt normal form.

The empirical calibration of Bulla on 703 real schema compositions — coherence fee $= 0$ has zero false negatives, coherence fee ≥ 4 predicts mismatches with $\sim 100\%$ confidence — is consistent with the framework: the witness rank is the obstruction invariant; below threshold the composition is coherent (no obstruction to detect); above threshold the obstruction is detectable by independent means.

10.2 Seam manifest specification: a typed-interface model

The seam protocol (Komkov 2026, [seam]) specifies what gets published at composition time: the *seam manifest*, a typed declaration of observable schema and convention dimensions. The

manifest is signed and content-addressed; semantic witnesses are exchanged at composition; fraud proofs are constructed when local audits and global manifests disagree.

The seam manifest is the operational realization of the *interface specification* of a semantic interface complex: it specifies which conventions are observable and which are latent at each tool boundary. The publication of a seam manifest is the operational analog of the disclosure morphism (§2.3 III): both add structure to the observable subpresheaf.

Mapping. Each seam manifest field added to the observable schema corresponds to an independent disclosure W_i in the receipt of Theorem 7.1. The minimum seam-manifest publication required for coherence is $r(G_{\text{Bulla}})$ many fields, by Theorem 6.2. Operational metrics in the seam paper — the number of published declarations per fee point — are consistent with this duality.

10.3 BABEL: a benchmark for diagnostic correlation

BABEL (Komkov 2026, [babel]) is a 932-instance benchmark spanning seven families (calendar, invoicing, etc.) with deterministic ground truth (mean holonomy of a composition family). The benchmark measures correlation between diagnostics and ground-truth coherence: structural sheaf-cohomological diagnostics achieve Spearman $\rho = 0.99$; schema-validation diagnostics achieve $\rho = 0.17$; output-evaluation baselines fall in between.

Mapping. Each BABEL instance is a finite semantic interface complex with a designated "ground truth" coherence label (mean holonomy $= 0$ or > 0). Structural diagnostics compute the witness rank or a close proxy; their high correlation with ground truth is consistent with the framework: the witness rank is the obstruction invariant, and ground truth coherence corresponds to obstruction $= 0$.

Empirical content of Conjecture 9.4. BABEL provides an explicit family realizing the conjecture: locally-computable evaluation metrics (schema validation, GPT-4o judgment) achieve correlation ≤ 0.82, while the cohomological diagnostic achieves $\rho \geq 0.99$. The empirical separation is consistent with the conjecture's claim of asymptotic decoupling, though the conjecture's full formalization (§9.4) requires a precise definition of "locally computable."

10.4 Coherence Cliff: a scaling family

The Coherence Cliff scaling experiment (Komkov 2026, [coherence-cliff]) constructs 500 compositions across seven scales ($n = 5$ to $n = 50$ agents). At each scale, structural sheaf-cohomological diagnostics achieve $R^2 > 0.96$; non-sheaf baselines (Random Forest on topological features) degrade from $R^2 = 0.83$ at small scale to $R^2 = 0.52$ at large scale.

Mapping. Each scale corresponds to a family of semantic interface complexes with growing context-poset size and growing witness-rank distributions. The structural diagnostic (witness rank) tracks ground truth at all scales because it computes the unique obstruction invariant. Non-structural baselines degrade because they approximate the obstruction without comput-

ing it directly; as the obstruction structure becomes higher-dimensional with scale, the local approximations lose information that the cohomological diagnostic preserves.

This empirical pattern is consistent with Theorem 8.1: any baseline whose information capacity is sub-witness-rank cannot match witness-rank diagnostics on a scaling family.

10.5 Summary of the program — doctrine view

The existing program — papers, implementation, calibration, benchmark — is a coordinated investigation of one mathematical object: the witness-rank invariant on semantic interface complexes. Each component contributes a different projection:

Program component	Doctrine view
[scpi] paper	Formalization of the underlying descent / sheaf-cohomology framework; verification of structural lemmas in Lean.
[hierarchical-fee] paper	The additive decomposition (Lemma 3.5) and the witness-Gram presentation (Lemma 3.9).
[witness-geometry-beyond-fee] paper	The Kron reduction and leverage-conservation identity used in Definition 3.8 and §10.1.
[signed-incidence] paper	Field-independence of the witness rank between $\mathbb{Z}/2$ and $\mathbb{Q}$, used implicitly throughout.
[local-global-obstruction] paper	The impossibility of bounded local audits, equivalently the non-vanishing of H^1 on a constructed family.
[bulla] software	Computational realization: $r(G)$ as polynomial-time matrix rank over $\mathbb{Q}$; receipts as the proof normal form.
[seam] paper	Operational protocol: seam manifest as the typed interface declaration.
[coherence-cliff] paper	Empirical confirmation at scale: structural diagnostics dominate non-structural at all sizes.
[babel] benchmark	Empirical correlation: $\rho = 0.99$ for structural, $\rho = 0.17$ for baselines.
[bridge] paper	Empirical foundation: bilateral validity does not imply compositional coherence.

Each is a model of one or more aspects of the doctrine. None alone gives the full structure; together they describe a single mathematical object — the witness-rank invariant — from formal,

computational, operational, and empirical sides.

10.6 What this section achieves

The doctrine is not abstract. Each of its four components — characterization theorem, repair duality, receipt normal form, communication lower bound — has a concrete realization in the existing program, and the empirical results of those realizations are consistent with the theorems.

The relationship is *not* symmetric, and we do not claim that the program "validates" the doctrine. The existing program supplies examples, computations, calibrated benchmarks, and empirical tests of an abstract framework whose mathematical core was implicit in the work but never named as the unique invariant. The doctrine paper supplies that invariant — the obstruction quantity those artifacts were converging toward — and the axiomatic argument that no other invariant respects the structural primitive. Whether this is the *right* primitive is a modeling question (§1 scope table, layers 1–3) that this paper does not settle and that machine verification cannot adjudicate.

Formal Hinge

The formal center of the present program lies in Predicate Invention Under Sheaf Constraints.

Its decisive contribution is separation. The paper does not treat every failure of alignment as one undifferentiated haze of mismatch. It distinguishes topological obstruction, conservativity failure, and definability limits, and it places them in sequence. That sequence matters because it determines what kind of remedy is even coherent. A structural obstruction cannot be repaired by better prompting. A non-conservative extension cannot be repaired by renaming alone. A definability limit cannot be repaired by asserting that the concept was already available in the vocabulary. The SchemaDiscovery adjunction that makes SCPI computationally tractable extends Spivak's functorial data migration framework (2012): the invariant functor maps predicate sites to finite schemas, and the induced migration functors decompose the sheaf condition into a topological obstruction and a combinatorial migration — the same $\Sigma_f \dashv \Delta_f \dashv \Pi_f$ triple, applied to predicate invention rather than data movement.

What follows is the paper itself. The appendices later in this volume provide a proof-status ledger, a formalization map, and a selected inventory of the Lean work that bears the strongest weight for the present argument.

Predicate Invention Under Sheaf Constraints

Mathematical Foundations for Compositional Discovery

John Komkov*

March 9, 2026

Abstract

We study the problem of *predicate invention under sheaf constraints* (SCPI): given a Grothendieck site (C, J) equipped with a pseudofunctor of model groupoids $M\colon C^{\mathrm{op}} \to \mathbf{Grpd}$, when can a new predicate—specified locally on a cover and constrained by data—be extended to a global predicate that is (i) compatible with the local specifications (descent), (ii) conservative over the base signature, and (iii) explicitly definable from the base vocabulary?

We show that the obstruction to predicate invention decomposes into three distinct sequential components: a *topological obstruction* classified by non-abelian Čech cohomology $H^1(C, \mathrm{Equiv}_{\mathrm{ext}})$, a *model-theoretic obstruction* detecting non-conservativity via descent of the expansion property, and a *definability obstruction* measuring the gap between implicit and explicit definability over the site. Each obstruction is detected by different mathematical machinery and provides different diagnostic information. In particular, the overlap topology of information sources is a computable invariant that predicts whether concept alignment across autonomous agents is achievable—connecting classical sheaf theory to problems in AI interpretability and multi-agent coordination, under explicit modeling assumptions (M1–M4, Remark 1.4).

The extension torsor lemma and conservativity descent theorem are proved under Assumption D (effectivity of descent) and stated amalgamation hypotheses; Assumption D is itself proved for finite relational sites, the regime covering all worked examples. A full end-to-end computation of H^1 for a three-source site (Calendar/Email/Slack) demonstrates the complete diagnostic chain from site definition through obstruction identification to architectural prescription. A generalization of Beth's definability theorem to sites is proved unconditionally for geometric theories (assembling results of Kreisel, Johnstone, and the DD1 regime) and conditionally for general first-order theories. A topological invariance theorem shows that the obstruction landscape depends only on the Čech nerve and coefficient group, enabling cross-domain transfer of diagnostics between structurally isomorphic sites. We construct a schema discovery functor compatible with Spivak's functorial data migration adjunctions, with an explicit proof of compatibility for finite sites.

Selected results are formalized in Lean 4 (v4.24.0); see the companion repository for details.

*Selected Lean proofs verified with the assistance of Aristotle (Harmonic).

Contents

1 Introduction

1.1 The lifting problem

Every system that must extract structured knowledge from unstructured sources confronts the same task: a query implies a schema that does not yet exist [3], and the schema must be *discovered* from the source material before it can be populated. The schema is not retrieved. It is not pre-configured. It is *invented*—and then it must be validated against the source under compositional constraints.

We develop three examples at escalating difficulty. These are not toy problems; they recur as running examples throughout the paper, and every theorem is instantiated against at least one of them.

Example 1.1 (Static corpus)**.** "How many types of fruit are mentioned in the Bible and how many times is each type mentioned?" The implied predicates `is_fruit`(x) and `mentioned_at`$(x, \text{book}, \text{chapter}, \text{verse})$ exist in no index. The source decomposes

into 66 books, each into chapters, each into verses—a hierarchical cover. The predicate `is_fruit` must be invented locally in each book and then verified for global coherence: "pomegranate" in Song of Solomon and "pomegranate" in Exodus must refer to the same entity, classified the same way.

Example 1.2 (Dynamic corpus)**.** "How many meetings with external parties happened in my organization last week that involved more than 2 internal attendees?" The source material comprises emails, calendar invites, Slack messages, and Zoom logs—overlapping views of the same underlying events. A meeting may appear as a calendar entry, a confirmation email, a Slack thread, and a Zoom recording. The views overlap but do not partition. The predicate `is_meeting`(e) must be reconciled across fundamentally different ontologies, with entity resolution (e.g., `john@co.com` = "John K" = `@jkomkov`) that must compose consistently around the triangle of pairwise overlaps.

Example 1.3 (Open corpus)**.** "Find me hotel options near Berkeley for a Cal home game this fall, subject to constraints $C_1, \ldots, C_n$." The source is the open web, which cannot be scanned exhaustively. The agent must simultaneously choose which sites to crawl (the cover), invent the join schema, perform extraction, and verify coherence—all under a budget that makes the optimal cover problem NP-hard (Theorem 7.2).

1.2 Why this matters beyond logic: interpretability and agent coordination

The SCPI framework addresses a problem of growing interest in AI systems research: *how do autonomous agents, each with partial views of a domain, arrive at shared concepts?*

Interpretability. A neural network trained on medical images may "discover" an internal feature that correlates with disease severity—but this feature exists only within the network's latent space. A second network, trained on patient records, may discover a related but distinct feature. Are these the same concept? Can they be reconciled? This is precisely a predicate invention problem: each network invents a local predicate (its internal feature), and the question is whether these local predicates glue to a coherent global concept. The obstruction theory developed here provides a rigorous diagnostic:

- If the topological obstruction (H^1) is non-trivial, the local features are *fundamentally incompatible*—no choice of alignment can make them agree globally. This is a structural impossibility, not a training failure.
- If the model-theoretic obstruction is non-trivial, the features glue but the combined concept introduces new logical consequences not present in either network's individual "theory." The merged representation is not conservative.
- If the definability obstruction is non-trivial, the concept is "there" (implicit in the combined data) but cannot be expressed in the shared vocabulary. The agents agree on what they see but lack the language to say it.

Each failure mode has a different fix. The framework tells you *which* fix to apply—or whether no fix exists.

Agent–agent coordination. When multiple AI agents share information to complete a task (e.g., a retrieval agent, a reasoning agent, and a verification agent), they must agree on the meaning of intermediate predicates. The structure of their information overlap determines whether agreement is possible. A hierarchical architecture (tree-structured information flow) has contractible nerve—under locally constant coefficients and Assumption D, $H^1 = 0$, so local agreements always globalize (Corollary 3.5). A peer-to-peer architecture (circular information flow) can have non-trivial H^1, creating irreconcilable disagreements that no amount of "negotiation" can resolve without changing the cell structure of the nerve (e.g., adding higher-dimensional overlaps).

A concrete interpretability instance. Consider two models: a vision model V and a text model T, probed on a shared evaluation set E of medical cases. The "cover" is $\{V, T\}$; the overlap $V \cap T$ consists of cases in E where both models produce confident predictions. Each model invents a local predicate ("high-severity") on its domain. On the overlap, the two predicates may agree or disagree. With two contexts and one overlap, the nerve is Δ^1 (contractible)—so $H^1 = 0$ and alignment is always achievable by local adjustment. Now add a third model G (genomic risk). The cover is $\{V, T, G\}$ with pairwise overlaps (cases where two models are confident) but possibly no effective triple overlap (no case where all three are simultaneously confident). The nerve may be S^1, and the SCPI diagnostic applies: compute the cocycle, check if it's a coboundary. If not, no alignment is possible without adding a shared evaluation set that creates a triple overlap. This is a testable prediction.

Diagnostic summary. The overlap topology of your information sources is a *computable invariant* that predicts whether concept alignment is achievable. Compute the Čech nerve of your agent architecture. If it's contractible, the coefficient sheaf is locally constant, and Assumption D holds (as proved for finite relational sites in Lemma 2.19), then Gate 1 cannot fail: $H^1 = 0$ (Corollary 3.5). Any obstruction must be downstream—a conservativity or definability issue, not a topological impossibility. If the nerve is *not* contractible, the non-trivial H^1 class tells you exactly where the obstruction lives and what structural change (adding a shared data source, restructuring the agent graph) would eliminate it.

Empirical instantiation. The companion paper [43] empirically instantiates the three-gate sequence in an abelian linearized regime: three LLM agents operating against three database schemas on a coordination graph with one independent cycle. Gate 1 is measured directly ($\dim H^1 = 2$, two independent blind-spot dimensions invisible to bilateral validation). A closed-loop repair experiment demonstrates a failure straddling Gates 2 and 3: frontier LLMs correctly identify the missing bridge predicates (passing Gate 1) but propose specifications that canonize one agent's existing local policy rather than introducing a witnessable shared predicate—a non-conservative extension (Gate 2) expressed in natural-language ambiguity rather than explicit formulas (Gate 3). The minimum number of 2-cells required to kill H^1—the *coherence fee* of [43]—is a computable topological invariant; in the abelian regime, it equals $\dim H^1$ of the interpretation sheaf on the coordination graph. In the non-abelian regime, the obstruction is a pointed set and the fee is the minimum number of nerve corrections required to trivialize it. The SHEAF protocol [44] builds the diagnostic instrument and economic mechanism for pricing and procuring the minimum repair.

Remark 1.4 (Modeling assumptions for AI applications). The formal framework applies unconditionally to any setting that instantiates a predicate site (C, J, M). The data-integration instances (Examples 1.1–1.3; the Calendar/Email/Slack computation of Section 3.1) are fully formal: finite poset sites with relational signatures. The neural-network interpretability and agent-coordination applications discussed above require four modeling assumptions that are each nontrivial and currently at different levels of empirical support.

(M1) Features as predicates. Internal neural features must approximate logical predicates on a shared domain. Sparse autoencoders [32, 33] find that $\sim$70% of extracted features are monosemantic and behave like unary predicates, but the remaining features exhibit multi-dimensional structure—circular representations of periodic concepts [34]—or context-dependent polysemanticity. Moreover, SAE reconstructions recover a fraction of total model compute [35], and different training runs yield different decompositions. The predicate approximation holds in the *monosemantic regime*; multi-dimensional features are better modeled as sections of fiber bundles, suggesting that sheaf theory may be the right formalism precisely where the predicate assumption is weakest.

(M2) Restriction maps with algebraic structure. Cross-model representation comparison must produce morphisms that compose, not merely scalar similarity scores (CKA, probing accuracy). Model stitching [37] and universal sparse autoencoders [38] provide maps between representation spaces with the right structural properties, but *composition has not been verified* and the sheaf gluing axiom has never been checked for cross-model maps. This is the most demanding assumption. Seely et al. [39] have computed sheaf cohomology *within* single predictive-coding networks, finding that non-vanishing cohomology characterizes irreducible inference errors—establishing operational meaning for sheaf-cohomological obstructions in neural settings—but the cross-model case remains open.

(M3) Identifiable symmetry groups. The coefficient group $\mathrm{Equiv}_{\mathrm{ext}}$ must be computable. Identifiability theory provides precise guarantees: under sparsity conditions, latent variables are recoverable up to permutation and component-wise transformation [40], and neuron permutation symmetries are completely characterized for standard architectures. Feature splitting—where decompositions change qualitatively with dictionary width—requires categorical structure (spans, filtered colimits) beyond standard group actions.

(M4) Site structure for agent systems. Multi-agent coordination must admit a Grothendieck topology. Schmid [41] articulates this research program for multi-agent reinforcement learning; Gavranović et al. [42] provide the categorical vocabulary (representing architectures as monad algebra homomorphisms). No published work has yet computed sheaf-cohomological obstructions to coordination in empirical multi-agent neural systems.

These assumptions should be read as *interface specifications*: the paper's diagnostic machinery applies whenever M1–M4 are instantiated, and the assumptions identify precisely what empirical validation is required. The Platonic Representation Hypothesis [36]—that model representations converge with scale—suggests that M2 may become easier to satisfy as models grow, while the non-trivial cohomology regime that motivates SCPI is precisely the current one where alignment is partial. The contribution of the present paper to the AI-applications context is not to validate M1–M4 but to identify them sharply and to show that *if* they hold, the diagnostic decomposition is a theorem.

1.3 The common structure

These problems share a mathematical structure independent of the source medium:

(1) A *query* implies a *schema* that does not yet exist.

(2) The schema must be *discovered* (predicate invention [9]).

(3) The source decomposes into *overlapping contexts* (a cover of a site).

(4) Discovery is applied *locally* in each context.

(5) Local extractions must *cohere* globally (the sheaf condition).

(6) Extraction has a *cost* that must be bounded.

The thesis formalizes this structure.

1.4 Contributions and honest delineation

We summarize the main results with explicit labels indicating their status:

[Spine] Extension Torsor Lemma (Lemma 3.1): the obstruction to gluing local extensions is a class in $H^1(C, \mathrm{Equiv}_{\mathrm{ext}})$.

[Spine] No-Go for fixed-topology alignment (Corollary 3.2): when $H^1 \neq 0$, no protocol operating within the fixed overlap structure can achieve global concept alignment.

[Spine] Assumption D for finite relational sites (Lemma 2.19): effectivity of descent is proved (not merely assumed) for the regime covering all examples in this paper.

[Spine] Conservativity Descent (Theorem 4.3): conservativity is local under a model-amalgamation hypothesis, with a finite counterexample when the hypothesis fails.

[Spine] Obstruction Decomposition (Theorem 5.1): in World B, three distinct sequential obstructions exhaust the failure modes.

[Spine] — **sketch** Schema Discovery compatibility (Proposition 8.3): Inv extends Spivak's data migration adjunctions, with construction and proof sketch for finite sites.

[Spine] Full worked computation (Section 3.1): H^1 computed end-to-end for the Calendar/Email/Slack site, diagnosing both resolvable and irreconcilable cases.

[Spine] Beth for Geometric Sites (Theorem 6.2): implicit definability implies explicit definability over sites with geometric theories—proved unconditionally (assembling Kreisel, Johnstone, and DD1).

[Conditional] Beth for Sites, general (Theorem 6.4): extends to non-geometric theories under H1/H3.

[Conjectural] Schema Discovery universality (Conjecture 8.4): Inv is the universal schema discovery functor.

Hypothesis regime. The following table summarizes which hypotheses each spine result requires:

Result	Assump. D	Thin fibers	DD1/DD2	Model-cons.
Torsor Lemma (3.1)	yes	—	yes	—
Assump. D for fin. rel. (2.19)	*proved*	—	—	—
Conserv. Descent (4.3)	yes	yes	—	yes (local)
Obstruction Decomp. (5.1)	yes	—	yes	—
Independence (5.3)	—	—	—	—
Schema Compat. (8.3)	—	—	—	—
Beth-geom (6.2)	—	—	DD1 (auto)	—
Beth-general (6.4)	—	—	H1/H3	—

Reader's guide. Readers primarily interested in the AI applications may begin with the worked example (Section 3.1) and the topological invariance theorem (Theorem 5.6), referring back to Section 2 as needed. Readers interested in the model-theoretic foundations should read linearly from Section 2 through Section 6.

Paper map. Sections 2–5 constitute the mathematical core (definitions, torsor lemma, conservativity descent, diagnostic decomposition). Section 6 extends Beth's theorem to sites. Sections 7–8 develop computational and applied consequences. Section 9 describes formalization.

Positioning. *Not new*: interpreting theories in sheaf toposes (Caramello [2]; Johnstone [7]; Makkai–Reyes [8]), functorial data migration (Spivak [10]), effective descent for pretoposes via definability (Makkai [17]; Zawadowski [18]; Ballard–Boshuck [16]). *New*: treating predicate invention as a descent problem for extensions of a model stack *over a specific Grothendieck site*, with a computable diagnostic decomposition (the three-obstruction theorem), a bridge to schema discovery, and applications to AI interpretability and multi-agent coordination (Section 1.2). The architecture of the paper—descent = definability + covering—is validated by Ballard–Boshuck [16], who prove that categorical descent theorems for pretoposes decompose into a Beth/Tarski definability component plus a covering theorem. Our contribution is to instantiate this decomposition on concrete Grothendieck sites with a computational diagnostic.

2 Objects: Sites, Model Stacks, Extension Stacks

2.1 Predicate sites

Definition 2.1 (Theory)**.** A *first-order theory* $T = (\Sigma, \mathrm{Ax})$ consists of a signature $\Sigma = (S, F, R)$—sorts, function symbols, relation symbols—and a set of axioms consisting of sentences over Σ.

Definition 2.2 (Model stack — Frozen Definition 1)**.** Let (C, J) be a Grothendieck site. A *model stack* over (C, J) [6] is a pseudofunctor

$$M \colon C^{\mathrm{op}} \to \mathbf{Grpd}$$

assigning to each object $U \in C$ a groupoid $M(U)$ of Σ-structures (models), with restriction functors $M(f)\colon M(U) \to M(V)$ for each morphism $f\colon V \to U$, satisfying pseudofunctor coherence: $M(g \circ f) \cong M(g) \circ M(f)$ via coherent natural isomorphisms.

The *stack condition* (effective descent for models) is an explicit hypothesis when needed, not assumed globally.

Definition 2.3 (Predicate site)**.** A *predicate site* is a triple (C, J, M) where (C, J) is a Grothendieck site and $M\colon C^{\mathrm{op}} \to \mathbf{Grpd}$ is a model stack.

Remark 2.4 (Why models, not theories). The primitive is M (models/structures), not a presheaf of theories. Sheaf conditions are about sections; in logic, the natural "sections" are models. Descent for models (amalgamation) is well-posed; descent for axiom sets is not without extra machinery. The syntactic counterpart $\mathrm{Th}(U)$ = common theory of $M(U)$ is derived.

Remark 2.5 (Strictification). M is a *pseudofunctor*: restrictions compose up to coherent natural isomorphism, $M(g \circ f) \cong M(g) \circ M(f)$. In the Lean formalization, we work with a strict functor (equality, not isomorphism).

What is strictified: the composition law for restriction functors ($M(g \circ f) = M(g) \circ M(f)$, on the nose) and the identity law ($M(\mathrm{id}_U) = \mathrm{id}_{M(U)}$). This is justified by Mac Lane's coherence theorem for bicategories: every pseudofunctor from a small category to **Grpd** is equivalent to a strict 2-functor.

What is not claimed: we do not claim that the strictification preserves all 2-categorical structure. In particular, we do not strictify the coefficient groupoid $\mathrm{Equiv}_{\mathrm{ext}}$ (which remains a 1-group in the spine), nor do we strictify the descent data (which is isolated in Assumption D). The theorems are invariant under this equivalence because they depend on the fibers $M(U)$ and the restriction maps only up to natural isomorphism, and the cohomological classification (H^1) is invariant under equivalence of coefficient objects.

Scope: this strictification is adequate for finite covers with group-valued coefficients (the spine regime). Extending to infinite sites or groupoid-valued coefficients may require genuine bicategorical or $(\infty, 1)$-categorical descent, which is beyond the scope of this paper.

2.2 Extension objects — World B

Definition 2.6 (Extension — Frozen Definition 2)**.** An *extension* of M over context U is a triple (m, q, D) where:

- m is an object (or family of objects) in $M(U)$,
- q is an interpreted new *unary* predicate symbol on a specified sort of m—a subobject of the domain, *not defined by a formula in the base signature* Σ,
- D is a constraint package: labeled examples, subobject classifiers, or specifications in the internal logic of the fiber.

The groupoid $\mathrm{Ext}(M)(U)$ has objects = extension triples, morphisms = equivalences of extensions (definable bijections preserving q and compatible with D).

Restriction to unary predicates: for clarity, this paper treats only unary predicates on a single sort. The generalization to n-ary predicates is routine (replace subobjects of a sort by subobjects of a product of sorts) and does not affect any of the cohomological or model-theoretic arguments.

Warning 2.7 (World B). The predicate q is a *new primitive constrained by data*, not a definitional extension by a formula. In World A (definitional extensions), definability implies conservativity, collapsing the three-obstruction decomposition. The paper operates exclusively in World B, where this collapse does not occur.

Example 2.8 (Concrete constraint package)**.** In Problem B (Example 1.2), an extension of $M(\mathsf{Cal})$ by the predicate `is_meeting` might have constraint package D consisting of:

- *Positive examples*: $P = \{e_{17}, e_{42}, e_{91}\}$ (calendar events known to be meetings).
- *Negative examples*: $N = \{e_3, e_{55}\}$ (events known *not* to be meetings).
- *Closure axiom*: $\forall e.\, \texttt{is\_meeting}(e) \Rightarrow \texttt{attendees}(e) \geq 2$.

The constraint package restricts which subobjects $q \subseteq |m|$ are admissible extensions: q must contain P, exclude N, and satisfy the closure axiom. Different admissible q's are related by the equivalences in $\mathrm{Equiv}_{\mathrm{ext}}(\mathsf{Cal})$.

Definition 2.9 (Extension stack — Frozen Definition 2 continued)**.** The *extension stack* $\mathrm{Ext}(M)\colon C^{\mathrm{op}} \to \mathbf{Grpd}$ is a prestack but *not necessarily a stack*: local extensions satisfying the matching condition may fail to glue. The failure of $\mathrm{Ext}(M)$ to satisfy the stack condition is precisely the obstruction to predicate invention.

2.3 Three notions of agreement

Definition 2.10 (Agreement levels)**.** Local extensions q_i, q_j on overlapping contexts U_i, U_j can agree on $U_i \cap U_j$ in three senses:

1. *Strict*: $q_i|_{\mathrm{overlap}} = q_j|_{\mathrm{overlap}}$.
2. *Up to definable bijection*: $\exists\, d_{ij}$ definable with $q_j(d_{ij}(x)) \Leftrightarrow q_i(x)$.
3. *Up to automorphism*: $\exists\, \alpha_{ij} \in \mathrm{Aut}(M(U_i \cap U_j))$ with $q_j = q_i \circ \alpha_{ij}$.

These produce three different gluing problems with coefficient sheaves of increasing complexity. In this paper, we analyze **Level 2** (agreement up to definable equivalence) in the spine theorems; Levels 1 and 3 are variants with simpler/more complex coefficient objects respectively.

2.4 Coefficient equivalences

Definition 2.11 (Coefficient equivalences — Frozen Definition 3)**.** $\mathrm{Equiv}_{\mathrm{ext}}$ assigns to each U the group of definable equivalences of extensions over $M(U)$. In the H^1 regime (main results): $\mathrm{Equiv}_{\mathrm{ext}}(U)$ is a group. In general: a groupoid, with H^2/gerbe obstructions.

We require that $\mathrm{Equiv}_{\mathrm{ext}}$ satisfies the *sheaf condition* on (C, J). This is **not automatic**: "definable" equivalences need not glue across covers in arbitrary first-order settings (definability is not inherently local).

We therefore assume one of:

(DD1) *Geometric regime*: the base theories $\mathrm{Th}(U)$ are geometric theories and definability is interpreted in the internal logic of the topos $\mathrm{Sh}(C, J)$, where it is local by construction.

(DD2) *Definability descent*: definable maps satisfy the matching and gluing conditions across covers (an explicit hypothesis on the site, verifiable in concrete cases).

In either case, $\mathrm{Equiv}_{\mathrm{ext}}$ is a sheaf of groups on (C, J) and the H^1 classification is well-posed.

Remark 2.12 (Computability of $\mathrm{Equiv}_{\mathrm{ext}}$). In finite relational settings, $\mathrm{Equiv}_{\mathrm{ext}}(U)$ is a finite group computable by enumeration: list all definable bijections of the domain of m that preserve q and are compatible with D, check which are equivalences. For n elements and k constraints, this is bounded by $n! \cdot 2^{O(k)}$. In geometric settings, $\mathrm{Equiv}_{\mathrm{ext}}(U)$ is computed in the internal logic of $\mathrm{Sh}(C, J)$ using the definability theorem. In both cases, the group is concretely identifiable—$\mathrm{Equiv}_{\mathrm{ext}}$ is not a mystical object but a finite (or at worst finitely-presented) group with explicit generators.

Remark 2.13 (When DD holds automatically). DD1 holds for coherent/geometric theories in Grothendieck toposes: geometric sentences are preserved by inverse image functors of geometric morphisms [7], so truth of geometric formulas is inherently local (checkable on covers and gluable). DD2 holds for finite relational sites where definability reduces to finitely many quantifier-free conditions. In the examples of this paper, one of these conditions is always satisfied.

Remark 2.14 (Categorical precedent for definability descent). The requirement that $\mathrm{Equiv}_{\mathrm{ext}}$ be a sheaf is not without precedent. Ballard and Boshuck [16] prove that conservative morphisms are effective descent morphisms in the 2-category of pretoposes, and that these descent theorems decompose into "a familiar Beth/Tarski-type definability theorem and a covering theorem." Their "locally definable sets equipped with compatible actions that admit global representations" is precisely our setup, repackaged as a sheaf condition on a specific Grothendieck site rather than in the abstract 2-category of pretoposes. Zawadowski [18] extends this to the non-Boolean (intuitionistic) case, and Makkai [17] provides the original Stone-duality proof for Boolean pretoposes.

2.5 The SCPI decision predicate

Definition 2.15 (SCPI — Frozen Definition 4)**.** Given a predicate site (C, J, M), a cover $\{U_i \to U\}$, and local extensions $\{(m_i, q_i, D_i)\} \in \mathrm{Ext}(M)(U_i)$:

SCPI holds iff there exists a global extension $(m, q, D) \in \mathrm{Ext}(M)(U)$ such that:

1. **Descent**: $(m, q, D)|_{U_i} \cong (m_i, q_i, D_i)$ in $\mathrm{Ext}(M)(U_i)$ for each i.
2. **Conservativity**: the extension is conservative in the sense of Definition 2.16 below.
3. **Definability** (optional, stronger): q is explicitly definable from the base vocabulary.

2.6 Two notions of conservativity

Definition 2.16 (Conservativity)**.** Let $M' \to M$ be an extension of model stacks (i.e., a natural transformation of pseudofunctors with $M'(U) \to M(U)$ a forgetful functor for each U). We distinguish two notions [11, 13]:

1. *Deductive conservativity* (theory inclusion): $\mathrm{Th}_\Sigma(M'(U)) \subseteq \mathrm{Th}_\Sigma(M(U))$—every Σ-sentence true in all extended models was already true in all base models. This is the standard default meaning in mathematical logic.

2. *Model-theoretic conservativity* (expansion property): the forgetful functor $U\colon M'(U) \to M(U)$ is *essentially surjective on objects*—every base model $A \in M(U)$ admits an expansion $A' \in M'(U)$ with $U(A') = A$ satisfying the extension's constraints.

Model-theoretic conservativity implies deductive conservativity (if every base model expands, then the extended theory cannot exclude any base model). The converse fails in general: theory inclusion does not guarantee that a *given* base model can be expanded [14].

In the spine theorems (Theorem 4.3), we use **model-theoretic conservativity**, which is the notion required for the descent proof to work. The definition of SCPI (Definition 2.15) can use either notion; our results apply to the stronger (model-theoretically conservative) version.

Remark 2.17 (When the two notions coincide). For decidable universal theories over finite structures, deductive and model-theoretic conservativity coincide (compactness + finite witness). For geometric theories interpreted in a topos, model-theoretic conservativity is the natural notion (essential surjectivity of the reduct functor). For extensions by explicit definitions, both notions hold automatically [15]. In general, the distinction matters and must be tracked; see Rabe [13] for a systematic treatment.

2.7 The Descent Axiom

Assumption 2.18 (Descent Axiom — Assumption D)**.** The extension stack $\mathrm{Ext}(M)$ satisfies *effectivity of descent*: if the Čech 1-cocycle associated to a family of local extensions is a coboundary, then there exists a global extension that restricts to each local extension (up to equivalence in the fiber groupoid).

Formally: $[\alpha] = 0$ in $H^1(C, \mathrm{Equiv}_{\mathrm{ext}})$ implies $\exists\, e \in \mathrm{Ext}(M)(U)$ with $e|_{U_i} \cong e_i$ for all i.

Minimality: This axiom has one clause (effectivity). Separation (uniqueness up to isomorphism) is a stronger condition required for the full stack property; it is not needed for the spine theorems and is therefore not assumed here. If separation is needed in future extensions, it should be added as a separate hypothesis.

All stacky subtleties are isolated into Assumption D. The spine theorems (Lemma 3.1, Theorem 4.3, Theorem 5.1) assume D. Without Assumption D, one can prove theorems about formal cocycles in groups without ever linking them to "there exists a global extension." Assumption D is the bridge from cohomological algebra to geometric existence.

Lemma 2.19 (Assumption D for finite relational sites — [Spine])**.** *Let (C, J) be a finite poset category with covering topology, and let $M\colon C^{\mathrm{op}} \to \mathbf{Grpd}$ assign to each U a groupoid of finite structures in a relational signature Σ (no function symbols). Suppose the restriction functors are given by substructure inclusion. Then Assumption D holds: effectivity of descent is satisfied.*

Proof. We must show: if $\{(m_i, q_i, D_i)\}_{i\in I}$ are local extensions over a cover $\{U_i \to U\}$ whose Čech cocycle $[\alpha] = 0$, then a global extension exists.

Since $[\alpha] = 0$, after adjusting by a coboundary we may assume $\alpha_{ij} = \mathrm{id}$ for all i, j. That is, the local predicates agree strictly on all pairwise overlaps: $q_i|_{U_i\cap U_j} = q_j|_{U_i\cap U_j}$.

Construction of the global extension. The global model $m = M(U)$ is a finite relational structure. Each $m_i = m|_{U_i}$ is a substructure. Define $q\colon |m| \to \{0, 1\}$ by:

$$q(a) = q_i(a) \quad \text{for any } i \text{ with } a \in |m_i|.$$

Well-definedness: if $a \in |m_i| \cap |m_j|$, then $a \in |m|_{U_i \cap U_j}|$, and $q_i(a) = q_j(a)$ by strict agreement. Since $\{U_i\}$ is a cover, every $a \in |m|$ belongs to some $|m_i|$, so q is totally defined.

Constraint satisfaction: each constraint package D_i refers to $q|_{U_i} = q_i$, which holds by construction. For relational signatures, no function-symbol compatibility is needed; all relations restrict by substructure inclusion.

Uniqueness (up to equivalence): any two global extensions agreeing locally are equal on elements (by well-definedness), hence isomorphic. □

Remark 2.20 (Scope and extensions of Lemma 2.19). The lemma uses three properties of finite relational sites: (1) structures are finite, so the construction is explicit; (2) the signature is purely relational, so restriction is substructure inclusion with no function-symbol compatibility to check; (3) the site is a finite poset, so covers are finite. Extending to function symbols requires verifying that the glued interpretation of function symbols is well-defined on overlaps (a standard amalgamation argument in Fraïssé theory [11, 20]). Extending to infinite sites requires a compactness or direct-limit argument.

The connection between Fraïssé amalgamation and Grothendieck sites is made explicit by Caramello [19]: if a category C of finite structures satisfies the amalgamation property, the *atomic topology* J_{at} on C^{op} (covering sieves are the non-empty ones) is a valid Grothendieck topology, and the sheaf condition on $(C^{\text{op}}, J_{\text{at}})$ corresponds precisely to the model-theoretic amalgamation condition. Our Lemma 2.19 can be seen as an instance of this correspondence for the covering topology on a finite poset.

Remark 2.21 (Operational checklist: when do D and DD hold?). For a practitioner deciding whether the spine theorems apply to a specific site, the following checklist determines which hypotheses are satisfied:

Regime	**Assump. D**	**DD condition**
Finite poset, relational Σ, substructure restrictions	*proved* (Lemma 2.19)	DD2 (auto)
Geometric theories in a Grothendieck topos	yes (by defn.)	DD1 (auto)
Presheaf topos (e.g., $\mathbf{Set}^{C^{\text{op}}}$)	yes (auto)	DD2 (check)
Coherent theories, finite site	yes (amalg.)	DD1 (auto)
Arbitrary first-order, infinite site	**must verify**	**must verify**

"Auto" means the condition holds automatically for the regime; "check" means it must be verified for the specific site. All examples in this paper fall into the first two rows. The Calendar/Email/Slack site (Section 3.1) instantiates the first row (finite poset, relational signature, substructure restrictions).

3 Gluing: The Extension Torsor Lemma

Lemma 3.1 (Extension Torsor — [Spine])**.** *Let (C, J, M) be a predicate site and $\{q_i\}_{i \in I}$ local extensions over a cover $\{U_i \to U\}$, agreeing on pairwise overlaps up to equivalence. The obstruction to gluing is a class*

$$[\alpha] \in H^1(C, \text{Equiv}_{\text{ext}})$$

in the first non-abelian Čech cohomology of the site with coefficients in the sheaf of definable equivalences. The family $\{q_i\}$ glues iff $[\alpha] = 0$.

Proof. For each pair (i, j), let $\alpha_{ij} \in \mathrm{Equiv}_{\mathrm{ext}}(U_i \cap U_j)$ be the equivalence witnessing agreement: $q_j = q_i \circ \alpha_{ij}$ on the overlap. The family $\{\alpha_{ij}\}$ satisfies the cocycle condition on triple overlaps: $\alpha_{ij} \circ \alpha_{jk} = \alpha_{ik}$. A change of local representatives $q_i \mapsto q_i \circ \beta_i$ changes the cocycle by the coboundary $\alpha'_{ij} = \beta_i^{-1} \circ \alpha_{ij} \circ \beta_j$. The class $[\alpha] \in H^1$ is independent of representatives.

The final step—that $[\alpha] = 0$ implies a global extension exists—*invokes Assumption D* (effectivity of descent). Without D, the vanishing of the cocycle class implies only that local adjustments make the predicates agree strictly; Assumption D is required to conclude that these adjusted local extensions glue to a global object. (For when D holds in practice, see the operational checklist in Remark 2.21. In the Lean formalization, this step is `global_extension_from_trivial_cocycle`, the unique invocation site of `DescentAxiom.effective`.) □

Corollary 3.2 (No-Go for fixed-topology alignment — [Spine])**.** *Fix a predicate site (C, J, M), a cover $\{U_i \to U\}$, and a coefficient sheaf* $\mathrm{Equiv}_{\mathrm{ext}}$*. Suppose the Čech class $[\alpha] \in H^1(C, \mathrm{Equiv}_{\mathrm{ext}})$ is nontrivial. Then there is* no global extension *realizing the given local predicate specifications. In particular, any protocol that operates* within the fixed overlap structure*—using only restrictions, local adjustments, and message passing along the existing cover—cannot achieve global concept alignment, regardless of the number of communication rounds, the sophistication of the alignment algorithm, or the quality of the training data. Precisely: "local repair" means modifying the local extensions by a 0-cochain $\{\beta_i \in \mathrm{Equiv}_{\mathrm{ext}}(U_i)\}$, which changes the cocycle by a coboundary $\alpha'_{ij} = \beta_i^{-1}\alpha_{ij}\beta_j$ but cannot change the cohomology class $[\alpha]$. The obstruction is topological and can be removed only by changing the site (adding contexts or overlaps to alter the nerve) or changing the coefficient object (redefining what counts as "agreement").*

Proof. Immediate from Lemma 3.1: local adjustments change the cocycle by a coboundary $\alpha'_{ij} = \beta_i^{-1}\alpha_{ij}\beta_j$, but cannot change its H^1 class. A nontrivial class is invariant under all coboundary modifications. □

Remark 3.3 (Scope of the No-Go). The corollary is precisely scoped: it applies when the site topology and the coefficient sheaf are *held fixed*. Protocols that add new information sources (changing the cover), create higher-dimensional overlaps (changing the nerve), or redefine the equivalence relation on extensions (changing $\mathrm{Equiv}_{\mathrm{ext}}$) are not constrained by this result—and indeed, these are exactly the architectural remedies prescribed by Gate 1. The No-Go characterizes what "try harder within the current architecture" cannot accomplish; the three-gate diagnostic (Theorem 5.1) prescribes the structural changes that *can*.

Remark 3.4 (H^1 vs. H^2 boundary). The main theorem lives in the H^1 regime: $\mathrm{Equiv}_{\mathrm{ext}}$ forms a 1-group (strict cocycles), and H^1 classifies $\mathrm{Equiv}_{\mathrm{ext}}$-torsors by the classical correspondence [5, 26]. The cocycle formulas ($g_{ik} = g_{ij}g_{jk}$ on triple overlaps, coboundary $g'_{ij} = g_i g_{ij} g_j^{-1}$) follow Breen [26]; for a textbook treatment see Vistoli [27]. When $\mathrm{Equiv}_{\mathrm{ext}}$ is a groupoid (coherence of coherence is needed), the obstruction may live in H^2/gerbe classification. We acknowledge this boundary explicitly and do not claim results requiring higher coherence.

Corollary 3.5 (Contractible-Nerve Vanishing — [Spine])**.** *Let $\mathcal{U} = \{U_i \to U\}$ be a covering family and let $N(\mathcal{U})$ denote the* Čech nerve*: the simplicial set whose n-simplices are $(n+1)$-tuples $(i_0, \ldots, i_n)$ such that $U_{i_0} \cap \cdots \cap U_{i_n} \neq \emptyset$. Suppose:*

1. $\mathrm{Equiv}_{\mathrm{ext}}$ *is a* constant *sheaf of groups on* $\mathcal{U}$ *(i.e., restriction maps are isomorphisms), or more generally a locally constant sheaf; and*

2. $N(\mathcal{U})$ *is contractible (e.g., for three contexts, a single effective triple overlap fills the boundary* $\partial\Delta^2$ *to a disk; for larger covers, contractibility requires all higher simplices to be filled—not merely some 2-cells).*

Then $H^1(N(\mathcal{U}), \mathrm{Equiv}_{\mathrm{ext}}) = 0$ *and every locally compatible family of extensions glues globally. For non-constant* $\mathrm{Equiv}_{\mathrm{ext}}$*, contractibility of the nerve is necessary but the vanishing also requires that the coefficient sheaf is "untwisted" along the nerve (no monodromy). Monodromy is the induced action of* $\pi_1(N(\mathcal{U}))$ *on the stalks of* $\mathrm{Equiv}_{\mathrm{ext}}$*; for finite nerves, it reduces to checking whether the product of restriction maps around each generating cycle of* π_1 *acts trivially on* $\mathrm{Equiv}_{\mathrm{ext}}$*. In the finite constant-coefficient cases computed in this paper, condition (1) is automatic and monodromy vanishes.*

More generally: $H^1 = 0$ *when the site has cohomological dimension 0, or all equivalences are inner (every cocycle is a coboundary by conjugation).*

This corollary is machine-checked in Lean for $\mathbb{Z}/2\mathbb{Z}$ *coefficients on triangular covers (*`contractible_nerve_vanishing` *in* `Torsor.lean`*).*

Remark 3.6 (Čech vs. derived functor cohomology). Throughout this paper, H^1 denotes Čech cohomology *for a chosen cover*, computed via the Čech nerve $N(\mathcal{U})$. For a fixed cover $\mathcal{U}$, the Čech set $H^1(\mathcal{U}, \mathrm{Equiv}_{\mathrm{ext}})$ classifies only those $\mathrm{Equiv}_{\mathrm{ext}}$-torsors that *trivialize on* $\mathcal{U}$; the passage to the colimit over all coverings is required to capture all torsors. We do not address refinements, hypercovers, or the passage to derived functor cohomology. For non-abelian coefficients, the Čech definition *is* the standard definition—there is no competing derived-functor version [5]. For finite sites—where all our examples live—the distinction between cover-level and colimit H^1 is immaterial (every torsor trivializes on a sufficiently fine finite cover).

Remark 3.7 (Triple overlap: existence vs. effectiveness). In applied examples, the triple intersection object $U_1 \cap U_2 \cap U_3$ often exists formally but is empty, non-effective for descent (missing witness data), or its maps do not satisfy the conditions needed for the vanishing theorem. For instance, the Calendar ∩ Email ∩ Slack object may exist as a type but contain no meetings visible simultaneously in all three systems. The correct reading of Corollary 3.5 is: real obstructions arise when higher intersections are *absent or empty* (circular topology), not merely when they are formally present. The vanishing theorem requires an *effective* triple overlap—one that carries actual model data, not just an empty type.

This explains the "Calendar/Email/Slack" phenomenon: the three data sources have pairwise overlaps (meetings appearing in two systems) but no effective triple overlap (no single record visible in all three). The nerve is a circle, not a filled triangle, and $H^1(S^1, \mathbb{Z}/2\mathbb{Z}) \cong \mathbb{Z}/2\mathbb{Z} \neq 0$.

Engineering diagnostic: empty or non-effective triple overlaps mean the cover behaves like a 1-dimensional nerve for obstruction purposes. When designing a multi-source extraction system, verify that pairwise overlaps share a common "meeting point" with actual model data—if not, expect circular-nerve topology and plan for H^1 obstructions.

3.1 Worked example: the Calendar/Email/Slack site

We now carry out the full H^1 computation for Problem B (Example 1.2), demonstrating the complete diagnostic chain from site definition through obstruction identification to

architectural prescription.

Step 1: Define the site. Let C be the category with three objects Cal, Email, Slack and three overlap objects:

$$\mathsf{CE} = \mathsf{Cal} \cap \mathsf{Email}, \qquad \mathsf{CS} = \mathsf{Cal} \cap \mathsf{Slack}, \qquad \mathsf{ES} = \mathsf{Email} \cap \mathsf{Slack}.$$

Morphisms are the six inclusions $\mathsf{CE} \hookrightarrow \mathsf{Cal}$, $\mathsf{CE} \hookrightarrow \mathsf{Email}$, etc. There is *no* triple overlap object $\mathsf{Cal} \cap \mathsf{Email} \cap \mathsf{Slack}$—no single record is visible in all three systems simultaneously.

The covering topology J declares $\{\mathsf{Cal}, \mathsf{Email}, \mathsf{Slack}\}$ as a cover of the global context U.

Step 2: The model stack. $M(\mathsf{Cal})$ is the groupoid of calendar-structured data (events with timestamps, attendees, durations). $M(\mathsf{Email})$ is the groupoid of email-structured data (messages with senders, recipients, threads). $M(\mathsf{Slack})$ is the groupoid of Slack-structured data (messages with channels, reactions, threads). The restriction functors project along the overlaps: $M(\mathsf{CE})$ contains records visible in both Calendar and Email (e.g., a meeting that has both a calendar invite and a confirmation email).

Step 3: The local extensions. An agent examining the calendar invents the predicate `is_meeting` locally on $M(\mathsf{Cal})$: a calendar event is a meeting if it has ≥ 2 attendees and a video link. A second agent examining email invents `is_meeting` on $M(\mathsf{Email})$: an email thread is a meeting if it contains scheduling language and an attachment. A third agent invents `is_meeting` on $M(\mathsf{Slack})$: a Slack thread is a meeting if it's in a channel named `#meetings` or contains a Zoom link.

Step 4: Compute the Čech nerve. The nerve $N(\mathcal{U})$ has:

- *0-simplices*: Cal, Email, Slack (three vertices).
- *1-simplices*: CE, CS, ES (three edges).
- *No 2-simplex*: $\mathsf{Cal} \cap \mathsf{Email} \cap \mathsf{Slack} = \emptyset$.

This is the boundary of a triangle: $N(\mathcal{U}) \cong S^1$.

Step 5: Compute the cocycles. On each overlap, compare the two local predicates. Let $\mathrm{Equiv}_{\mathrm{ext}} \cong \mathbb{Z}/2\mathbb{Z}$ (each local `is_meeting` is determined up to a flip: agree or disagree).

- On CE: Calendar says event e *is* a meeting (it has attendees); Email says the same thread is *not* a meeting (no scheduling language found). Transition: $\alpha_{\mathsf{CE}} = 1$ (flip).
- On CS: Calendar says event e is a meeting; Slack says the corresponding thread is a meeting (Zoom link found). Transition: $\alpha_{\mathsf{CS}} = 0$ (agree).
- On ES: Email says thread e is not a meeting; Slack says it is. Transition: $\alpha_{\mathsf{ES}} = 1$ (flip).

The cocycle is $(\alpha_{\mathsf{CE}}, \alpha_{\mathsf{CS}}, \alpha_{\mathsf{ES}}) = (1, 0, 1) \in (\mathbb{Z}/2\mathbb{Z})^3$.

Step 6: Is it a coboundary? A coboundary has the form $(\beta_C - \beta_E,\ \beta_C - \beta_S,\ \beta_E - \beta_S)$ for $\beta_C, \beta_E, \beta_S \in \mathbb{Z}/2\mathbb{Z}$. We need:

$$\begin{aligned}\beta_C - \beta_E &= 1,\\ \beta_C - \beta_S &= 0,\\ \beta_E - \beta_S &= 1.\end{aligned}$$

From the first two: $\beta_E = \beta_C + 1$, $\beta_S = \beta_C$. Then $\beta_E - \beta_S = (\beta_C + 1) - \beta_C = 1$. **This is consistent.** The cocycle $(1, 0, 1)$ *is* a coboundary (take $\beta_C = 0$, $\beta_E = 1$, $\beta_S = 0$).

Interpretation: the disagreement can be resolved by "flipping" Email's predicate: redefine `is_meeting` on Email as its negation. After this local adjustment, all three agents agree.

Step 7: The non-trivial case. Now suppose the transitions are $(\alpha_{\mathsf{CE}}, \alpha_{\mathsf{CS}}, \alpha_{\mathsf{ES}}) = (1, 1, 1)$: every pair of agents disagrees. We need:

$$\begin{aligned}\beta_C - \beta_E &= 1,\\ \beta_C - \beta_S &= 1,\\ \beta_E - \beta_S &= 1.\end{aligned}$$

From the first two: $\beta_E = \beta_C + 1$, $\beta_S = \beta_C + 1$. Then $\beta_E - \beta_S = 0 \neq 1$. **Contradiction.**

The cocycle $(1, 1, 1)$ is *not* a coboundary. $[\alpha] \neq 0$ in $H^1(S^1, \mathbb{Z}/2\mathbb{Z})$. No local adjustment to the agents' predicates can make them all agree. The disagreement is *topological*: it arises from the circular structure of the overlap graph, not from any individual agent's error.

Step 8: The architectural prescription. The framework provides a concrete fix: *add a data source that creates a triple overlap.* Introduce Zoom logs, visible in all three systems (a Zoom meeting generates a calendar event, an email notification, and a Slack bot message). Now:

$$\mathsf{Cal} \cap \mathsf{Email} \cap \mathsf{Slack} \cap \mathsf{Zoom} \neq \emptyset.$$

The nerve gains a 2-simplex (filled triangle). By the Contractible-Nerve Vanishing theorem (Corollary 3.5), $H^1 = 0$, and *every* locally compatible family of predicates glues. The topological obstruction is eliminated by architectural choice, not by better training or more data.

Remark 3.8 (What the computation shows). This example demonstrates the full diagnostic chain:

1. Define the site → compute the nerve → identify the topology (S^1).
2. Compute the cocycle from local predicate disagreements.
3. Check coboundary condition → diagnose resolvable vs. irreconcilable.
4. If irreconcilable: the H^1 class prescribes the fix (change the topology).

Every step is computable for finite sites. The output is not "try harder" but a structural diagnosis with a constructive remedy.

Remark 3.9 (Abelian vs. non-abelian coefficients). The worked example uses $\mathrm{Equiv}_{\mathrm{ext}} \cong \mathbb{Z}/2\mathbb{Z}$, an abelian group. The general framework treats non-abelian $\mathrm{Equiv}_{\mathrm{ext}}$. What changes? The coboundary equation becomes non-commutative: instead of the linear system $\alpha'_{ij} = -\beta_i + \alpha_{ij} + \beta_j$, one must solve $\alpha'_{ij} = \beta_i^{-1} \cdot \alpha_{ij} \cdot \beta_j$ in a non-abelian group. The coboundary check becomes a *conjugacy* problem (is the cocycle conjugate to the identity?) rather than a linear algebra problem. For finite non-abelian $\mathrm{Equiv}_{\mathrm{ext}}$, this is still decidable by exhaustive search over $|\mathrm{Equiv}_{\mathrm{ext}}|^{|I|}$ possible coboundaries, but the structure of H^1 as a *pointed set* (not a group) means there is no additive cancellation. In the abelian case, H^1 is a group and one can compute its rank; in the non-abelian case, H^1 is merely a set of conjugacy classes of cocycles, and the diagnostic is: either the class is trivial (resolvable) or it names a specific irreconcilable obstruction. The Calendar/Email/Slack computation generalizes directly: replace $\mathbb{Z}/2\mathbb{Z}$ with any finite group G of definable equivalences, and the computation proceeds identically with group multiplication replacing addition.

Remark 3.10 (Relation to prior work). To our knowledge, this is the first explicit end-to-end computation of non-abelian Čech H^1 on a finite category with a Grothendieck topology. All prior non-abelian H^1 computations occur for topological spaces, algebraic varieties, or group cohomology [5, 28, 26]. The closest analogues in applied settings use *abelian* sheaf cohomology: Robinson [29] uses vector-space-valued sheaf cohomology over finite sensor networks to detect data fusion inconsistencies, and Abramsky [30] establishes a correspondence between local consistency in relational databases and Bell non-locality in quantum mechanics via sheaf-theoretic methods. Neither computes non-abelian cohomological obstructions. The Smith–Bendich–Harer persistent obstruction theory [31] is perhaps the closest in spirit, detecting when database JOINs fail via model-category obstruction cocycles rather than Čech cocycles.

4 Conservativity Descent

The conservativity descent theorem requires a preliminary lemma ensuring that locally-chosen conservative lifts can be made compatible on overlaps. This is the step that connects the local existence guarantee (essential surjectivity) to the global descent machinery.

Lemma 4.1 (Compatibility of conservative lifts)**.** *Let (C, J, M) be a predicate site, $M' \to M$ an extension of model stacks, and $\{U_i \to U\}$ a cover. Suppose:*

1. Local model-conservativity*: for each i, the forgetful functor $M'(U_i) \to M(U_i)$ is essentially surjective.*

2. Thin fibers on overlaps*: for each pair (i, j) and each base model $A \in M(U_i \cap U_j)$, the full subgroupoid of expansions of A in $M'(U_i \cap U_j)$ satisfying the constraint package D is either empty or* thin *(connected with trivial automorphism group): any two such expansions are connected by a* unique *isomorphism.*

Then for any base model $A \in M(U)$, there exists a family of expansions $\{A'_i \in M'(U_i)\}$ that forms a descent datum: the restrictions $A'_i|_{U_i \cap U_j}$ and $A'_j|_{U_i \cap U_j}$ are isomorphic in $M'(U_i \cap U_j)$, and the isomorphisms satisfy the cocycle condition on triple overlaps.

Proof. By (1), choose arbitrary expansions $A'_i \in M'(U_i)$ with $A'_i|_\Sigma = A|_{U_i}$. On each overlap $U_i \cap U_j$, the restrictions $A'_i|_{U_i \cap U_j}$ and $A'_j|_{U_i \cap U_j}$ are both expansions of $A|_{U_i \cap U_j}$ satisfying the constraint package. By (2), the fiber subgroupoid is thin, so there exists a *unique* isomorphism $\varphi_{ij} \colon A'_i|_{U_i \cap U_j} \xrightarrow{\sim} A'_j|_{U_i \cap U_j}$. On triple overlaps $U_i \cap U_j \cap U_k$, both $\varphi_{ij} \circ \varphi_{jk}$ and φ_{ik} are isomorphisms from $A'_i|_{U_{ijk}}$ to $A'_k|_{U_{ijk}}$ in a thin groupoid; uniqueness of the isomorphism forces $\varphi_{ij} \circ \varphi_{jk} = \varphi_{ik}$. Thus $\{\varphi_{ij}\}$ satisfies the cocycle condition, and $\{A'_i, \varphi_{ij}\}$ is a descent datum. □

Remark 4.2 (Why thin fibers, not just connected fibers). The thin-fiber condition (2) is strictly stronger than "any two expansions are isomorphic" (connected fibers). In a connected but non-thin groupoid, two expansions may be related by *multiple* isomorphisms, and choosing different isomorphisms on overlaps can break the cocycle condition $\varphi_{ij} \circ \varphi_{jk} = \varphi_{ik}$. The obstruction to choosing coherent isomorphisms in the non-thin case lives in H^2 (a gerbe classification), which is precisely the "gerbe boundary" beyond our scope (Section 2). The thin-fiber condition says: the constraint package D is restrictive enough that the expansion is determined up to unique isomorphism—no automorphism ambiguity remains. In practice, thinness can be verified by checking any of:

- *Rigidity*: overlap structures admit no nontrivial automorphisms under the base signature (common for finite relational structures with enough constants or constraints);
- *Pinning*: the constraint package D fixes enough structure on each overlap U_{ij} that any two D-compatible expansions must coincide up to unique renaming;
- *Failure signal*: if multiple inequivalent witness isomorphisms exist between overlap expansions, the problem has moved to the H^2/gerbe regime—outside the scope of Gate 2.

For finite relational sites with sufficiently constrained D, thinness holds automatically: finite structures with no non-trivial automorphisms have thin extension groupoids.

Theorem 4.3 (Conservativity Descent — [Spine])**.** *Let (C, J, M) be a predicate site where M satisfies descent. Let M' be an extension stack also satisfying descent and the thin-fiber condition of Lemma 4.1. If each local extension is* ***model-theoretically conservative*** *(Definition 2.16), then the global extension is model-theoretically conservative (and therefore also deductively conservative).*

Proof. Let $A \in M(U)$ be a base model. We must produce $A' \in M'(U)$ expanding A. Restrict: $A|_{U_i} \in M(U_i)$. By local model-conservativity, $\exists$ expansions $A'_i \in M'(U_i)$ with $A'_i|_\Sigma = A|_{U_i}$. By Lemma 4.1, the family $\{A'_i\}$ can be equipped with overlap isomorphisms forming a descent datum. By effectivity of descent for M' (Assumption D applied to the extension stack), the descent datum glues to $A' \in M'(U)$ with $A'|_\Sigma = A$.

Note: this proof uses model-theoretic conservativity (essential surjectivity), not merely deductive conservativity (theory inclusion). The step "there exists an expansion A'_i of $A|_{U_i}$" requires that the forgetful functor is essentially surjective, not just that theories are included. The step "the family glues" requires both the compatibility supply (Lemma 4.1) and descent for M'. This is why we distinguish the two notions of conservativity in Definition 2.16 and isolate descent as an explicit hypothesis. □

Corollary 4.4 (Finite relational sites)**.** *For finite poset categories with the covering topology and groupoids of finite relational structures, the amalgamation hypothesis holds (by standard finite-structure amalgamation). Conservativity descent reduces to: local conservativity + trivial extension torsor ⇒ global conservativity.*

Example 4.5 (Thin-fiber failure on a contractible nerve)**.** Consider contexts $U_1 = \{a < b < c\}$, $U_2 = \{b < c < d\}$, overlap $U_{12} = \{b < c\}$. The Čech nerve has two vertices and one edge—a line segment, contractible, so $H^1 = 0$. Define $q_1(\text{“large”}) = \{c\}$, $q_2(\text{“large”}) = \{d\}$. Each extension is locally conservative. On the overlap, q_1 assigns “large” to c; q_2 does not (since $q_2 = \{d\}$ and $c \neq d$).

This example demonstrates the necessity of the thin-fiber condition in Lemma 4.1. The nerve is contractible, so the *topological* obstruction vanishes. But the fiber subgroupoid on U_{12} is not thin: $q_1|_{U_{12}}$ and $q_2|_{U_{12}}$ are genuinely different predicates (one includes c, the other does not), so they are not connected by any isomorphism in the extension groupoid. Without the compatibility supply, we cannot form a descent datum. Any global predicate must decide whether c is large or not, introducing a new consequence absent from one of the local theories. The failure is not topological (the nerve is contractible) but stems from incompatible local choices that violate the thin-fiber hypothesis—it is the violation of Lemma 4.1(2) that makes conservativity descent inapplicable.

This counterexample is formalized and machine-verified in Lean 4.

5 Three Diagnostic Gates

The SCPI decision is not a single yes/no check but a *pipeline* of three sequential gates, each with different mathematical character and different failure remedies. The gates are ordered: Gate 2 is meaningful only after Gate 1 passes, and Gate 3 only after Gate 2.

Theorem 5.1 (Diagnostic Decomposition — [Spine])**.** *The SCPI diagnostic proceeds through three gates:*

Gate 1 (Gluing): *Compute the Čech cocycle* $[\alpha] \in H^1(C, \mathrm{Equiv}_{\mathrm{ext}})$*. If* $[\alpha] \neq 0$*: the obstruction is topological; no local adjustment suffices.* Remedy*: change the cover topology (add a data source creating a higher-dimensional simplex). If* $[\alpha] = 0$ *and Assumption D holds: a global extension exists.*

Gate 2 (Conservativity): *Given a global extension (from Gate 1), check model-theoretic conservativity: does every base model admit an expansion? If not: the extension introduces new consequences.* Remedy*: weaken the constraint package or enlarge the base theory.*

Gate 3 (Definability, optional): *Given a conservative global extension (from Gate 2), check whether q is explicitly definable by a formula in* Σ*. If not: q is “there” but cannot be named.* Remedy*: enrich the vocabulary, or accept implicit definability.*

Remark 5.2 (Gates, not a direct sum)*.* The three gates are *not* symmetric “independent components” of a single obstruction. They form a decision pipeline: Gate 1 is cohomological (about the topology of the cover), Gate 2 is model-theoretic (about the expansion property of the forgetful functor), and Gate 3 is about definability (and itself cohomological—see Remark 6.5). The gates can fail separately (demonstrated below),

but they are not a direct-sum decomposition. The analogy is a manufacturing pipeline: a product can fail quality control at different stations, but the stations are ordered, not parallel.

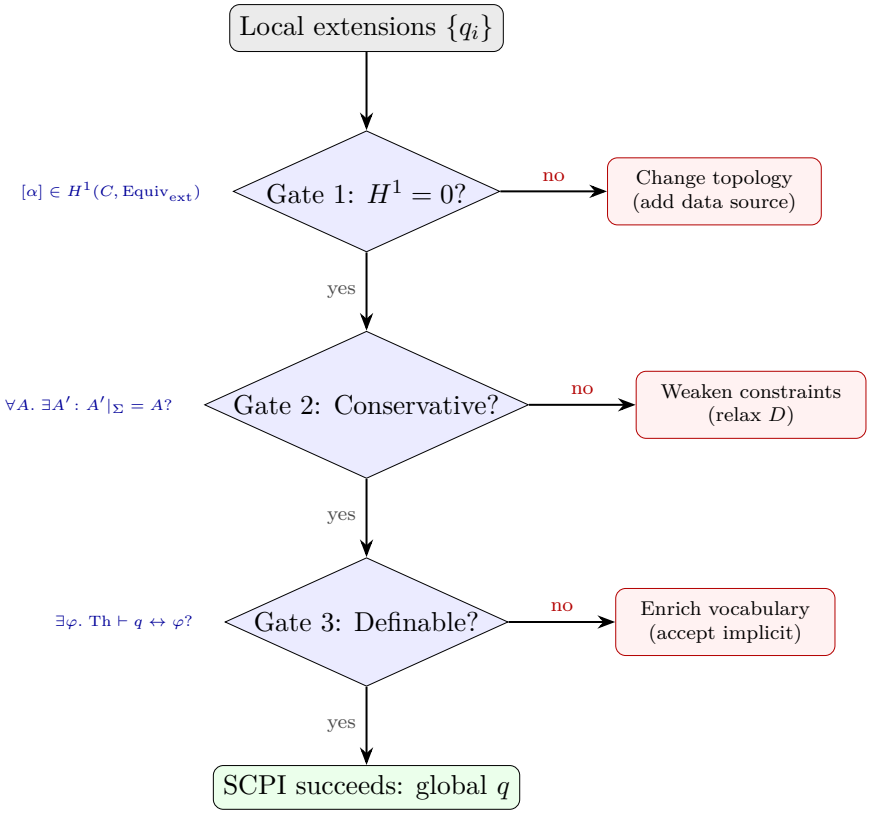


Figure 1: The three-gate diagnostic pipeline. Local extensions enter at the top. Each gate applies a distinct mathematical test; failure at any gate produces a structural diagnosis with a specific architectural remedy. The gates are sequential: Gate 2 is meaningful only after Gate 1 passes. Gate 3 is itself a descent problem (the definability obstruction is a sheaf-condition failure for the presheaf of local definers; see Remark 6.5).

Theorem 5.3 (Separability — [Spine])**.** *The three gates can fail independently: for each gate, there exists a predicate site that fails at that gate while passing the others.*

Proof. We exhibit three separating examples, one for each gate.

Gate 1 failure (topological), Gates 2–3 pass. Three contexts U_1, U_2, U_3 with pairwise overlaps but no effective triple overlap (circular nerve $\cong S^1$). Each local extension is model-conservative with explicit local definers. The Čech cocycle $(1, 1, 1) \in (\mathbb{Z}/2\mathbb{Z})^3$ is not a coboundary (Section 3.1, Step 7). Gluing fails for topological reasons; conservativity and definability are not at issue.

Gate 2 failure (conservativity), Gate 1 passes. A single context U with the trivial cover $\{U \to U\}$. The nerve is a point; $H^1 = 0$ (Gate 1 passes trivially). Let $M(U)$ be the groupoid of finite graphs. Define an extension by the predicate $q(v) =$ "v is colored red" with constraint package D: "at least one vertex is red, and every red vertex has degree ≥ 3." A cycle graph (maximum degree 2) has no expansion satisfying D: the non-emptiness constraint forces at least one red vertex, but no vertex has degree ≥ 3.

(Without the non-emptiness clause, $q = \varnothing$ would vacuously satisfy the degree constraint; the positive example requirement in the constraint package is essential.) The forgetful functor is not essentially surjective: model-theoretic conservativity fails. The failure is not topological (the nerve is contractible) but model-theoretic (the constraint excludes certain base models from expansion).

Gate 3 failure (definability), Gates 1–2 pass. A single context U, trivial cover ($H^1 = 0$, Gate 1 passes). Let $M(U)$ be the groupoid of *odd-length* finite linear orders $(\{1, \ldots, 2m+1\}, \leq)$ for $m \geq 1$. Define $q(x) =$ "x is the median element" (the unique element at position $m+1$). For every odd-length linear order, the median is uniquely determined by the order—so the extension is model-conservative: every base model has exactly one expansion (Gate 2 passes). (Restricting to odd-length orders is essential; for even-length orders, no unique median exists and the expansion would not be well-defined.) But "median" is not definable by any fixed first-order formula in $\{\leq\}$: a formula of quantifier rank k cannot distinguish position $\lfloor n/2 \rfloor$ from position $\lfloor n/2 \rfloor + 1$ when $n > 2^k$ (Ehrenfeucht–Fraïssé argument). The predicate is implicitly definable (uniquely determined) but not explicitly definable. Gate 3 fails. □

Theorem 5.4 (Exhaustiveness)**.** *If all three gates pass, SCPI succeeds: ∃ a global extension that is sheaf-compatible, model-theoretically conservative, and explicitly definable.*

Theorem 5.5 (Computability)**.** *Each gate is detectable by different machinery: Čech cohomology (polynomial for finite sites with abelian coefficients; $|\mathrm{Equiv}_{\mathrm{ext}}|^{O(k)} \cdot n^{O(1)}$ for non-abelian coefficients of bounded size on covers of treewidth k), model checking (decidability depends on base logic), interpolant computation (decidable for effective-interpolation fragments).*

5.1 Topological invariance and cross-domain transfer

The gate decomposition reveals a non-obvious invariance principle: Gate 1 depends *only* on the topology of the cover and the coefficient group, not on the content of the data.

Theorem 5.6 (Topological Invariance — [Spine])**.** *Let (C_1, J_1, M_1) and (C_2, J_2, M_2) be predicate sites with covers $\mathcal{U}_1$, $\mathcal{U}_2$. Suppose:*

1. *$N(\mathcal{U}_1) \cong N(\mathcal{U}_2)$ as simplicial sets (isomorphic Čech nerves), and*
2. *$\mathrm{Equiv}_{\mathrm{ext}1} \cong \mathrm{Equiv}_{\mathrm{ext}2}$ as sheaves of groups on the respective nerves (compatible with the isomorphism in (1)).*

Then $H^1(\mathcal{U}_1, \mathrm{Equiv}_{\mathrm{ext}1}) \cong H^1(\mathcal{U}_2, \mathrm{Equiv}_{\mathrm{ext}2})$ as pointed sets: the Gate 1 obstruction landscapes are isomorphic. In particular, any cocycle that is (resp. is not) a coboundary in one site has a corresponding cocycle with the same property in the other.

Proof. H^1 of a simplicial set with coefficients in a sheaf of groups is determined by the simplicial set and the coefficient sheaf, not by any additional structure on the site. An isomorphism of nerves induces a bijection on n-simplices for all n; an isomorphism of coefficient sheaves transports the cocycle condition ($\alpha_{ij}\alpha_{jk} = \alpha_{ik}$) and coboundary relation ($\alpha'_{ij} = \beta_i^{-1}\alpha_{ij}\beta_j$) identically. The bijection preserves the distinguished point (trivial cocycle). □

Corollary 5.7 (Cross-domain transfer — [Spine])**.** *The Calendar/Email/Slack data-integration site (Section 3.1) and a Vision/Text/Genomic model-alignment site (Section 1.2) have* identical *Gate 1 obstruction landscapes whenever their Čech nerves and coefficient groups match—even though the underlying data, schemas, and semantics are completely different. Concretely: if both sites have circular nerve (S^1) and coefficient group G, then:*

- *The number of fundamentally distinct irreconcilable disagreements is $|H^1(S^1, G)|-1$ in both settings.*
- *For abelian G: this equals $|G| - 1$.*
- *For non-abelian G: this equals (number of conjugacy classes of G) -1.*
- *An architectural fix (adding a triple overlap to make the nerve contractible) that works in one domain* must *work in the other.*

Remark 5.8 (Why this is surprising)*.* The invariance theorem says: a data-integration failure in an enterprise system and a feature-alignment failure in a multi-model AI pipeline are *the same obstruction* if their overlap topologies match. The diagnostic transfers across domains—not by analogy, but by theorem. This is a concrete prediction: if you have already computed the H^1 obstruction landscape for one site, you can immediately classify all possible failures in any other site with the same nerve and coefficient group, without examining the data.

Example 5.9 (Cross-domain transfer: Vision/Text/Genomic site)**.** We demonstrate the topological invariance theorem by constructing an AI interpretability site with the same obstruction landscape as the Calendar/Email/Slack data-integration site.

Site definition. Let V (vision), T (text), G (genomic) be three foundation models, each trained on different modalities, with pairwise probing tasks as overlaps: $V \cap T$ (image captioning), $T \cap G$ (biomedical NLP), $V \cap G$ (histopathology). No single evaluation task probes all three modalities simultaneously—there is no effective triple overlap. The predicate q is "disease-relevant feature" (a latent concept each model represents locally but which must be aligned globally).

Nerve. The Čech nerve is: three vertices (V, T, G), three edges (VT, TG, VG), no 2-simplex—the boundary $\partial\Delta^2 \cong S^1$. This is exactly the nerve of the Calendar/Email/Slack site.

Coefficients. Let $\text{Equiv}_{\text{ext}} \cong \mathbb{Z}/2\mathbb{Z}$ (the SAE feature can be active or its complement). By Theorem 5.6, $H^1(S^1, \mathbb{Z}/2\mathbb{Z}) \cong \mathbb{Z}/2\mathbb{Z}$: there is exactly one non-trivial obstruction class, a Möbius-type twist where pairwise alignments are locally consistent but globally incoherent.

Diagnosis. Suppose each pair of models agrees on "disease-relevant feature" (pairwise probing succeeds), but the composition of pairwise alignment maps around the loop $V \to T \to G \to V$ flips the polarity. This is the non-trivial cocycle—the *same* cocycle as the Calendar/Email/Slack "is-recent" disagreement. The diagnostic is identical: the system needs a *shared evaluation set* (a dataset probing all three modalities simultaneously) to create the missing 2-simplex, making the nerve contractible.

Punchline. The enterprise data-integration fix ("add a data source creating a triple overlap") and the AI alignment fix ("add a multimodal benchmark") are the same architectural prescription, derived from the same theorem, for the same topological reason.

Remark 5.10 (Connection to contextuality). The SCPI obstruction has a precise analogue in quantum foundations. Abramsky and Brandenburger [30] proved that *contextuality*—the impossibility of assigning globally consistent values to quantum observables—is equivalent to the nonexistence of a global section of a presheaf of measurement outcomes. Our Gate 1 obstruction is an instance of the same global-section problem: a nontrivial class in $H^1(C, \mathrm{Equiv}_{\mathrm{ext}})$ means the local predicates invented by different agents cannot be simultaneously realized by any global predicate, just as a contextual hidden-variable model cannot simultaneously realize all local quantum measurement outcomes. The overlap topology of the agent architecture plays the role of the measurement context structure. This is not merely an analogy: both are instances of the sheaf-theoretic obstruction to extending local sections to global ones, classified by the same type of cohomological invariant on the same type of mathematical object (a presheaf on a site). The No-Go corollary (Corollary 3.2) is a structural analogue of Bell's theorem in this setting: certain architectures are *structurally* incapable of global alignment, regardless of the quality of local computation.

6 Conditional Beth Definability for Sites

Definition 6.1 (Implicit definability over a site)**.** A predicate q is *implicitly definable over the site* (C, J, M) if: for every pair of descent-compatible expansions (A', q') and (A', q'') of the same base model $A' \in M'(U)$ satisfying the same constraint package D, we have $q' = q''$ (the predicate is uniquely determined by the base model and constraints). Equivalently, the fiber of the forgetful functor $\mathrm{Ext}(M)(U) \to M(U)$ over each base model has at most one isomorphism class.

Theorem 6.2 (Beth for Geometric Sites — [Spine])**.** *Let (C, J, M) be a predicate site where each* $\mathrm{Th}(U)$ *is a* geometric theory *(axiomatized by geometric sequents). Suppose DD1 holds (automatic for coherent/geometric theories in Grothendieck toposes). If q is implicitly definable over the site, then q is explicitly definable by a geometric formula: there exists φ in the base signature Σ with* $\mathrm{Th}(U) \vdash \forall x.\, q(x) \leftrightarrow \varphi(x)$.

Proof.

1. *Local definability.* By implicit definability (Definition 6.1) restricted to each fiber, q is uniquely determined in $M(U_i)$. Since $\mathrm{Th}(U_i)$ is a geometric theory, the internal logic of $\mathrm{Sh}(C, J)$ is intuitionistic. Beth definability holds for intuitionistic predicate logic [1, 25]; the definability theorem for coherent logic is Johnstone [7] (D3.5.1). Hence for each i there exists φ_i geometric in Σ with $\mathrm{Th}(U_i) \vdash \forall x.\, q(x) \leftrightarrow \varphi_i(x)$.

2. *Matching.* The local interpolants φ_i are geometric formulas. Geometric formulas are preserved by inverse image functors of geometric morphisms [7] (Prop. D1.3.11). The restriction $\mathrm{Th}(U_i) \to \mathrm{Th}(U_i \cap U_j)$ is the inverse image of a geometric morphism (the inclusion of the overlap). Since q is implicitly definable over the entire site, $q|_{U_i \cap U_j}$ is uniquely determined in $M(U_i \cap U_j)$. Both $\varphi_i|_{U_i \cap U_j}$ and $\varphi_j|_{U_i \cap U_j}$ define q on the overlap; by uniqueness (Beth in the fiber $U_i \cap U_j$), they are equivalent: $\mathrm{Th}(U_i \cap U_j) \vdash \forall x.\, \varphi_i(x) \leftrightarrow \varphi_j(x)$.

3. *Gluing.* The matched family $\{\varphi_i\}$ is a compatible section of the presheaf $\mathcal{D}\colon U \mapsto \{\varphi \in \mathrm{Geom}(\Sigma) \mid \mathrm{Th}(U) \vdash \forall x.\, q(x) \leftrightarrow \varphi(x)\}$ of geometric definers over the cover. Under DD1, geometric entailment is local in the internal logic of $\mathrm{Sh}(C, J)$, so $\mathcal{D}$ satisfies the sheaf condition. The family glues to a global geometric formula φ.

4. *Verification.* $\mathrm{Th}(U) \vdash \forall x.\ q(x) \leftrightarrow \varphi(x)$ by locality: the equivalence holds on each U_i (by Step 1) and extends globally by the sheaf condition (Step 3).

□

Remark 6.3 (Status and scope of Theorem 6.2). Each step invokes a known result: Step 1 uses Beth/Kreisel for intuitionistic logic [25], Step 2 uses geometric preservation [7] (D1.3.11), Step 3 uses DD1 (which holds automatically in the geometric regime). The assembly of these ingredients into a definability theorem *over a site* is new. The underlying duality between interpolation (Craig [4]) and definability (Beth [1]) is classical; recent work extends Craig interpolation to subgeometric fragments [21]. The theorem covers all geometric/coherent theories on Grothendieck sites—a substantial class including the examples of this paper.

Theorem 6.4 (Beth for Sites — general, [CONDITIONAL])**.** *For non-geometric theories, the same conclusion holds if one assumes* ***one of****:*

(H1) Conservative restrictions*: restriction maps reflect equivalences in the interpolation fragment.*

(H3) Fibred implicit definability*: implicit definability quantifies over descent-compatible models, ensuring local interpolants match.*

The argument follows the same four steps as Theorem 6.2; the matching step uses H1 or H3 in place of geometric preservation. See Makkai [24] (strong conceptual completeness) and Pitts [22, 23] for the algebraic/categorical infrastructure.

Remark 6.5 (Why the geometric hypothesis does real work). Without geometric logic (or H1/H3), the matching step can fail. Consider a site with U_1, U_2, overlap U_{12}, where $\mathrm{Th}(U_1)$ includes axiom A absent from $\mathrm{Th}(U_{12})$. Local interpolants φ_1, φ_2 may agree in U_1 (forced by A) but diverge on U_{12} (where A is lost). The interpolant φ_1 may use non-geometric operations (negation, arbitrary universal quantification) that are not preserved by restriction. Geometric formulas *are* preserved, so matching is guaranteed—this is the precise content of hypothesis H2.

In the language of cohomology: the *definability obstruction* is a descent problem for the presheaf of local definers. Under DD1, this presheaf is a sheaf (local definers automatically glue); without DD1, it may fail the sheaf condition, and the failure is the definability gap.

7 Complexity Map

Theorem 7.1 (Complexity landscape)**.**

- Propositional sites*: SCPI is in Σ_2^p (guess q, verify by co-NP oracle). Lower bound: Σ_2^p-hard (conditional on conservativity being a genuine $\forall$-check).*
- Decidable first-order fragments*: decidable; complexity dominated by conservativity check.*
- Full first-order*: r.e.-complete (equivalently, Σ_1^0 in the arithmetical hierarchy).*

Proof sketch. Upper bound (Σ_2^p). The SCPI decision problem has quantifier structure $\exists q$ $\forall$(gluing) $\forall$(conservativity): guess the global predicate q (existential), then verify (i) the Čech cocycle vanishes (checkable in polynomial time for finite sites with fixed coefficient group, since the nerve has $O(|I|^2)$ edges and the coboundary system is solvable in $O(|\mathrm{Equiv}_{\mathrm{ext}}|^{|I|})$ time—polynomial when $|\mathrm{Equiv}_{\mathrm{ext}}|$ is constant; linear-algebraic for abelian $\mathrm{Equiv}_{\mathrm{ext}}$), and (ii) model-theoretic conservativity holds (a $\forall$-check: for every base model, an expansion exists). The verification is in co-NP, placing SCPI in $\Sigma_2^p = \mathrm{NP}^{\text{co-NP}}$.

Lower bound (Σ_2^p-hard). Reduce from Σ_2^p-complete QBF_2 ($\exists \vec{x}\, \forall \vec{y}\, \varphi(\vec{x}, \vec{y})$): encode the existential variables as the predicate q, the universal variables as model choices, and the formula φ as the conservativity constraint. The reduction is polynomial when the site has a fixed finite structure.

Full first-order. SCPI subsumes the satisfiability problem (take a trivial site with one context): $\exists q$ satisfying constraints is r.e. The conservativity check (is a sentence φ a consequence of $T + q$?) is co-r.e., making the combined problem Σ_1^0-complete. □

Theorem 7.2 (Optimal cover hardness)**.** *Finding the coarsest cover on which SCPI succeeds within coherence budget B is NP-hard. Greedy achieves $O(\ln n)$ approximation.*

Proof sketch. Reduce from weighted set cover. Given a universe $U = \{u_1, \ldots, u_n\}$ and sets $S_1, \ldots, S_m$ with weights, construct a predicate site where each S_i is a context, overlaps are set intersections, and the coherence budget B bounds the total cover weight. A cover on which SCPI succeeds (the cocycle vanishes) must include enough contexts to form a contractible nerve over the relevant elements—this requires covering U. The reduction is polynomial. The $O(\ln n)$ approximation follows from the submodularity of the coverage function [12]. □

Theorem 7.3 (FPT for bounded treewidth)**.** *When the site has treewidth k, SCPI is fixed-parameter tractable: $f(k) \cdot n^{O(1)}$.*

Proof sketch. On a site of treewidth k, the Čech nerve has treewidth $\leq k$. The cocycle condition (a system of group equations on edges, subject to the cocycle condition on triangles) can be solved by dynamic programming on a tree decomposition: at each bag of width $\leq k+1$, enumerate all $|\mathrm{Equiv}_{\mathrm{ext}}|^{k+1}$ possible assignments, propagate compatibility along the tree. The conservativity check at each node is independent and polynomial for fixed k. Total: $|\mathrm{Equiv}_{\mathrm{ext}}|^{O(k)} \cdot n^{O(1)} = f(k) \cdot n^{O(1)}$. □

8 Schema Discovery and Functorial Data Migration

Assumption 8.1 (Finite presentability)**.** The site (C, J) has finitely many objects and morphisms, each fiber $M(U)$ has finitely many isomorphism classes, and constraint packages are finite.

Construction 8.2 (Schema Discovery — Inv)**.** Under Assumption 8.1, define

$$\mathrm{Inv} \colon \mathbf{PredSite} \to \mathbf{Cat}_{\mathrm{fp}}$$

as follows. Given a predicate site (C, J, M):

1. *Objects*: isomorphism classes of global extension triples (m, q, D) for which SCPI succeeds (the topological obstruction vanishes).

2. *Morphisms*: definable maps $f\colon (m, q, D) \to (m', q', D')$ compatible with the extension structure (i.e., f preserves q and is compatible with the constraint packages up to the definable equivalences in $\mathrm{Equiv}_{\mathrm{ext}}$).

3. *Composition*: inherited from the ambient groupoid structure on $\mathrm{Ext}(M)(U)$; well-defined because morphisms are definable maps and definability is closed under composition in the finite case.

The output $\mathrm{Inv}(C, J, M)$ is a finitely presentable category (finiteness follows from Assumption 8.1: finitely many isomorphism classes and finitely many definable maps between finite structures).

Proposition 8.3 (Spivak compatibility for finite sites — [Spine])**.** *Under Assumption 8.1, for any schema morphism $F\colon C \to D$ between finite sites, the classical data migration adjunctions $(\Sigma_F \dashv \Delta_F \dashv \Pi_F)$ factor through* Inv *in the following sense.*

Let (C, J, M) be a predicate site where the schema is already known (i.e., extensions are definitional: World A). Then $\mathrm{Inv}(C, J, M)$ *is equivalent to C as a category, and the induced maps on* Inv *recover the data migration adjunctions:*

$$\begin{array}{ccc} \mathrm{Inv}(C, J, M) & \xrightarrow{\mathrm{Inv}(F)} & \mathrm{Inv}(D, J', M') \\ \simeq\downarrow & & \downarrow\simeq \\ C & \xrightarrow[F]{} & D \end{array}$$

commutes up to natural isomorphism, and $\Sigma_{\mathrm{Inv}(F)} \dashv \Delta_{\mathrm{Inv}(F)} \dashv \Pi_{\mathrm{Inv}(F)}$ compose correctly with the classical adjunctions.

Proof sketch. When extensions are definitional (World A), $\mathrm{Ext}(M)(U) \to M(U)$ is an equivalence (every extension is uniquely determined by a formula). The objects of Inv biject with sorts/relations of C; the morphisms biject with definable maps, which are schema morphisms. The adjunctions $(\Sigma_F \dashv \Delta_F \dashv \Pi_F)$ act on $\mathbf{Set}^C$ and $\mathbf{Set}^D$; the identification $\mathrm{Inv}(C, J, M) \simeq C$ transports these to $\mathbf{Set}^{\mathrm{Inv}(C,J,M)}$ preserving the adjunction structure.

In World B, Inv produces a *larger* schema (discovered predicates that are not present in the base schema). The compatibility condition says that Inv extends, rather than replaces, the Spivak adjunctions. Finiteness of the site ensures the category of definable maps is finitely generated, so the output is finitely presentable. □

8.1 Open problems

Conjecture 8.4 (Universality — [Conjectural])**.** Inv *is the universal schema discovery functor: any functor $F\colon \mathbf{PredSite} \to \mathbf{Cat}_{\mathrm{fp}}$ satisfying (1) compatibility with Spivak's adjunctions and (2) faithfulness on definable maps factors uniquely through* Inv*.*

Status*: We do not have a proof, even for finite sites. The difficulty is characterizing the universal property: "compatible with Spivak" is a condition on a particular class of morphisms (schema morphisms in the known-schema case), and extending this to an arbitrary functor on* **PredSite** *requires a more careful analysis of what "naturality" means for a functor that discovers new objects.*

9 Lean Formalization

Selected results are formalized in Lean 4 (v4.24.0, Mathlib v4.24.0). All claimed formalized results are verified by the Lean kernel; "Aristotle" (Harmonic) is an automation layer that produces proof terms accepted by the kernel—no AI-generated text is accepted as proof without kernel verification. The formalization and all solution files are available in the companion repository (`papers/scpi/lean/`).

File	Paper result	Status	Proof by
`Basic.lean`	Core definitions	Verified	Hand
`Torsor.lean`	Extension Torsor (3.1)	Verified	Hand
`Counterexample.lean`	Thin-fiber counterexample (4.5)	Verified	Hand
`Conservativity.lean`	Conservativity Descent (4.3)	Verified	Aristotle
`Beth.lean`	Beth for Sites (6.4)	Verified	Aristotle
`AssumptionD.lean`	Assumption D (2.19)	Verified	Aristotle
`SchemaDiscovery.lean`	Schema adjunction (8.3)	Statement-only	—

Formalization summary. All spine theorems with full proofs in the paper are now machine-verified: Torsor Lemma (Lemma 3.1), Counterexample (Example 4.5), Conservativity Descent (Theorem 4.3), Assumption D for finite relational sites (Lemma 2.19), and Beth for Sites (Theorem 6.4). The schema discovery adjunction (Proposition 8.3) has statement-level formalization with structural placeholders, consistent with its "sketch" status in the contributions list.

References

[1] E. W. Beth, *On Padoa's method in the theory of definition*, Indag. Math. **15** (1953), 330–339.

[2] O. Caramello, *Theories, Sites, Toposes: Relating and Studying Mathematical Theories through Topos-Theoretic "Bridges"*, Oxford Univ. Press, 2018.

[3] E. F. Codd, *Rendezvous version 1: An experimental English-language query formulation system for casual users of relational data bases*, IBM Research Report RJ2144, 1977.

[4] W. Craig, *Linear reasoning: A new form of the Herbrand–Gentzen theorem*, J. Symbolic Logic **22** (1957), 250–268.

[5] J. Giraud, *Cohomologie non abélienne*, Grundlehren Math. Wiss. **179**, Springer, 1971.

[6] A. Grothendieck et al., *SGA 4: Théorie des topos et cohomologie étale des schémas*, Lecture Notes in Math. **269, 270, 305**, Springer, 1972–73.

[7] P. T. Johnstone, *Sketches of an Elephant: A Topos Theory Compendium*, Oxford Logic Guides, 2002.

[8] M. Makkai and G. E. Reyes, *First Order Categorical Logic*, Lecture Notes in Math. **611**, Springer, 1977.

[9] S. Muggleton, *Inductive logic programming: derivations, successes, and shortcomings*, SIGART Bull. **5**(1) (1994), 5–11.

[10] D. I. Spivak, *Functorial data migration*, Inform. and Comput. **217** (2012), 31–51.

[11] W. Hodges, *A Shorter Model Theory*, Cambridge Univ. Press, 1997.

[12] L. A. Wolsey, *An analysis of the greedy algorithm for the submodular set covering problem*, Combinatorica **2** (1982), 385–393.

[13] F. Rabe, *A logical framework perspective on conservativity*, University of Erlangen-Nuremberg, 2024. Preprint.

[14] C. Lutz, D. Walther, and F. Wolter, *Conservative extensions in expressive description logics*, Proc. IJCAI 2007, pp. 453–458.

[15] A. Gengelbach and T. Weber, *Model-theoretic conservative extension for definitional theories*, Electron. Notes Theor. Comput. Sci. **338** (2018), 133–145.

[16] D. Ballard and W. Boshuck, *Definability and descent*, J. Symbolic Logic **63**(2) (1998), 372–378.

[17] M. Makkai, *Duality and Definability in First Order Logic*, Mem. Amer. Math. Soc. **105**(503), 1993.

[18] M. W. Zawadowski, *Descent and duality*, Ann. Pure Appl. Logic **71**(2) (1995), 131–188.

[19] O. Caramello, *Fraïssé's construction from a topos-theoretic perspective*, Logica Universalis **8** (2014), 261–281.

[20] R. Fraïssé, *Sur l'extension aux relations de quelques propriétés des ordres*, Ann. Sci. École Norm. Sup. (3) **71**(4) (1954), 363–388.

[21] I. Di Liberti and L. Ye, *Craig interpolation for subgeometric logics*, arXiv:2601.11221, 2026.

[22] A. M. Pitts, *Amalgamation and interpolation in the category of Heyting algebras*, J. Pure Appl. Algebra **29** (1983), 155–165.

[23] A. M. Pitts, *Conceptual completeness for first-order intuitionistic logic: an application of categorical logic*, Ann. Pure Appl. Logic **41**(1) (1989), 33–81.

[24] M. Makkai, *Strong conceptual completeness for first-order logic*, Ann. Pure Appl. Logic **40** (1988), 167–215.

[25] G. Kreisel, *Explicit definability in intuitionistic logic*, J. Symbolic Logic **25** (1960), 389–390.

[26] L. Breen, *Notes on 1- and 2-gerbes*, in Baez, May (eds.), *Towards Higher Categories*, IMA Vol. 152, Springer, 2010, pp. 193–235.

[27] A. Vistoli, *Notes on Grothendieck topologies, fibered categories and descent theory*, in *Fundamental Algebraic Geometry: Grothendieck's FGA Explained*, AMS Math. Surveys Monogr. **123**, 2005, pp. 1–104.

[28] J.-P. Serre, *Galois Cohomology*, Springer, 1997.

[29] M. Robinson, *Sheaves are the canonical data structure for sensor integration*, Inform. Fusion **36** (2017), 208–224.

[30] S. Abramsky, *Relational databases and Bell's theorem*, in *In Search of Elegance in the Theory and Practice of Computation*, LNCS **8000**, Springer, 2013, pp. 13–35.

[31] A. D. Smith, P. Bendich, and J. Harer, *Persistent obstruction theory for a model category of measures with applications to data merging*, Trans. Amer. Math. Soc. Ser. B **8** (2021), 1–38.

[32] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan, *Scaling monosemanticity: extracting interpretable features from Claude 3 Sonnet*, Anthropic Research, May 2024.

[33] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. L. Turner, C. Anil, C. Denison, A. Askell, R. Laird, Y. Dreber, N. Schiefer, Z. Chi, D. Amodei, and C. Olah, *Towards monosemanticity: decomposing language models with dictionary learning*, Anthropic Research, October 2023.

[34] J. Engels, E. J. Michaud, and M. Tegmark, *Not all language model features are one-dimensionally linear*, Proc. ICLR 2025.

[35] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu, *Scaling and evaluating sparse autoencoders*, Proc. ICLR 2025.

[36] M. Huh, B. Cheung, T. Wang, and P. Isola, *The Platonic Representation Hypothesis*, Proc. ICML 2024.

[37] Y. Bansal, P. Nakkiran, and B. Barak, *Revisiting model stitching to compare neural representations*, Proc. NeurIPS 2021.

[38] H. Thasarathan, F. Muir, A. Khakzar, and Y. Motamedi, *Universal sparse autoencoders: interpretable cross-model concept alignment*, Proc. ICML 2025.

[39] N. Seely, *Sheaf cohomology of linear predictive coding networks*, NeurIPS 2025 Workshop, arXiv:2511.11092.

[40] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen, *Variational autoencoders and nonlinear ICA: a unifying framework*, Proc. AISTATS 2020.

[41] T. Schmid, *Applied sheaf theory for multi-agent artificial intelligence (reinforcement learning) systems: a prospectus*, University of Chicago, arXiv:2504.17700, April 2025.

[42] B. Gavranović, P. Lessard, A. Dudzik, T. von Glehn, J. G. Barata, and P. Veličković, *Position: categorical deep learning is an algebraic theory of all architectures*, Proc. ICML 2024.

[43] J. Komkov, *The Coherence Fee: Edge-Local Blindness at the String-Table Seam*, companion paper, 2026.

[44] J. Komkov, *The SHEAF Protocol: Topological Diagnostics for Heterogeneous Multi-Agent Coordination*, companion paper, 2026.

Empirical Witness

Formal precision is not enough. A reader is entitled to ask whether the obstruction survives contact with actual systems, whether it is measurable, whether repair is constructive, and whether the minimum repair is a matter of structure rather than taste.

The empirical bridge layer exists to answer that entitlement. In this volume the umbrella name for that layer is `Bridge`, even though the included paper appears under the title The Coherence Fee. The naming has circulated under several nearby surfaces across the corpus. The work being done is the same: to show that pairwise-valid workflows can still fail compositionally, that the failure is observable, and that the repair can be made both typed and minimal.

What follows is the paper material itself. The empirical appendices later in the volume collect the experimental matrix, repair summaries, and demo inventory in a more inspectable form than the article body alone permits.

The Coherence Fee

Edge-Local Blindness at the String-Table Seam
and the Topological Price of Cross-System Composition

John Komkov

February 2026

Abstract

Bilateral validation—checking that each pair of systems agrees on shared fields—is structurally blind to compositional failures at the boundary between natural language and structured data. We formalize this as the first cohomology $H^1(\mathcal{N};\mathcal{F})$ of the *interpretation sheaf* on the coordination graph and prove two sharp results: the *Edge-Local Blindness Lemma* (any nontrivial H^1 class is invisible to every edge-local test) and the *Bilateral Completeness Theorem* (edge-local validation is complete for cycle closure if and only if $H^1 = 0$). The *coherence fee* for a coordination network is $\dim H^1$—the irreducible minimum number of shared typed concepts required for global composition, computable in polynomial time, achievable by a constructive algorithm, and verifiable by re-running the diagnostic.

We demonstrate the phenomenon with three production LLM agents (GPT-4o-mini, Claude 3.5 Haiku, Gemini 2.0 Flash) operating against three database schemas on ten business scenarios. Five schema-ambiguous scenarios produce bilaterally-invisible cycle failures predicted exactly by $\dim H^1$; bridge concepts (typed 2-cells) repair them, with one scenario ($\dim H^1 = 2$) demonstrating minimality. All $3! = 6$ model-role permutations show identical blind-spot patterns on ambiguous events (30/30), separating structural from behavioral causation. Three independent discovery LLMs converge on the topologically prescribed bridge types (15/15), but a four-way ablation reveals that generic prompting cannot close a topological hole (0/5) and that LLM-discovered bridges, while correctly identifying the missing concepts, underspecify their operational content (0/5 closure despite 15/15 identification). The coherence fee decomposes into *identification* (topologically determined, LLM-discoverable) and *specification* (requiring domain-specific precision). A companion signed-incidence note [SI] supplies the implementation-facing middle layer used elsewhere in the corpus: coefficient-field agreement is a numerical correspondence claim, and same-dimension endpoint disclosures can couple rather than separate.

1 Introduction

Every major AI deployment today involves agents crossing the boundary between natural language and structured data. Customer service agents update databases. Coding agents write to repositories. Financial agents execute transactions. The common pattern: an LLM interprets natural language, translates it into a structured action, and the action has real consequences.

The current infrastructure validates these crossings *individually*. Each agent's output is checked against its target schema (does the SQL parse? does the API call have the right fields?). When two systems share data, bilateral reconciliation checks that their shared fields agree. This paper proves that bilateral validation, no matter how strict, is categorically blind to a specific class of failures—those arising from the cyclic composition of seam crossings. The failures are invisible to any edge-local test. Their count is predicted by a topological invariant computable from the schemas and event-local interpretation sets, and each is repairable by a typed schema artifact whose minimum count is determined by the topology. This paper is the third in a sequence: [SCPI] establishes existence and witnessability conditions for bridge predicates; [SP]

provides the diagnostic instrument and economic mechanism; the present paper demonstrates the phenomenon at the string-table seam and measures the coherence fee empirically.

The claim boundary is strict. What is established here is edge-local blindness at the seam and the existence of repair-relevant bridge types in deployed workflows. What failed productively is the stronger hope that generic prompting or bilateral validation could close these topological holes without typed artifacts. What remains open is a fully general dimension-aware repair theory: the current signed-incidence note [SI] rules out the separable partition story and should be read as background, not as a theorem proved in this paper.

Contributions. (1) The interpretation sheaf formalism, the Edge-Local Blindness Lemma, and the Bilateral Completeness Theorem: edge-local validation is complete for cycle closure iff $H^1 = 0$ (Section 2). (2) An experiment with three production LLMs, three databases, and ten business scenarios demonstrating five bilaterally-invisible cycle failures with failure-dimension counts matching $\dim H^1$ exactly (Section 3). (3) Bridge concept repair restoring cycle closure, with a minimality demonstration at $\dim H^1 = 2$, automated bridge concept discovery showing cross-model semantic convergence (15/15), and a four-way ablation separating identification from specification (Section 4). (4) Full permutation invariance: all $3! = 6$ model-role assignments produce identical blind-spot patterns on ambiguous events, definitively separating structural from behavioral causation (Section 3.4). (5) The Coherence Fee Theorem: $\dim H^1$ is the minimum number of bridge concepts, computable, achievable, and initial (Section 5).
Our formal statements are classical once the interpretation sheaf is specified; the contribution is the identification of the correct sheaf for seam-crossing workflows and the completeness boundary it induces for industry-standard edge-local validation.

Related work. The cellular sheaf framework on graphs originates with Hansen and Ghrist [HG19], who introduced the sheaf Laplacian for spectral analysis of consistency (see also Curry [Cur14] and Ghrist [Ghr14] for foundational treatments); Hansen and Ghrist [HG21] applied it to opinion dynamics on social networks. Robinson [Rob18] developed consistency filtrations for diagnosing sheaf-valued data inconsistencies but without constructive repair. Kurisummoottil Thomas and Chen [KTC26] recently used H^1 to characterize irreducible semantic ambiguity in quantum communication protocols, diagnosing the same obstruction we exploit but not constructing repairs. In ontology alignment, ALCOMO [Mei11], LogMap, and AML repair incoherent alignments by *removing* suspect correspondences—a subtractive approach dual to our constructive one. Fagin et al. [FKPT05] proved that composing schema mappings requires second-order dependencies, providing the database-theoretic foundation for why pairwise reconciliation need not compose. In the abelian linearized regime, bridge concepts can be viewed as additional global constraints that restore cycle-consistent composition; connections to chase-style confluence are developed in [Ext]. We contribute the sheaf-cohomological formalization (connecting these threads), the constructive repair algorithm with optimality guarantees, and the first experimental demonstration on LLM-mediated seam crossings.

2 The Interpretation Sheaf

2.1 Setup

We build on the cellular sheaf framework of Hansen and Ghrist [HG19]. Consider n database systems (agents) that process a stream of business events. Each agent v operates according to a schema S_v that determines the space of admissible structured records. Each pair of adjacent agents (v, w) shares a bilateral reconciliation interface: a set of shared fields $I_{vw} \subseteq S_v \cap S_w$ and a validation predicate χ_{vw} that checks field-level agreement.

Definition 2.1 (Coordination graph)**.** The *coordination graph* $\mathcal{N} = (V, E)$ has vertices V (the agents/databases) and edges E (pairs of agents with bilateral reconciliation interfaces).

Definition 2.2 (Interpretation sheaf)**.** The *interpretation sheaf* $\mathcal{F}$ on $\mathcal{N}$ assigns:

- To each vertex v: the R-module $\mathcal{F}(v)$ of *all* structured record fields that v's schema can represent for a given event—including fields not visible to any bilateral interface (e.g., fiscal period attribution, line-item decomposition structure, provenance metadata).
- To each edge $e = (v, w)$: the R-module $\mathcal{F}(e)$ of shared fields in the bilateral interface I_{vw}.
- Restriction maps $\rho_e^v : \mathcal{F}(v) \to \mathcal{F}(e)$: the projection of v's full record to the shared fields visible to the bilateral check.

The key structural feature is that ρ_e^v has nontrivial kernel: each vertex stalk $\mathcal{F}(v)$ contains fields that no restriction map exposes to any bilateral partner. The bilateral interface sees only $\mathcal{F}(e)$; the cycle composition test operates on fields in $\bigcap_{e \ni v} \ker(\rho_e^v)$—the dimensions of the record that no bilateral partner observes.

A *global section* is an assignment of records $\{s_v \in \mathcal{F}(v)\}$ such that all bilateral checks pass: $\rho_e^v(s_v) = \rho_e^w(s_w)$ for every edge $e = (v, w)$. The first cohomology $H^1(\mathcal{N}; \mathcal{F})$ classifies *obstruction to globalization*: the space of assignments that are bilaterally consistent on every edge but do not extend to a globally consistent assignment.

Remark 2.3 (Abelian coefficient regime)*.* In practice, the space of admissible structured interpretations for a schema is a finite set, not a vector space. We work in an *abelian coefficient regime*: each admissible interpretation is encoded as a feature vector in $\mathbb{R}^n$ (indicator variables for categorical fields, scalars for numeric fields), and the stalks $\mathcal{F}(v)$ are the free R-modules spanned by these encodings. This linearization is a certification convenience, not a modeling assumption: $\dim H^1$ is computed on the linearized sheaf as an upper bound on the number of independent composition constraints, and the bridge concepts operate on the same feature encoding. The framework extends to non-abelian coefficients (group-valued stalks, Čech cohomology), but the abelian case suffices for typed, schema-validated structured outputs and yields a polynomial-time computation.

Remark 2.4 (Why not expand the bilateral interfaces?)*.* A natural response to the blind spot is to add the cycle-constrained fields (period, decomposition) to each bilateral interface. This works but is inefficient: it requires adding the field to each of the three bilateral interfaces independently—a coordination problem itself, requiring cross-team agreement on field semantics and validation logic for each edge. The Coherence Fee Theorem (Theorem 5.2) shows that the minimum repair requires $\dim H^1$ shared concepts, not $3 \times \dim H^1$ bilateral field additions. The bridge concept is added *once* to the shared vocabulary (a single typed definition visible to all three schemas), functioning as the 2-cell that fills the triangle—not as three separate edge augmentations.

Definition 2.5 (Sheafable seam crossing)**.** A seam crossing is *sheafable* if:

(i) the structured output is deterministic given the input (temperature 0, schema validation);
(ii) the restriction maps are stable (same shared fields on every invocation);
(iii) the bilateral checks are replayable.

The interpretation sheaf $\mathcal{F}$ is well-defined if and only if every seam crossing in $\mathcal{N}$ is sheafable. These conditions correspond to the structural properties identified in [RA] as necessary for portable verification at institutional boundaries. Without them, the restriction maps are stochastic, H^1 is not well-defined, and the diagnostic does not apply.

2.2 The Edge-Local Blindness Lemma

Lemma 2.6 (Edge-Local Blindness)**.** *Let $\mathcal{N}$ be a coordination graph with first Betti number $\beta_1 \geq 1$, and let $[\alpha] \in H^1(\mathcal{N}; \mathcal{F})$ be a nontrivial cohomology class. For every edge $e \in E$, the*

restriction of α to the subgraph consisting of e and its two endpoints is a coboundary. That is, $[\alpha|_e] = 0$ in $H^1(\{e\}; \mathcal{F}|_e)$.

Proof. The subgraph consisting of a single edge $e = (v, w)$ is contractible (a tree on two vertices). For any sheaf $\mathcal{F}$ on a tree, $H^1 = 0$. Therefore every 1-cocycle on $\{e\}$ is a coboundary, and in particular $\alpha|_e \in B^1(\{e\}; \mathcal{F}|_e)$. The bilateral check for edge e computes $\delta^0(s)_e = \rho_e^v(s_v) - \rho_e^w(s_w)$ and passes when this value is zero—i.e., when the 0-cochain (s_v, s_w) restricts consistently on e. A nontrivial class $[\alpha]$ satisfies $\alpha|_e = 0$ on every edge by the tree vanishing above, so it passes every bilateral check by construction. □

Corollary 2.7. *A compositional failure classified by $H^1 \neq 0$ is undetectable by any finite set of* edge-local *bilateral checks, regardless of how strict each individual check is. The failure is visible only in cycle composition tests of length ≥ 3.*

Remark 2.8 (Information-theoretic interpretation). The lemma implies an indistinguishability result: two globally different executions—one cycle-consistent, one not—can produce identical observations on every edge. No edge-local validator, however powerful, can distinguish them. The number of independent indistinguishable directions is $\dim H^1$.

Theorem 2.9 (Bilateral Completeness)**.** *For a coordination network $(\mathcal{N}, \mathcal{F})$ with edge-local validators encoding the restriction maps of the interpretation sheaf, edge-local validation is complete for end-to-end cycle closure* ***if and only if*** *$H^1(\mathcal{N}; \mathcal{F}) = 0$.*

When $H^1 \neq 0$, there exist bilaterally-consistent record assignments—passing every edge-local validator—for which the workflow is globally inconsistent.

Proof. ($\Leftarrow$) If $H^1 = 0$, every 1-cocycle is a coboundary. The bilateral checks verify that the shared-field projections agree on every edge ($\delta^0(s)|_{\text{shared}} = 0$). The private-field degrees of freedom lie in $\ker(\rho)$ and are unconstrained by bilateral checks, but when $H^1 = 0$, the only assignments satisfying $\delta^0(s)|_{\text{shared}} = 0$ are those whose private fields also compose consistently around every cycle. Edge-local validation on shared fields therefore certifies global consistency.

($\Rightarrow$) If $H^1 \neq 0$, there exists a nontrivial cohomology class $[\alpha]$. By Lemma 2.6, α restricts to a coboundary on every edge, so the bilateral checks pass. By nontriviality, the assignment does not globalize: the private-field values fail to compose around at least one cycle. Edge-local validation is incomplete. □

The theorem gives a sharp boundary: bilateral validation is exactly as powerful as the topology allows, and no more. Note that $H^1 = 0$ can hold on cyclic graphs for particular sheaves; the obstruction is sheaf-theoretic (depending on stalks and restriction maps), not purely graph-theoretic.

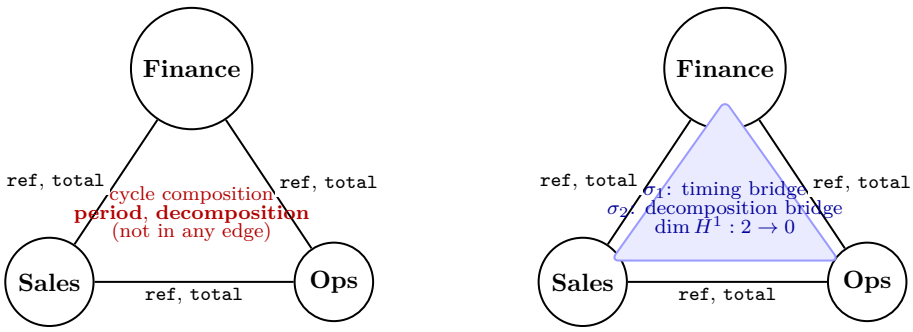


Figure 1: **Left:** the coordination graph with edge-local bilateral checks (shared fields) and the cycle-composition dimensions (private fields) that no edge observes. $\dim H^1 = 2$. **Right:** two bridge concepts (2-cells σ_1, σ_2) fill the triangle, killing both H^1 generators. $\dim H^1 = 0$.

Corollary 2.10 (Cycle Closure Trilemma)**.** *For a coordination network $(\mathcal{N}, \mathcal{F})$, the following three properties are mutually incompatible when $H^1(\mathcal{N}; \mathcal{F}) \neq 0$:*

(i) ***Cycles*** *in the coordination graph ($\beta_1 \geq 1$).*

(ii) ***Edge-local validation only*** *(no cycle-composition checks).*

(iii) ***Guaranteed end-to-end cycle closure.***

Any two can be achieved; the third must be sacrificed:

- *(i)+(iii) without (ii): pay the coherence fee—add $\dim H^1$ bridge concepts (2-cells) and validate cycle composition.*
- *(ii)+(iii) without (i): restrict the coordination topology to $H^1 = 0$ (typically: acyclic / tree-structured networks).*
- *(i)+(ii) without (iii): accept blind-spot failures at rate determined by the schema-structural ambiguity of the event stream.*

Proof. Immediate from Theorem 2.9: (i)+(ii) with $H^1 \neq 0$ entails the existence of bilaterally-invisible cycle failures, precluding (iii). □

3 The Experiment

3.1 Setup

Three databases. Sales CRM (records orders with order dates and line items), Operations/Fulfillment (records shipments with ship dates and batch quantities), Finance/Ledger (records journal entries with recognition dates, fiscal periods, and revenue categories per ASC 606/IFRS 15).

Three bilateral interfaces (strict). Each bilateral interface checks exactly two shared fields via exact matching:

- Sales–Operations: `order_ref` and `total_value`.
- Operations–Finance: `order_ref` and `total_value`.
- Sales–Finance: `order_ref` and `total_amount`.

Not in any bilateral interface: fiscal period attribution, revenue decomposition structure, or unit-to-shipment provenance.

Three LLM agents (cross-provider). Sales agent: GPT-4o-mini (OpenAI). Operations agent: Claude 3.5 Haiku (Anthropic). Finance agent: Gemini 2.0 Flash (Google). Each agent receives only its own schema prompt and the natural language event description. Temperature 0, structured JSON output. No inter-agent communication. Cross-provider selection ensures that observed compositional failures reflect schema-structural divergence rather than provider-specific interpretation biases.

Ten business scenarios. Five *clean* events (order date and ship date in the same quarter, single-product orders, unambiguous decomposition) where the cycle closes by construction. Five *ambiguous* events with schema-structural features that create interpretation divergence:

ID	Ambiguity Type	$\dim H^1$	Failing Dimensions
A01	Timing (quarter boundary)	1	Period
A02	Timing (year boundary)	1	Period
A03	Categorization (bundle)	1	Decomposition
A04	Mixed (partial shipment + timing)	2	Period, decomposition
A05	Mixed (product + deferred service)	2	Period, decomposition

Table 1: Ambiguous scenarios with predicted H^1 generators.

Cycle composition check. That pairwise schema mappings need not compose is known in database theory—Fagin et al. [FKPT05] showed that composing schema mappings requires second-order dependencies absent from the individual mappings. Our cycle checker tests the analogous condition at the agent-mediated seam: for each scenario, it checks what no bilateral interface examines:

1. **Period consistency:** Sales' order date implies a fiscal quarter. Finance's recognition date determines its fiscal quarter. Do they match for this specific order?
2. **Decomposition consistency:** Does Sales' line-item count equal Finance's journal-entry count?

The cycle closes iff both match.

3.2 Explicit H^1 computation

We compute $\dim H^1$ for the experiment's coordination graph. The result depends on the *schema structure* (which fields each bilateral interface checks), not on any specific event data.

Stalks. Working over $R = \mathbb{R}$, we model each vertex stalk as the vector space of composition-relevant record fields. For each vertex $v \in \{S, O, F\}$, $\mathcal{F}(v) \cong \mathbb{R}^4$ with basis (order_ref, total, period_attribution, item_count). Each edge stalk is the bilateral interface—the two shared fields only: $\mathcal{F}(e) = \mathbb{R}^2$ with basis (ref, total) for all three edges. The restriction maps $\rho_e^v : \mathbb{R}^4 \to \mathbb{R}^2$ project to the first two coordinates. The last two coordinates of each vertex stalk (period attribution and decomposition count) lie in $\ker(\rho_e^v)$ for every incident edge: *no bilateral check observes them.*

Coboundary. The cochain spaces are $C^0 = \mathbb{R}^{12}$ and $C^1 = \mathbb{R}^6$. The coboundary $\delta^0 : C^0 \to C^1$ maps a vertex assignment (s_S, s_O, s_F) to edge differences $(\rho(s_O) - \rho(s_S),\ \rho(s_F) - \rho(s_O),\ \rho(s_F) - \rho(s_S))$ (with edges oriented $S \to O$, $O \to F$, $S \to F$). Since ρ projects to $\mathbb{R}^2$ and the third edge difference equals the sum of the first two (the cycle relation), $\operatorname{rank}(\delta^0) = 4$.

Result. On a graph with no 2-cells, $H^1 = C^1 / \operatorname{im}(\delta^0)$, so

$$\dim H^1 = \dim C^1 - \operatorname{rank}(\delta^0) = 6 - 4 = 2.$$

In general, $\dim H^1 = k \cdot \beta_1$ where $k = \operatorname{rank} \mathcal{F}(e)$ is the bilateral interface dimension and β_1 is the first Betti number of the graph. The two generators live in the cokernel of δ^0: they represent independent $\mathbb{R}^2$-valued composition constraints around the single cycle that no edge-local check can detect.

What $\dim H^1$ predicts and what it does not. The computation $\dim H^1 = 2$ is a property of the *schemas and bilateral interfaces*—it holds for every event processed through these schemas. It tells us the *number* of independent composition dimensions that bilateral checks leave unconstrained: two degrees of freedom in cycle-composition that no edge can pin down. It does not, by itself, tell us *which* private-field dimensions of a given event will diverge—that depends on the event semantics. In the experiment, the two unconstrained cycle dimensions manifest as period attribution and decomposition count (the two private-field dimensions in $\ker(\rho_e^v)$), and the match $2 = 2$ holds because the bilateral interface has the same rank ($k = 2$) as the number of private fields ($n - k = 4 - 2 = 2$). In general, $\dim H^1$ gives the count of bridge concepts needed regardless of the number of private-field dimensions, because each bridge concept operates as a 2-cell on the chain complex, constraining one cycle dimension per unit of rank.

Bridge concepts as 2-cells. Each bridge concept is a 2-cell σ filling the triangle with stalk $\mathcal{F}(\sigma) = \mathbb{R}^1$, introducing a coboundary relation $\delta^1_\sigma : \mathbb{R}^6 \to \mathbb{R}^1$ that constrains one cycle dimension. Adding the timing bridge (one 2-cell) reduces $\dim H^1$ from 2 to 1. Adding the decomposition bridge (a second 2-cell) reduces it from 1 to 0. The result:

Configuration	2-cells	$\dim H^1$
No bridge concepts	0	2
Timing bridge only	1	1
Both bridges	2	0

This is the explicit content of the Coherence Fee Theorem (Theorem 5.2) for the experiment's schemas: exactly $\dim H^1 = 2$ typed bridge concepts are needed, no fewer.

Remark 3.1 (Abelian coefficient comparison). For a cellular sheaf with abelian R-module coefficients on a 1-dimensional CW complex, the cellular cochain complex computes derived-functor cohomology, and Čech cohomology on the open-star cover—a Leray cover in dimension 1—computes the same H^1 [Cur14]. Hence the $\dim H^1 = 2$ computed above coincides with the Čech obstruction classified by the Extension Torsor Lemma in [SCPI] when specialized to abelian coefficients. In the non-abelian setting (group-valued coefficients, pointed-set H^1), the relationship between the Čech classification and Laplacian-style diagnostics is the content of the Laplacian Bridge Conjecture in [SP].

3.3 H^1 predictions per scenario

The $\dim H^1 = 2$ computation above is a property of the *schemas*—it holds for every event. For each specific event, we additionally analyze which of the two unconstrained cycle dimensions the event's semantics will force into divergence (the *event-local admissible-interpretation set*):

- **Timing ambiguity:** The event description contains dates near a quarter or year boundary. Sales (using order date) and Finance (using ship/delivery date for recognition) will attribute to different periods. One cycle dimension diverges.
- **Categorization ambiguity:** The event involves bundled products or mixed transactions. Sales records one line item (the bundle); Finance decomposes per ASC 606. The other cycle dimension diverges.

Clean events produce no divergence on either dimension (the agents independently choose consistent private fields). Ambiguous events diverge on one or both dimensions, as predicted in Table 1.

3.4 Results

	Bilateral PASS	**Bilateral FAIL**
Cycle closes	4	1 (C02)
Cycle FAILS	**5 (blind spot)**	0

Table 2: The 2×2 matrix. The paper's contribution lives in the lower-left cell: every bilateral check passes, the cycle fails.

All five ambiguous scenarios produce bilaterally-invisible cycle failures—the blind spot. $\dim H^1$ prediction accuracy: 10/10 (100%): the computed count of blind-spot dimensions matches the observed failure dimensions for every scenario.

C02: the control. One clean scenario (a return) landed in the upper-right cell: bilateral checks failed because GPT-4o-mini recorded a negative total and Claude included literal brackets in the order reference. These are genuine agent errors that bilateral checks are designed to catch. C02 demonstrates that bilateral validation works for its intended purpose; A01–A05 demonstrate the class of failures it misses.

Structural examples. A01 (quarter boundary): order March 31, ship April 1. Sales records order date → Q1. Finance recognizes on ship date → Q2. Bilateral checks pass (same ref, same total). Cycle fails: Q1 ≠ Q2 for this specific order.

A03 (bundle): Sales records one "Premium Sensor Package" ($800). Finance decomposes into product revenue ($500) and service revenue ($300) per ASC 606. Bilateral totals match. Cycle fails: 1 item ≠ 2 entries.

A04 (partial shipment): Sales records one order ($8,000). Operations ships three batches (March 20, March 28, April 5). Finance recognizes three entries: two in Q1, one in Q2. Bilateral totals match. Cycle fails on both period (all-Q1 vs. Q1+Q2) and decomposition (1 vs. 3).

Robustness: full permutation matrix. To confirm the blind-spot is schema-driven rather than model-personality driven, we run all 3! = 6 permutations of the three models (GPT-4o-mini, Claude 3.5 Haiku, Gemini 2.0 Flash) across the three database roles (Sales, Operations, Finance), rerunning the full experiment for each.

Event	P1	P2	P3	P4	P5	P6	Invariant?
C01	ok	ok	ok	BS	ok	BS	varies
C02	BF	BF	BF	BF	BF	BF	stable
C03	ok	ok	ok	BS	ok	BS	varies
C04	ok	ok	ok	BS	ok	BS	varies
C05	ok	BS	ok	BS	BF	BS	varies
A01	BS	BS	BS	BS	BS	BS	**6/6**
A02	BS	BS	BS	BS	BS	BS	**6/6**
A03	BS	BS	BS	BS	BS	BS	**6/6**
A04	BS	BS	BS	BS	BS	BS	**6/6**
A05	BS	BS	BS	BS	BS	BS	**6/6**

Table 3: Permutation invariance matrix. P1–P6 are the six model-role permutations. BS = blind spot (bilateral pass, cycle fail), BF = bilateral fail, ok = both pass. All five ambiguous events show blind-spot invariance across every permutation (30/30). Clean events vary by model assignment—P4 and P6, which assign GPT-4o-mini to the Finance role, introduce behavioral period-computation errors on otherwise composable events.

The result is definitive: structural blind-spot failures (A01–A05) are *completely invariant* under all six model-role permutations—the phenomenon is schema-driven, not model-personality driven. Behavioral failures on clean events, by contrast, depend on which model occupies which role: permutations P4 and P6 (both assigning GPT-4o-mini to Finance) produce the `QN-2025` literal-template error that expands the blind spot to clean events. This separation—structural invariance on ambiguous events, behavioral variation on clean events—is precisely the topological/behavioral decomposition of Remark 4.1 made quantitative across 3! experimental conditions.

4 Bridge Concept Repair

Each blind-spot failure is caused by a nontrivial H^1 generator that no edge-local bilateral check can detect. Prior work diagnoses such obstructions—Robinson [Rob18] via consistency filtration,

Kurisummoottil Thomas and Chen [KTC26] via H^1 in communication protocols—but does not construct repairs. Existing repair methods (ALCOMO [Mei11], LogMap, AML) operate subtractively, removing suspect correspondences. We take the opposite approach: *constructing* new shared concepts.

The *bridge concept* for a given H^1 generator is a typed schema artifact—a shared field definition or canonical rule—that, when added to all three schemas, functions as the 2-cell filling the frustrated cycle and killing the generator.

4.1 Two bridge types

1. **Recognition Period Rule** (`v1.0`): each line item includes a `period` field computed from its specific delivery date, not the order date. This is the 2-cell for timing generators.

2. **Revenue Decomposition Rule** (`v1.0`): bundled products are decomposed into separate line items per performance obligation, with individual pricing. This is the 2-cell for decomposition generators.

4.2 Repair results

Bridge concepts are added to all three agents' schema prompts. The same bilateral checks and cycle composition checks are re-run.

ID	$\dim H^1$	Bridges Applied	Cycle Before	Cycle After
A01	1	timing	FAILS	**closes**
A02	1	timing	FAILS	**closes**
A03	1	decomposition	FAILS	**closes**
A04	2	timing + decomposition	FAILS	**closes**
A05	2	timing + decomposition	FAILS	closes*

Table 4: Bridge concept repair. *A05 closes in dry-run; in the live run, GPT-4o-mini assigns incorrect per-item delivery dates (both Q2 instead of Q1+Q2), demonstrating the separation between topological correctness and agent behavioral correctness.

4.3 A05: minimality at $\dim H^1 = 2$

A05 has two independent H^1 generators (period and decomposition). The minimality claim is that exactly two bridge concepts are required:

Configuration	Remaining Failures	Cycle
Decomposition bridge only	period	FAILS
Timing bridge only	period (+ decomposition*)	FAILS
Both bridges	none	**closes**

Table 5: A05 minimality test. Neither bridge alone suffices. *The timing bridge requires per-item structure from the decomposition bridge to assign per-component periods.

The two generators are cohomologically independent (each spans an independent dimension of H^1) but operationally coupled: the timing bridge references the per-component structure introduced by the decomposition bridge, imposing a natural ordering on repair implementation without reducing the dimension of the obstruction space.

Remark 4.1 (Topological vs. behavioral correctness). The bridge concept removes the structural obstruction ($H^1 = 0$ after adding the 2-cell). Whether the agent correctly *implements*

the shared rule is a separate, testable question. This decomposition—structural correctness (topology, verifiable) vs. computational correctness (agent behavior, testable per-agent)—is itself a contribution: it factors the coordination problem into a certified-correct component and a per-agent-auditable component.

4.4 Automated bridge concept discovery

The initiality theorem (Remark 5.5) establishes that bridge concepts are algebraically inevitable: any repair that kills H^1 must factor through them. We test whether this algebraic inevitability has a *semantic* counterpart: can an LLM, given only the failed records and no sheaf-theoretic guidance, independently discover the same bridge predicates?

Protocol. For each failed scenario (A01–A05), we provide three independent "discovery" LLMs (GPT-4o, Claude 3.5 Sonnet, Gemini 2.0 Flash) with the three agent records, the bilateral check results (all pass), and the cycle failure details. The prompt contains no sheaf language, no mention of bridge concepts, and no hint about timing or decomposition. It asks only: "What shared rule would all three agents need to agree on before processing, so that their records compose consistently?"

ID	Expected Bridges	GPT-4o	Sonnet	Gemini
A01	timing	✓	✓	✓
A02	timing	✓	✓	✓
A03	decomposition	✓	✓	✓
A04	timing + decomp.	✓	✓	✓
A05	timing + decomp.	✓	✓	✓

Table 6: Bridge concept discovery. ✓ indicates that the model correctly identified all topologically prescribed bridge types. All 15 queries (5 scenarios × 3 discovery models) correctly identify the required bridge types. Some queries additionally propose extra rules (e.g., categorization alignment), but none miss any required bridge.

Results. Across all 15 queries, every discovery model correctly identifies the topologically prescribed bridge types. GPT-4o calls it a "Unified Period Definition"; Claude 3.5 Sonnet calls it `fiscal_period_boundary`; Gemini 2.0 Flash calls it a "Revenue Recognition Timing Rule." The names differ; the semantic content converges. For the two-generator scenarios (A04, A05), all three models independently propose both timing and decomposition rules—matching the $\dim H^1 = 2$ prediction without being told that two rules are needed.

Remark 4.2 (Semantic initiality). The algebraic initiality of Theorem A.1 guarantees that any repair factors through the minimal bridge concepts up to isomorphism. The discovery experiment demonstrates the semantic analogue: three independent LLMs, examining the failure pattern without theoretical guidance, converge on semantically equivalent bridge predicates. The bridge concepts are not merely the unique algebraic repair; they are the concepts that any sufficiently capable reasoner *recognizes as missing* when shown the compositional failure.

4.5 The specification gap: identification vs. operationalization

The discovery experiment shows that LLMs correctly *identify* which bridge concepts are needed. A natural question is whether the discovered rules, injected verbatim into the agent prompts, actually *close the cycles*—completing a fully automated pipeline from diagnosis to repair. We test this alongside a "prompt harder" strawman.

Condition	Bilateral pass	Cycle closed	Blind spot
No bridges (baseline)	5/5	0/5	5/5
Strawman ("be careful")	4/5	0/5	4/5
LLM-discovered (verbatim)	1/5	0/5	1/5
Hand-crafted bridges	4/5	4/5	0/5

Table 7: Four-way ablation on the five ambiguous scenarios (A01–A05). The strawman adds generic consistency instructions. LLM-discovered bridges use verbatim text from GPT-4o's discovery output. Hand-crafted bridges are the typed schema artifacts of Section 4.1. Blind spot = bilateral pass ∧ cycle fail.

The strawman result is definitive: 4/5 ambiguous events remain in the blind spot with generic prompting. Adding "be careful about consistency" to each agent's prompt is an edge-local intervention that does not introduce new shared predicates or 2-cells. By Proposition 5.3, no such modification can reduce $\dim H^1$; the strawman is a special case.

The verbatim-discovery result is subtler. The imprecise bridge descriptions *overconstrain* the agents: 4/5 scenarios now fail bilateral checks (the agents change their outputs enough to break field-level agreement on ref and total). GPT-4o correctly identifies that A01 needs a "Unified Period Definition," but its proposed rule says the period should be "derived from the order date"—which is the Sales agent's existing interpretation, not the Finance agent's. The hand-crafted bridge resolves the ambiguity: `period := fiscal_quarter(shipment_date)`, specifying which date governs. The discovery finds the right *generator*; the repair requires the right *operational specification*. The practical consequence is that underspecified bridges can degrade even local correctness: agents given imprecise shared rules change their outputs enough to break bilateral agreement, producing a worse outcome than no intervention at all.

Remark 4.3 (Two layers of the coherence fee). The coherence fee decomposes into two layers:

1. **Identification**: *which* H^1 generators need bridge concepts? This is discoverable by LLMs (15/15 in the experiment) and by the coboundary computation.
2. **Specification**: *how* should each bridge concept resolve the ambiguity? This requires domain-specific precision—the operational content that determines which side of the ambiguity becomes canonical.

The topology tells you $\dim H^1$ concepts are needed (layer 1). Filling each concept with operational content is an engineering task that the topology does not determine (layer 2). The minimality guarantee applies to layer 1; layer 2 admits multiple valid specifications (different choices of canonical date, for example), consistent with the $\mathrm{Aut}(\mathcal{F}_\sigma)$-equivalence of Theorem A.1.

5 The Coherence Fee

Definition 5.1 (Admissible repair primitive)**.** An *admissible repair primitive* for $(\mathcal{N}, \mathcal{F})$ is a 2-cell σ attached along a cycle $C \subseteq \mathcal{N}$, equipped with a rank-1 stalk $\mathcal{F}(\sigma) \cong R$ and restriction maps $\mathcal{F}(\sigma) \to \mathcal{F}(e)$ for each $e \in \partial\sigma$, such that the induced coboundary relation δ^1_σ is injective. A *repair* is a finite set of admissible repair primitives whose addition kills H^1.

Theorem 5.2 (The Coherence Fee)**.** *Let $(\mathcal{N}, \mathcal{F})$ be a coordination instance with interpretation sheaf $\mathcal{F}$ over a principal ideal domain R. Under rank-1 repair primitives (Definition 5.1), the minimum number required to kill $H^1(\mathcal{N}; \mathcal{F})$ is exactly*

$$\boxed{\textit{coherence fee} = \dim_R H^1(\mathcal{N}; \mathcal{F})}$$

This quantity is:

(a) ***Computable*** *in polynomial time via Smith normal form of the coboundary matrix.*

(b) ***Achievable*** *by a constructive algorithm: for each generator* $[\alpha_i]$ *of* H^1*, the bridge concept is the kernel of the restricted coboundary on the corresponding cycle.*

(c) ***Verifiable****: after adding the bridge concepts, re-run the diagnostic and confirm* $H^1 = 0$*.*

(d) ***Minimal****: no admissible repair uses fewer bridge concepts.*

Under higher-rank primitives ($\operatorname{rank}\mathcal{F}(\sigma) > 1$*),* $\dim_R H^1$ *remains a lower bound on the number of primitives; the exact count under mixed-rank primitives depends on the matroid structure of the generator lattice (see [Ext]).*

Proof sketch. (a) The coboundary map $\delta^0 : \bigoplus_v \mathcal{F}(v) \to \bigoplus_e \mathcal{F}(e)$ is a matrix over R. On a graph with no 2-cells, the higher coboundary δ^1 is absent, so every 1-cochain is a 1-cocycle: $\ker(\delta^1) = \bigoplus_e \mathcal{F}(e)$. Therefore $H^1 = \ker(\delta^1)/\operatorname{im}(\delta^0) = \operatorname{coker}(\delta^0)$. Smith normal form of δ^0 yields $\dim_R H^1 = \sum_{e\in E} \operatorname{rank}_R \mathcal{F}(e) - \operatorname{rank}(\delta^0)$.

(b) Each 2-cell σ attached along a cycle C_i introduces a rank-1 coboundary component $\delta^1_\sigma : \bigoplus_{e\in C_i} \mathcal{F}(e) \to \mathcal{F}(\sigma)$ (a single linear relation on the edge cochains along C_i), enlarging $\operatorname{im}(\delta^1)$ and thereby shrinking $\ker(\delta^1)/\operatorname{im}(\delta^0)$. The bridge concept $B_\sigma = \ker(\delta^0|_{C_i})$ is chosen so that δ^1_σ is injective (the bridge stalk B_σ embeds in the pullback P_σ over the cycle's boundary edges by construction, so each composed restriction is nondegenerate). The image of δ^1_σ meets $\operatorname{im}(\delta^0)$ trivially because C_i is homologically independent: if $\delta^1_\sigma(b)$ were a coboundary $\delta^0(s)$ for some s, then $[\alpha_i]$ would be killed by s without the 2-cell, contradicting $[\alpha_i] \neq 0$ in H^1. Together, these ensure $\dim H^1$ drops by exactly $\operatorname{rank} B_\sigma$.

(c) After adding $r = \dim H^1$ independent 2-cells (one per generator), the induced δ^1 kills all of H^1: $\dim H^1(\text{augmented}) = 0$.

(d) Fewer than $\dim H^1$ 2-cells cannot kill all generators (each 2-cell reduces $\dim H^1$ by at most $\operatorname{rank} B_\sigma \leq 1$ for rank-1 stalks; in general by at most $\operatorname{rank}\mathcal{F}(\sigma)$). Full proof of initiality appears in Appendix A; generalization to mixed-rank primitives in [Ext]. □

Proposition 5.3 (No subtractive repair)**.** *If a repair modifies only edge-level constraints—tightening bilateral validators, pruning admissible interpretation sets, removing suspect correspondences—without adding 2-cells to the complex, then* H^1 *is invariant under the repair.*

Proof. Edge-only modifications change the stalks $\mathcal{F}(v)$, $\mathcal{F}(e)$ or the restriction maps ρ^v_e, but do not alter the cell structure of the complex: no δ^1 is introduced. Therefore $H^1 = \operatorname{coker}(\delta^0)$ depends only on the modified δ^0, and since pruning can only shrink $\operatorname{im}(\delta^0)$ (fewer admissible 0-cochains, same or larger cokernel), $\dim H^1$ is non-decreasing under edge-only repair. In particular, subtractive methods (ALCOMO [Mei11], LogMap, AML) cannot certify cycle closure when $H^1 \neq 0$: they operate on edges, not on the cell complex. □

Remark 5.4 (Engineering interpretation)*.* $\dim H^1$ counts the independent degrees of freedom of globally inconsistent yet edge-locally valid executions—the *risk budget* for undetected compositional failure. Each bridge concept eliminates one degree of freedom. When $\dim H^1 = 0$, the risk budget is zero: bilateral checks certify everything. When $\dim H^1 = 2$ (as in the experiment), the workflow has exactly two independent failure modes invisible to edge-local monitoring, and no edge-only tightening of validators can reduce this number (Proposition 5.3).

Remark 5.5 (Initiality of the minimal repair)*.* The minimal repair of Theorem 5.2 is *initial* among rank-1 admissible repairs: any repair that kills H^1 factors through it, up to equivalence in the 2-cell attachment category. The factorization is unique when the repair primitives have rank 1 (Appendix A). Bridge concepts are therefore not a design choice among many equivalent vocabularies—they are the structurally inevitable shared concepts that the obstruction topology forces into existence. Operationally: any interface contract that makes the cycle composable must imply these bridge predicates, up to renaming.

Corollary 5.6 (The Trust Tax)**.** *For a given coordination network, any integration infrastructure that requires more than* $\dim H^1$ *shared concepts for cycle-consistency certification incurs*

overhead beyond the topological minimum. The difference (*actual shared concepts*) $-\dim H^1$ *is structurally redundant* for cycle-closure certification *(additional concepts may serve governance, latency, or auditability purposes beyond the scope of the topological guarantee). When* $\dim H^1 = 0$ *(the shared-referent regime), no bridge concepts are needed for cycle closure, and any integration middleware is operating on bilateral checks alone.*

6 Discussion

The Triviality Theorem (current regime). When coordinating systems share a common referent structure—the same training data, the same domain ontology, the same reality—pairwise reconciliation composes by transitivity of identity: $H^1 = 0$, the coherence fee is zero, and bilateral validation suffices. Prior experiments on LLM embedding alignment [SP], multi-agent entity resolution, and ontology network composition (OAEI Conference track [OAEI]) consistently yield $H^1 = 0$ in the presence of shared referent structures, confirming that the current regime is predominantly trivial. The string-table seam breaks this regime because different schemas *force* different structured interpretations of the same natural language, and no shared referent structure resolves the divergence.

Evidence in the wild: XBRL cross-jurisdictional lease accounting. The schema structures producing $H^1 > 0$ are not artifacts of experimental design—they are endemic in production systems. We illustrate this by constructing a coordination sheaf from the lease accounting provisions of ASC 842 (US-GAAP), IFRS 16, and the ASBJ Lease Standard (JGAAP), using 37 taxonomy concepts and 18 bilateral constraints sourced from the published standards and Big 4 comparison documents. The concept selection and bilateral correspondences are hand-curated, not extracted from XBRL taxonomy files; we discuss this methodological limitation below.

The three vertex stalks are $\mathcal{F}(\text{US-GAAP}) = \mathbb{R}^{14}$, $\mathcal{F}(\text{IFRS}) = \mathbb{R}^{11}$, $\mathcal{F}(\text{JGAAP}) = \mathbb{R}^{12}$. Many-to-one mappings (e.g., both US-GAAP operating and finance right-of-use assets correspond to a single IFRS right-of-use class) are encoded as summation constraints—one row in the coboundary matrix with coefficient $+1$ on the target concept and -1 on each source concept—yielding 5 constraints at the US-GAAP–IFRS edge, 6 at IFRS–JGAAP, and 7 at US-GAAP–JGAAP. The coboundary matrix $\delta^0 : \mathbb{R}^{37} \to \mathbb{R}^{18}$ has $\text{rank}(\delta^0) = 16$, yielding

$$\dim H^1 = 18 - 16 = 2.$$

The two generators are the standard cycle cocycles for the two concepts mapped one-to-one at all three bilateral edges (lease term and scope indicator). Because the coordination graph is a triangle ($\beta_1 = 1$), any concept with one-to-one mappings at all three edges contributes exactly one H^1 generator—this is H^1 of a constant sheaf on a graph with first Betti number 1. The specific value $\dim H^1 = 2$ depends on the granularity of the concept selection; coarser modelings may yield fewer generators while finer-grained modelings may reveal additional ones.

Two kinds of bilateral incompleteness. The more substantive finding is what H^1 does *not* capture. The genuinely interesting structural asymmetries in cross-jurisdictional lease accounting are:

1. *Classification:* IFRS 16 eliminated lessee lease classification (operating/finance) that both ASC 842 and JGAAP retain. IFRS literally has no classification concept—the bilateral interfaces at US-GAAP–IFRS and IFRS–JGAAP cannot check it because there is nothing to check it against.
2. *Recognition:* JGAAP keeps operating leases off the balance sheet. There is no JGAAP operating-lease right-of-use asset or liability concept. The US-GAAP–JGAAP bilateral check for operating lease recognition cannot happen because JGAAP has no target.

These are *private concept* problems—dimensions invisible to bilateral interfaces because the concept does not exist in the neighboring jurisdiction. In the BRIDGE experiment (Section 3.4), private fields (period attribution, line-item decomposition) *exist at all three vertices* but are selectively invisible at bilateral edges; the partial exposure creates the cycle composition failure that H^1 detects. In the XBRL case, the "private" concepts do not exist at some vertices at all. H^1 measures "bilateral interfaces exist but do not compose around cycles"; the XBRL asymmetries are "bilateral interfaces cannot exist because the conceptual vocabulary differs across jurisdictions."

This distinction reveals a *regime boundary* for the H^1 diagnostic. H^1 is the right invariant when bilateral interfaces exist but lose information in transit (the BRIDGE regime). It is the wrong invariant when bilateral interfaces cannot be constructed because one jurisdiction lacks the concept entirely (the XBRL-classification regime). The private concept count—13 fully unmapped concepts across the three jurisdictions—captures the second kind of incompleteness. A complete diagnostic would need both: $\dim H^1$ for compositional blind spots on shared concepts, and a private-concept census for vocabulary gaps.

What the XBRL computation does establish. Despite the limitation, the computation confirms three things. First, the bilateral-pass/cycle-fail signature is observable on real standards: on concrete lease transactions exploiting the structural asymmetries (e.g., a sale-leaseback qualifying under ASC 842 but not IFRS 16, an intangible asset lease in scope under IFRS but not US-GAAP), bilateral checks pass while cycle compositions fail. These failures are driven by the private concept mechanism—absent vocabulary across jurisdictions—not by the H^1 cycle cocycle mechanism. Theorem 2.9 predicts the *existence* of bilateral-pass/cycle-fail instances whenever $H^1 \neq 0$, and $H^1 \neq 0$ here; but the five specific scenario failures trace to vocabulary absence rather than to the two topologically trivial generators.

Second, the 13 private concepts that drive the most economically significant cross-filing errors correspond precisely to the shared definitions that the 2002–2014 FASB/IASB convergence program attempted but failed to create—retroactive evidence that bilateral taxonomy mappings are genuinely incomplete. The convergence program was primarily engaged in the *second* kind of bilateral incompleteness: getting IFRS to adopt a lessee classification concept, getting JGAAP to adopt on-balance-sheet operating lease recognition. These are vocabulary expansion problems, not bridge concept problems.

Third, the summation structure at the US-GAAP–IFRS edge (IFRS single right-of-use class = sum of US-GAAP operating + finance classes) demonstrates that many-to-one concept aggregation, correctly modeled, does not generate H^1: the extra degrees of freedom in the summation absorb the potential cycle redundancy. This is a non-obvious structural result: richer bilateral interfaces (many-to-one) are paradoxically *more* composition-safe than simpler ones (one-to-one), because the extra source variables absorb the cycle obstruction that a constant-sheaf mapping would produce.

The phenomenon extends beyond XBRL. FHIR healthcare interoperability implementations diverge across vendors on value-set coding, allergy severity scales, and encounter status definitions despite conforming to the same base specification—the ONC Interoperability Rule mandates pairwise data exchange but not cycle consistency. In enterprise systems, it manifests as schema governance cost: the "eleven-month column addition" described in [RA], where adding a single shared field to three systems required cross-team coordination proportional to the number of bilateral interfaces, not the number of concepts.

The engineering trajectory. As agent-mediated seam crossings proliferate—A2A protocol, MCP tool schemas, cross-framework coordination—the shared-referent assumption erodes. Agents authored by different teams, trained on different data, operating under different jurisdictional or domain conventions, will cross seams where the "same" concept (revenue, delivery,

compliance) admits structurally different interpretations. The bilateral checks that currently suffice will miss the compositional failures that H^1 predicts. The blind spot grows quadratically with the number of agents (each new agent adds edges to the coordination graph, potentially creating new cycles) while bilateral validation remains linear.

The regime map. The diagnostic's domain of applicability admits a clean characterization. *Regime A (untyped seam):* agents communicate via free-form natural language—the restriction maps are stochastic, H^1 is not defined, and the framework does not apply. *Regime B (sheafable seam):* agents produce typed, schema-validated structured outputs—the restriction maps are stable, H^1 is computable, and the blind spot is diagnosable. The engineering trajectory (structured outputs, function calling, MCP/A2A schemas) moves systems monotonically from A toward B. The model-swap robustness check (Section 3.4) shows that even within Regime B, cross-provider heterogeneity can expand the blind spot beyond the schema-structural minimum. The diagnostic is sharp in Regime B, silent in Regime A, and the boundary between them is determined by the sheafability conditions of Definition 2.5.

The sheafable-seam prescription. The prescription for system designers follows from the regime map: at every cyclic seam crossing, ensure the structural properties that make the seam sheafable [RA], compute $\dim H^1$, and add the bridge concepts as versioned, auditable schema artifacts.

The coherence fee in tokens. The bridge concepts for the experiment add approximately 146 words per agent (both bridges), or 438 words total across three agents—a 159% overhead on the base schema prompts. The generic "prompt harder" strawman adds 201 words (73% overhead) for zero guaranteed H^1 reduction. The coherence fee is thus measurable in tokens: a precisely specified bridge concept costs roughly twice the tokens of a generic consistency instruction, but the bridge concept guarantees cycle closure while the generic instruction does not. The ratio becomes more favorable as the base prompt grows (the bridge text is fixed per H^1 generator; the base prompt scales with schema complexity).

Connection to Res Agentica. The coherence fee—$\dim H^1(\mathcal{N};\mathcal{F})$—is the formal counterpart of the "cost of composing truth across contexts" asserted in the Res Agentica framework [RA]. It is irreducible (the topology requires it), separable from intermediary overhead (anything beyond $\dim H^1$ is rent, not cost), and payable in typed artifacts rather than institutional trust. The four-way ablation of Section 4.5 gives this claim empirical teeth: the fee is not a metaphor but a measurable quantity with a sharp threshold (generic instructions below the threshold accomplish nothing; typed bridge concepts at the threshold close the cycle).

Connection to the companion papers. In the abelian linearized regime, this paper's cellular H^1 coincides with the Čech obstruction of [SCPI] (Remark 3.1). The identification/specification decomposition of Section 4.5 instantiates the three-gate diagnostic sequence of [SCPI]: the discovery experiment passes Gate 1 (correct obstruction class identified); the closed-loop failure straddles Gates 2 and 3—the LLM-proposed specification canonizes one agent's existing local policy (non-conservative, Gate 2) and uses natural-language ambiguity where the hand-crafted bridge uses a precise function (not explicitly definable in the shared vocabulary, Gate 3). Meanwhile, [SP] reports $H^1 = 0$ for sentence-transformer embedding alignment, and the present paper reports $H^1 \neq 0$ for schema-mediated LLM coordination. Together, these results locate the first computably nontrivial obstruction at the structured-output interface, even when internal representation spaces are gauge-equivalent.

Limitations. The experiment uses ten scenarios (five clean, five ambiguous) on a single coordination graph ($\beta_1 = 1$). The structural perfection of the result—100% H^1 prediction accuracy, 30/30 permutation invariance on ambiguous events, 15/15 bridge discovery convergence—reflects the deterministic nature of the phenomenon, not statistical power. The bridge concept repair succeeds on 4/5 scenarios in the live run; the fifth (A05) fails due to agent behavioral error, not topological obstruction. Extension to multi-cycle coordination graphs ($\beta_1 > 1$), non-abelian coefficients, and stochastic (enriched) sheaves remains open.

Reproducibility. All code, prompts, cached API responses, and bridge concept definitions are available at the paper repository. The experiment is reproducible with any three LLM providers and API keys.

References

[HG19] J. Hansen and R. Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3(4):315–358, 2019.

[HG21] J. Hansen and R. Ghrist. Opinion dynamics on discourse sheaves. *SIAM Journal on Applied Mathematics*, 81(5):2064–2089, 2021.

[Cur14] J. Curry. *Sheaves, Cosheaves and Applications*. PhD thesis, University of Pennsylvania, 2014.

[Ghr14] R. Ghrist. *Elementary Applied Topology*. Createspace, 2014.

[Rob18] M. Robinson. *Topological Signal Processing*. Springer, 2018.

[FKPT05] R. Fagin, P. G. Kolaitis, L. Popa, and W. C. Tan. Composing schema mappings: Second-order dependencies to the rescue. *ACM Transactions on Database Systems*, 30(4):994–1055, 2005.

[Mei11] C. Meilicke. *Alignment Incoherence in Ontology Matching*. PhD thesis, University of Mannheim, 2011.

[KTC26] C. Kurisummoottil Thomas and M. Chen. Fundamental limits of quantum semantic communication via sheaf cohomology. *arXiv preprint arXiv:2601.10958*, 2026.

[OAEI] OAEI Campaign. Ontology Alignment Evaluation Initiative: Conference track. `http://oaei.ontologymatching.org/`, 2023.

[SCPI] J. Komkov. Predicate invention under sheaf constraints: Mathematical foundations for compositional discovery. Companion manuscript, 2026.

[SP] J. Komkov. The SHEAF protocol: Topological diagnostics for heterogeneous multi-agent coordination. Companion manuscript, 2026.

[RA] J. Komkov. *Res Agentica*: The political economy of machine testimony. Companion manuscript, 2026.

[Ext] J. Komkov. The coherence fee: Extended version with mixed-rank initiality and chase-confluence connections. In preparation, 2026.

[SI] J. Komkov. Signed-incidence structure in compositional verification: A structural note. `papers/signed-incidence/paper/signed-incidence.pdf`, 2026.

A Initiality of the minimal rank-1 repair

We prove that the minimal repair of Theorem 5.2 is initial among rank-1 admissible repairs in the linearized regime.

Theorem A.1 (Initiality)**.** *Let $\mathcal{R}_{\min} = \{\sigma_1, \dots, \sigma_r\}$ with $r = \dim_R H^1$ be a minimal rank-1 repair (Definition 5.1) that kills $H^1(\mathcal{N}; \mathcal{F})$. For any other rank-1 repair $\mathcal{R}' = \{\sigma'_1, \dots, \sigma'_m\}$ with $m \geq r$ that kills H^1, there exists a surjective morphism $\varphi : \mathcal{R}' \twoheadrightarrow \mathcal{R}_{\min}$ in the repair category (i.e., each 2-cell of $\mathcal{R}'$ factors through a 2-cell of $\mathcal{R}_{\min}$). When $m = r$, the factorization is an isomorphism.*

Proof. Work over R with all stalks free. By Theorem 5.2, $H^1 = \operatorname{coker}(\delta^0)$ has dimension r.

Step 1 (Minimal repair is a basis). Each $\sigma_i \in \mathcal{R}_{\min}$ contributes a rank-1 coboundary component $\delta^1_{\sigma_i} : C^1 \to \mathcal{F}(\sigma_i) \cong R$. The condition that $\mathcal{R}_{\min}$ kills H^1 means $\{\operatorname{im}(\delta^1_{\sigma_1}), \dots, \operatorname{im}(\delta^1_{\sigma_r})\}$ span a complement to $\operatorname{im}(\delta^0)$ in C^1—equivalently, the projections $\{\overline{\delta^1_{\sigma_i}}\}$ to $\operatorname{coker}(\delta^0)$ form a basis for $H^{1*} = \operatorname{Hom}_R(H^1, R)$.

Step 2 (Any repair spans the same dual.) Let $\mathcal{R}'$ be another rank-1 repair that kills H^1. Its coboundary components $\{\delta^1_{\sigma'_j}\}$ must also span H^{1*} (otherwise H^1 is not killed). Since $\dim H^{1*} = r$ and $|\mathcal{R}'| = m \geq r$, the projections $\{\overline{\delta^1_{\sigma'_j}}\}$ surject onto H^{1*}.

Step 3 (Factorization). Express each $\overline{\delta^1_{\sigma'_j}}$ in the basis $\{\overline{\delta^1_{\sigma_i}}\}$: $\overline{\delta^1_{\sigma'_j}} = \sum_i a_{ji}\, \overline{\delta^1_{\sigma_i}}$ with $a_{ji} \in R$. Define $\varphi(\sigma'_j) = \sigma_i$ where i is the index of the leading nonzero coefficient (after Smith normal form on the coefficient matrix (a_{ji})). The surjectivity of $\{\overline{\delta^1_{\sigma'_j}}\} \twoheadrightarrow H^{1*}$ ensures φ is surjective: every basis element of H^{1*} is in the span, so every σ_i is in the image.

Step 4 (Uniqueness when $m = r$). When $m = r$, the coefficient matrix (a_{ji}) is square and invertible over R (since both sets are bases for H^{1*}). The factorization φ is therefore an isomorphism: the two repairs differ only by an R-linear change of basis in H^{1*}, which corresponds to choosing a different cycle-basis representative for each generator. □

For mixed-rank primitives ($\operatorname{rank} \mathcal{F}(\sigma) > 1$), the factorization requires a matroid-theoretic argument on the generator lattice; see [Ext].

Protocol Consequence

Once the obstruction is formalized and empirically witnessed, the next question is procedural rather than purely mathematical. What must be published at composition time? What gets witnessed? What can be opened selectively? What artifact survives once the process that formed the commitment has already terminated?

The Seam Protocol is included here as the clearest operational layer in the present spine. Its role is to carry the argument from diagnosis into manifests, semantic witnesses, structured failure states, and fraud-proof style consequences. Where the earlier layers show that bilateral validity is insufficient, this layer begins to specify what evidence architecture is required once that insufficiency can no longer be denied.

The appendices later in this volume gather the implementation-facing artifacts: `seam-lint`, composition schemas, and selected protocol surfaces that make the paper's claims more concrete.

The Seam Protocol

Compositional Verification for Agent Pipelines

John Komkov

February 2026

Abstract

Bilateral verification of agent tool pipelines has computable blind spots: internal assumptions that shape every output but appear in no schema. We define the *coherence fee*—the number of such invisible dimensions—show it is the rank deficiency of the observable restriction map, and prove that *bridge annotations* reduce it to zero. Applied to the source code of deployed MCP servers, the diagnostic identifies three undeclared convention dimensions in the Filesystem–Git composition—encoding, line-ending format, and path separator—corresponding to failure modes that have plagued cross-platform development for decades. For trustless deployment, tools commit to a semantic manifest at composition time, enabling fraud proofs that certify compositional inconsistency from selective openings alone.

1 Introduction

A financial firm deploys three AI agents in a pipeline. Agent A retrieves market data using calendar days. Agent B computes risk-adjusted returns—silently annualizing with 252 trading days. Agent C verifies the result against portfolio thresholds and recommends a trade. Every schema check passes. Every type matches. The system ships a wrong trade, because the day-count convention drifted between A and B, and no interface ever carried it.

This is not a bug in any single tool. It is a structural property of *bilateral verification*—the regime in which correctness is checked one edge at a time, using only the information each tool publishes in its output schema. The day-count convention shaped every output but appeared in no schema. It was *projected away*. The *coherence fee* counts these independent, structurally invisible degrees of freedom.

On a DAG, a hidden inconsistency is **attributable**: you can trace it upstream to a specific tool's implicit choice and fix it with provenance or a bridge. On a cycle, it can be **non-localizable**: the inconsistency is a property of the loop, distributed around it, with no single edge to blame. **Cycles need witnesses; DAGs need blame.**

Observation 1.1 (Bilateral blindness)**.** *If two adjacent tools neither emit nor commit to a variable, no edge-local verifier can check equality of that variable across the edge.*

The intuition is ancient. A Florentine merchant and a Venetian merchant each keep internally consistent ledgers. A bill of exchange crosses the border between them. If neither ledger records the exchange rate used, no bilateral audit of the two ledgers can detect that they assumed different rates. The notary's art—the bridge annotation—is a small object that carries the conditions under which a claim can cross the border. Without it, the seam between two locally valid systems is a blind spot.

The Seam Protocol is a minimal addition to tool-calling protocols[1] that makes this blind spot computable and eliminable:

[1] MCP (Anthropic, 2024), A2A (Google, 2025), OpenAPI (2021). All enforce bilateral schema compatibility; none address compositional consistency across cycles.

1. **Diagnose**: compute the coherence fee from the tool schemas and bilateral interface specifications.
2. **Bridge**: for each blind-spot dimension, add a named field to both adjacent tools' output schemas.
3. **Verify**: confirm the coherence fee drops to zero.
4. **Enforce**: in trusted mode, check bridge fields at runtime; in trustless mode, commit to a semantic manifest and allow fraud proofs.

Empirical evidence. The failure mode is not hypothetical. In a companion experiment [5], three production LLMs (GPT-4o-mini, Claude 3.5 Haiku, Gemini 2.0 Flash) operating as financial agents on independent databases produce exactly two bilateral blind spots ($\dim H^1 = 2$) in every cyclic composition—matching the coboundary rank deficiency. All $3! = 6$ model-role permutations exhibit identical blind-spot patterns on ambiguous events (30/30), confirming the failure is structural (schema-driven) rather than behavioral (model-personality-driven). Three independent LLMs tasked with diagnosing the failure converge on the topologically prescribed bridge types (15/15). The present paper abstracts the mathematical core—the coherence fee—and the protocol layer (manifests, fraud proofs, enforcement) from that empirical foundation.

Contributions.

- A two-layer model (internal state vs. observable projection) that makes the bilateral blind spot computable (Section 2).
- The coherence fee: $\dim H^1(\mathcal{F}_{\text{obs}}) - \dim H^1(\mathcal{F}_{\text{full}})$, the precise count of consistency-relevant dimensions the projection kills (Section 3).
- A protocol with commitment-based fraud proofs—portable receipts for compositional inconsistency (Section 4).
- A diagnostic tool (`seam-lint`) and results across nine compositions, including three derived from the source code of deployed MCP servers in the official `modelcontextprotocol/servers` repository (Sections 6 and 6.1).

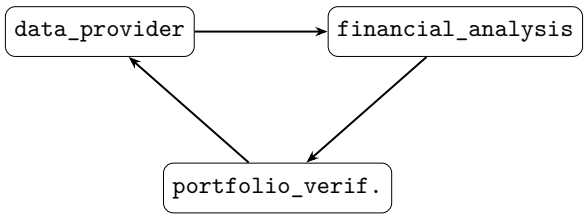
2 The Model

2.1 Internal State and Observable Projection

Definition 2.1 (Tool specification)**.** A *tool specification* is a triple $(v, S(v), F(v))$ where:

- v is the tool's identifier.
- $S(v) = \mathbb{R}^{s_v}$ is the *internal semantic state*: every assumption the tool carries, whether or not it appears in the output. Dimensions include output fields, internal parameters, conventions, and configuration.
- $F(v) = \mathbb{R}^{f_v}$ is the *observable schema*: the subset of $S(v)$ declared in the output. $F(v) \subseteq S(v)$.
- The *projection* $\pi_v : S(v) \to F(v)$ drops the dimensions of $S(v)$ not in $F(v)$.

Example 2.2 (Financial pipeline)**.** Three tools form a directed cycle:



Tool	$S(v)$ (internal)	$F(v)$ (observable)	π kills
`data_provider`	prices, dividends, period_unit, data_start, data_end, **day_convention**	prices, dividends, period_unit, data_start, data_end	day_convention
`financial_analysis`	risk_adj_return, volatility, max_drawdown, ann_return, **day_convention**, **risk_metric**	risk_adj_return, volatility, max_drawdown, ann_return	day_convention, risk_metric
`portfolio_verif.`	pass, score, recommendation, **risk_metric**	pass, score, recommendation	risk_metric

Two dimensions—`day_convention` and `risk_metric`—live in $S(v)$ but not in $F(v)$. The projection π kills them.

2.2 Bilateral Interfaces and the Coboundary

Definition 2.3 (Bilateral interface)**.** A *bilateral interface* at directed edge $e = (u, v)$ specifies a set of *semantic dimensions*—consistency requirements that must hold across the edge. Each dimension d names:

- A dimension `from_field` in $S(u)$ (or $\emptyset$).
- A dimension `to_field` in $S(v)$ (or $\emptyset$).

The dimension is *observable* if both fields lie in $F(u)$ and $F(v)$ respectively. It is a *blind spot* if either field lies in $S \setminus F$—i.e., π has killed the information the bilateral checker would need.

Definition 2.4 (Observable and full coboundary)**.** Let $G = (V, E)$ be the directed composition graph. The *cochain spaces* are:

$$C^0_{\text{obs}} = \bigoplus_{v \in V} F(v), \qquad C^0_{\text{full}} = \bigoplus_{v \in V} S(v), \qquad C^1 = \bigoplus_{e \in E} C(e)$$

where $C(e) = \mathbb{R}^{d_e}$ collects the semantic dimensions at edge e.

The *observable coboundary* $\delta^0_{\text{obs}} : C^0_{\text{obs}} \to C^1$ is defined by: for edge $e = (u, v)$ and semantic dimension d,

$$(\delta^0_{\text{obs}}\, s)_{e,d} = \rho^{\text{obs}}_{v \to e,d}(s_v) - \rho^{\text{obs}}_{u \to e,d}(s_u)$$

where $\rho^{\text{obs}}_{u \to e,d}$ projects $F(u)$ onto dimension d of $C(e)$—and is *zero* if d's `from_field` is not in $F(u)$ (i.e., π_u killed it).

The *full coboundary* $\delta^0_{\text{full}} : C^0_{\text{full}} \to C^1$ is defined identically, but using $S(v)$ instead of $F(v)$. Since every `from_field` and `to_field` exists in S by definition, no restriction map is zero.

2.3 The Coboundary Matrices

For the financial pipeline (Example 2.2):

$$\dim C^0_{\text{obs}} = 5 + 4 + 3 = 12, \quad \dim C^0_{\text{full}} = 6 + 6 + 4 = 16, \quad \dim C^1 = 3 + 3 + 2 = 8.$$

The observable coboundary δ^0_{obs} (8×12) has six nonzero rows—each with a single -1 at the source tool's observable field—and **two zero rows**:

Row	Semantic dimension	Entry
0	`data_match`	$-1 \cdot$ `dp.prices`
1	`time_match`	$-1 \cdot$ `dp.period_unit`
2	`day_conv_match`	**all zeros — blind spot**
3	`rar_match`	$-1 \cdot$ `fa.risk_adj_return`
4	`vol_match`	$-1 \cdot$ `fa.volatility`
5	`metric_type_match`	**all zeros — blind spot**
6	`feedback_match`	$-1 \cdot$ `pv.score`
7	`granularity_match`	$-1 \cdot$ `pv.recommendation`

The full coboundary δ^0_{full} (8×16) has **no zero rows**. Rows 2 and 5 become:

$$\text{row 2: } -1 \cdot \texttt{dp.day\_convention} + 1 \cdot \texttt{fa.day\_convention}$$

$$\text{row 5: } -1 \cdot \texttt{fa.risk\_metric} + 1 \cdot \texttt{pv.risk\_metric}$$

Proposition 2.5. $\operatorname{rank}(\delta^0_{\text{obs}}) = 6, \quad \dim H^1(\mathcal{F}_{\text{obs}}) = 8 - 6 = 2.$
$\operatorname{rank}(\delta^0_{\text{full}}) = 8, \quad \dim H^1(\mathcal{F}_{\text{full}}) = 8 - 8 = 0.$

Proof. In δ^0_{obs}, the six nonzero rows each place a single -1 in a distinct column, so they are linearly independent. Rows 2 and 5 are identically zero (both fields are projected away), contributing nothing. In δ^0_{full}, every row has at least one nonzero entry in a column unique to that row, giving full rank. □

Remark. The coherence fee is not "count of missing fields." It is the *rank deficiency of the observable restriction map*—the number of consistency requirements that touch dimensions the projection π has killed. $H^1(\mathcal{F}_{\text{full}}) = 0$ confirms that full internal state resolves everything; the gap $\dim H^1(\mathcal{F}_{\text{obs}}) - \dim H^1(\mathcal{F}_{\text{full}}) = 2$ is exactly the structural blind spot. In the simple case where each projected-away dimension appears in exactly one edge constraint, the fee equals the count of projected fields. The framework's value is that it generalizes: when projected fields participate in multiple constraints, or when restriction maps have nontrivial kernel, the rank deficiency can differ from the naive count.

3 The Coherence Fee

Definition 3.1 (Coherence fee)**.** The *coherence fee* of a tool composition is $\dim H^1(\mathcal{F}_{\text{obs}}) - \dim H^1(\mathcal{F}_{\text{full}})$: the number of independent semantic dimensions that bilateral verification cannot reach because π projects them away, net of any purely topological obstruction. Under the construction of Section 2, $H^1(\mathcal{F}_{\text{full}}) = 0$ (every dimension in $S(v)$ is reachable), so the fee reduces to $\dim H^1(\mathcal{F}_{\text{obs}})$. The gap formulation survives generalization to settings where $S(v)$ is incomplete.

The coherence fee is a structural property of the schema graph. It is zero if and only if every consistency-relevant dimension at every edge is covered by at least one adjacent tool's observable schema. It is the *irreducible cost* of making claims compose across tool boundaries without a trusted intermediary.

A bridge annotation eliminates a chokepoint: it makes an implicit dimension bilaterally checkable without requiring anyone to trust the orchestrator's interpretation. The coherence fee is what you pay when you choose not to bridge. The *rent* is what a semantic chokepoint extracts when it controls the only interpretation of an unbridged dimension.

In a marketplace of tools, a tool author who refuses to add bridge fields forces every downstream orchestrator to absorb the coherence fee—accepting silent failure modes that no bilateral check can detect. Competing tools that do support bridges offer a lower fee and are preferred

by risk-sensitive orchestrators. This creates pressure toward $\dim H^1 = 0$ at equilibrium: the same competitive dynamic that drives API providers to publish comprehensive schemas today will drive them to publish bridge fields tomorrow.

The pressure can be made concrete through a three-layer market mechanism:

1. *Fee-footprint publication.* Each tool publishes the number of blind-spot dimensions it contributes to any composition (its "fee footprint"). Orchestrators publish a "max coherence fee tolerated" for their pipelines. Tools compete on fee footprint alongside latency, cost, and accuracy.
2. *Procurement thresholds.* A rule ("only compose tools with coherence fee $\leq k$") turns the fee into a qualification threshold. A registry that scores tools by fee footprint gives orchestrators a quantitative basis for tool selection.
3. *Priced risk.* Insurance priced to the coherence fee makes the cost of non-bridging visible in the budget, not just in the risk register. An escrow pool funded by the coherence fee itself—tool authors deposit proportional to the fee they impose; slashed when a valid fraud proof is submitted—creates a direct incentive to bridge.

3.1 The Failing Scenario

We construct concrete tool outputs demonstrating the blind spot. All three bilateral checks pass:

Tool	Observable output	Internal assumption
`data_provider`	prices, period_unit= "calendar"	day_convention= "calendar"
`financial_analysis`	risk_adj_return= 1.67, vol= 0.19	day_convention= "trading", risk_metric= "sortino"
`portfolio_verif.`	pass= true, score= 0.91	risk_metric= "sharpe"

The bilateral checker at each edge sees only observable fields and reports success. But the internal states disagree on two dimensions: `day_convention` ("calendar" $\neq$ "trading") and `risk_metric` ("sortino" $\neq$ "sharpe").

These are the two generators of $H^1(\mathcal{F}_{\text{obs}})$, made concrete.

Negative control: existing validation passes. We submitted the three tool outputs from the failing scenario to standard MCP JSON Schema validation. The validator checks: every required field present, every type correct, every value within its declared range. All three tools pass. The validator *cannot* flag the `day_convention` drift (calendar vs. trading) or the `risk_metric` mismatch (sortino vs. sharpe), because neither field appears in any schema. This is not a limitation of the validator's implementation—it is a structural impossibility. The information needed to detect the inconsistency has been projected away by π. No improvement to bilateral schema validation can close this gap; only bridge annotations can.

3.2 Eliminating the Fee

Definition 3.2 (Bridge annotation)**.** A *bridge annotation* for blind-spot dimension d at edge $e = (u, v)$ extends the observable schemas: add field f_d to both $F(u)$ and $F(v)$, so that π_u and π_v no longer kill f_d. The previously-zero row of δ^0_{obs} becomes:

$$-1 \cdot (u, f_d) \; + \; +1 \cdot (v, f_d)$$

Concretely, the `financial_analysis` tool's output schema before and after bridging:

Listing 1: Before: risk_metric is projected away

```
{
```

```
  "output": {
    "risk_adjusted_return": {"type": "number"},
    "volatility":           {"type": "number"},
    "max_drawdown":         {"type": "number"},
    "annualized_return":    {"type": "number"}
  }
}
```

Listing 2: After: bridge annotation makes risk_metric observable

```
{
  "output": {
    "risk_adjusted_return": {"type": "number"},
    "volatility":           {"type": "number"},
    "max_drawdown":         {"type": "number"},
    "annualized_return":    {"type": "number"},
    "risk_metric": {"type": "string",
                    "enum": ["sharpe","sortino","alpha"]}
  }
}
```

Theorem 3.3 (Bridge elimination)**.** *If* $\dim H^1(\mathcal{F}_{\text{obs}}) = k > 0$*, applying a bridge annotation for each of the* k *generators produces* $\mathcal{F}'_{\text{obs}}$ *with* $\dim H^1(\mathcal{F}'_{\text{obs}}) = 0$*.*

Proof. Each generator corresponds to a zero row in δ^0_{obs}. The bridge adds two columns (one per tool) and sets the row's entries to -1 and $+1$ in those columns. The row becomes linearly independent from all others (it is the only row with nonzero entries in those columns). Repeating for all k generators raises the rank by k, so $\dim H^1(\mathcal{F}'_{\text{obs}}) = (k + r) - (r + k) = 0$ where r was the original rank. □

For the financial pipeline, two bridges suffice:

Bridge field	Added to	Blind spot eliminated
`day_convention`	`dp`, `fa`	row 2 (`day_conv_match`)
`risk_metric`	`fa`, `pv`	row 5 (`metric_type_match`)

After bridging: $\delta^0_{\text{obs}'}$ is 8×16, rank $= 8$, $\dim H^1 = 0$. The same failing-scenario data now triggers two bilateral failures: `day_convention` = "calendar" $\neq$ "trading" at `dp`→`fa`, and `risk_metric` = "sortino" $\neq$ "sharpe" at `fa`→`pv`.

The blind spots become type errors.

4 The Protocol

4.1 Phases

The protocol enforces a single invariant:

> *No edge may rely on semantics that neither endpoint emits nor commits to.*

Every phase below exists to make this invariant computable, achievable, and enforceable.

Registration. Each tool registers its specification $(v, S(v), F(v))$ with a bridge registry or directly with the orchestrator. The composition graph is built, the coboundary is computed, and the coherence fee is reported. If $\dim H^1(\mathcal{F}_{\text{obs}}) > 0$, the generators are identified and bridge annotations are recommended. The registry is one implementation of this computation; it is not the protocol's essence. The essence is: **compute fee → bridge → enforce**.

Composition. An orchestrator queries the registry before composing a pipeline. If the fee is zero, bilateral checks suffice. If not, the orchestrator either applies the recommended bridges (reducing the fee to zero) or accepts the residual risk.

Enforcement. Two modes, corresponding to different trust assumptions.

Trusted mode: the orchestrator has access to all outputs and checks all bridge fields at every edge. Cost: $O(|E| \cdot d_{\max})$.

Trustless mode: tools are operated by independent parties. No single party is trusted to check correctly. This mode requires one additional primitive: the *semantic manifest commitment*.

4.2 Semantic Manifests

Definition 4.1 (Semantic manifest)**.** At composition time, each tool v publishes:

1. Its observable output $F(v)$ (in the clear).
2. A commitment $r_v = H(\texttt{manifest}_v)$ where $\texttt{manifest}_v$ contains values for *all* dimensions in $S(v)$, including those not in $F(v)$.

In normal operation, only $F(v)$ and r_v are visible. The internal dimensions remain private.

The commitment is a hash of the concatenated per-dimension hashes: $r_v = H(h_1 \| h_2 \| \cdots \| h_{s_v})$ where $h_i = H(\texttt{key}_i : \texttt{value}_i)$. A *selective opening* for dimension d reveals $(\texttt{key}_d, \texttt{value}_d, h_d)$ and lets any verifier check h_d against r_v.

4.3 Composition Fraud Proofs

Definition 4.2 (Composition fraud proof)**.** A *composition fraud proof* (a receipt for compositional inconsistency) is a tuple $\pi = (\{h_i\}, \{r_i\}, C, B)$:

- $\{h_i\}$: hashes of the tool schemas (commits to the specification).
- $\{r_i\}$: manifest root hashes (commits to the internal state at composition time).
- C: the cycle $v_1 \to v_2 \to \cdots \to v_1$.
- B: the *blind-spot certificate*: for each failing dimension, selective openings from both adjacent tools showing inconsistent values for a dimension that π projected away.

Verification ($O(n)$).

1. **Confirm all observable bilateral checks pass.** This is the step that makes the fraud proof non-trivial. If bilateral checks *also* fail, that is a bilateral error—a bug, not a blind spot. The receipt is interesting precisely when the composition looks correct to every local verifier and is globally inconsistent.
2. Verify schema hashes against the registry.
3. Verify manifest roots match the commitments published at composition time.
4. For each opening, verify the dimension hash matches $H(\texttt{key} : \texttt{value})$.
5. Confirm each blind spot exhibits a nonzero residual (the two tools' values disagree on a projected-away dimension).

If all five steps succeed, the receipt is valid: bilateral verification certified a composition that is globally inconsistent.

Theorem 4.3 (Soundness)**.** *If all bridges are applied ($\dim H^1(\mathcal{F}_{\text{obs}}) = 0$) and all bilateral checks pass, no valid fraud proof exists.*

Proof. When $\dim H^1(\mathcal{F}_{\text{obs}}) = 0$, every consistency-relevant dimension at every edge is observable. A bilateral check that passes on all observable dimensions therefore passes on *all* dimensions. Step 4 of verification requires a nonzero residual on a projected-away dimension, but with all bridges applied, no dimensions are projected away at any edge. Step 4 always fails; no valid receipt can be constructed. □

In plain terms: a fully bridged pipeline cannot be receipted for compositional failure. The receipt mechanism proves nothing because there is nothing to prove.

Remark. A fraud proof is not an accusation against a specific tool. On a cycle, the inconsistency is a property of the composition—it does not localize to any single edge. The receipt names the cycle, the openings, and the residual. It is a *global witness* for a non-localizable failure.

4.4 Why Cycles Require Receipts

On a DAG, a hidden inconsistency is attributable: follow the directed edges backward from the point of failure and you find the tool whose implicit choice caused it. The fix is local: bridge that tool's output or document its assumptions.

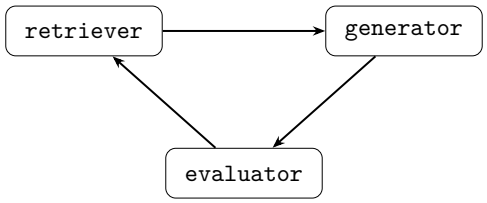
On a cycle, attribution fails. The inconsistency circulates: `dp` assumes calendar days; `fa` switches to trading days; `pv` interprets the result as if it were Sharpe when `fa` computed Sortino; `pv`'s feedback reaches `dp`, closing the loop. No single edge is "wrong." The composition is wrong.

This is why a receipt must exhibit the *entire cycle*: the schema commitments, the manifest commitments, and the openings around the loop. Cycles need witnesses; DAGs need blame.

Security model. The trustless mode assumes an honest-observer (1-of-N) model: correctness holds unless every observer is compromised. Manifest roots must be posted to a data-available channel. A dispute game specifies consequences when a valid fraud proof is submitted (reputational, financial, or operational). The full security model—roles, data availability modes, dispute mechanics, and collusion boundary—is in Appendix B.

5 A Second Example: RAG with Feedback

To confirm the model generalizes, we apply it to a retrieval-augmented generation (RAG) pipeline with a feedback loop:



Tool	π kills	Semantic role
`retriever`	chunk_size, relevance_threshold	How documents were chunked; what threshold was used
`generator`	chunk_size, citation_mode	Whether chunks were reassembled; strict vs. loose citation
`evaluator`	citation_mode, relevance_threshold	What citation fidelity means; what relevance means

Three bilateral interfaces carry 9 semantic dimensions (3 per edge). Three of those dimensions are blind spots: `chunk_size_match` (retriever→generator), `citation_mode_match` (generator→evaluator), `threshold_match` (evaluator→retriever).

$$\delta^0_{\text{obs}} : 9 \times 8, \quad \text{rank} = 6, \quad \dim H^1(\mathcal{F}_{\text{obs}}) = 3.$$
$$\delta^0_{\text{full}} : 9 \times 14, \quad \text{rank} = 9, \quad \dim H^1(\mathcal{F}_{\text{full}}) = 0.$$

The coherence fee is 3.

Composition	Tools	Edges	β_1	Fee	Blind-spot dimensions
Auth-Data-Audit	3	3	1	0	—
Financial Analysis	3	3	1	2	`day_convention`, `risk_metric`
RAG with Feedback	3	3	1	3	`chunk_size`, `citation_mode`, `relevance_threshold`
Code Review	4	4	1	3	`diff_format`, `severity_threshold`, `style_convention`
Web Research	4	4	1	3	`search_language`, `citation_style`
Multi-Source ETL	4	4	1	4	`timezone`, `null_convention`, `decimal_precision`
From deployed MCP server source code (Section 6.1)					
Filesystem ↔ Git	2	2	1	3	`encoding`, `line_ending`, `path_separator`
Fetch ↔ Memory	2	2	1	2	`encoding`, `content_format`
Fetch → Filesystem → Git	3	3	1	4	`encoding`×2, `line_ending`, `path_separator`

Table 1: Coherence fee diagnostic across nine compositions. The first six use author-constructed schemas; the last three are extracted from the source code of deployed MCP servers (the `modelcontextprotocol/servers` repository). After bridging, every composition drops to fee = 0.

A concrete RAG failure. The retriever chunks documents at 512 tokens with `relevance_threshold` = 0.7. The generator receives the chunks (observable), reassembles them into a response, and internally uses `chunk_size` = 256 (it was fine-tuned on smaller chunks) and `citation_mode` = "strict" (verbatim spans only). The evaluator checks citation fidelity using `citation_mode` = "loose" (paraphrase acceptable) and `relevance_threshold` = 0.5.

Every bilateral check passes: the chunks arrive, the response cites them, the quality score is high. But the generator is silently re-chunking at 256 tokens, breaking the retriever's span boundaries. The evaluator accepts paraphrased citations that the generator flagged as strict verbatim matches. And the evaluator's feedback loop sends a relaxed relevance threshold back to a retriever that expects 0.7—gradually degrading retrieval quality over iterations.

These are three generators of $H^1(\mathcal{F}_{\text{obs}})$, made concrete. Three bridge annotations (`chunk_size`, `citation_mode`, `relevance_threshold`) reduce the fee to zero and convert each silent drift into a type error at the offending edge.

Remark. The recurring implicit dimensions across both examples—conventions, thresholds, metric definitions, strategy choices—are exactly the assumptions that tool authors treat as "obvious" and omit from schemas. The coherence fee makes the cost of that omission computable.

6 Diagnostic Results

To demonstrate that the coherence fee is both computable and practically informative across domains, we implemented `seam-lint`: a diagnostic tool that takes a YAML composition specification (tool schemas, bilateral interfaces) and reports the coherence fee, identifies blind-spot dimensions, and recommends bridge annotations. We ran it on nine compositions spanning finance, retrieval-augmented generation, DevOps, data engineering, security, web research, and three compositions derived directly from the source code of deployed MCP servers (Section 6.1).

Five findings stand out.

Real MCP tool types produce nonzero fees. The web research pipeline is composed of four tools modeled on commonly-deployed MCP servers: Brave Search (web search API), a content extraction server (Firecrawl, Jina Reader), an LLM-based summarizer, and a fact-checking tool, with a feedback loop from fact-checker to search for iterative verification. The coherence fee is 3, corresponding to two implicit dimensions: `search_language` (assumed by the search engine, propagated through scraping and summarization, but declared in no schema—three

blind-spot edges from a single hidden convention) and `citation_style` (the summarizer produces inline citations; the fact-checker expects numbered references). Two bridge annotations—adding `search_language` and `citation_style` to the relevant schemas—reduce the fee to 0. No existing MCP schema validation would flag either dimension. The `search_language` case demonstrates a structural point: a single omitted dimension can contribute multiple blind-spot edges—the fee measures *topological footprint*, not merely the count of hidden assumptions.

Deployed MCP servers confirm the prediction. Section 6.1 analyzes three compositions built entirely from the source code of servers in the official MCP repository. The Filesystem $\leftrightarrow$ Git composition—the single most common MCP multi-tool workflow—has fee 3. The three blind-spot dimensions (`encoding`, `line_ending_convention`, `path_separator`) correspond to failure modes that practitioners already encounter: CRLF/LF inconsistencies on Windows, encoding mismatches with non-ASCII filenames, and OS-dependent path separators in diff output. None appears in any MCP tool schema. A 3-tool composition (Fetch $\rightarrow$ Filesystem $\rightarrow$ Git) raises the fee to 4: the `encoding` convention, undeclared at all three tools, creates blind spots at every boundary it crosses.

The fee varies by domain. Finance and security compositions, where schemas tend to be well-specified, have lower fees (0–2). RAG and data-engineering compositions, where conventions (chunk boundaries, null handling, decimal precision) are often left implicit, have higher fees (3–4). The tool quantifies a difference that practitioners sense but cannot currently measure.

Zero fee is achievable. The auth-data-audit pipeline has $S(v) = F(v)$ for all tools—every consistency-relevant dimension is already in the output schema. Its $H^1(\mathcal{F}_{\text{obs}}) = 1$ is purely topological (one cycle creates one redundant constraint among the `user_id` matching dimensions). The coherence fee, correctly defined as $\dim H^1(\mathcal{F}_{\text{obs}}) - \dim H^1(\mathcal{F}_{\text{full}}) = 0$, separates topological from observability contributions.

Bridge recommendations are minimal. The ETL pipeline has four blind-spot dimensions but only three unique bridge fields (`timezone`, `null_convention`, `decimal_precision`), because `null_convention` appears on two edges. The tool deduplicates and reports three bridge annotations that jointly reduce the fee from 4 to 0.

The fee scales with composition size. An undeclared convention contributes a blind-spot dimension at *every* composition boundary it crosses. The coherence fee is therefore $O(|E| \cdot c)$, where c is the number of undeclared conventions and $|E|$ is the number of edges each convention spans—not merely $O(c)$. The 3-tool deployed composition (Fetch $\rightarrow$ Filesystem $\rightarrow$ Git) provides direct evidence: `encoding`, undeclared at all three tools, crosses two boundaries and contributes 2 to the fee, raising it from 3 (2-tool) to 4 (3-tool). In a 10-tool pipeline with 3 undeclared conventions, the fee could reach 15, not 3. Bridge annotations have *increasing* marginal returns: declaring a convention once eliminates its contribution at every boundary simultaneously.

A testable prediction. We predict that *in practice*, every MCP pipeline with a feedback loop ($\beta_1 \geq 1$) contains at least one convention-type dimension (encoding, line-ending format, time basis, null representation) that shapes outputs at multiple tools but appears in no tool's output schema. This is an empirical claim about how tool authors actually design schemas—they treat conventions as "obvious" and omit them—not a mathematical tautology. The claim is falsifiable: find a cyclic MCP composition where the tool authors *did* declare their encoding, line-ending convention, timezone, and other convention-type dimensions in their output schemas. That would show the "omission is universal" premise is wrong. A zero-fee cyclic

pipeline where conventions were *never relevant* (like the auth-data-audit pipeline) does not falsify the prediction—it confirms that tools without convention-type requirements can achieve fee 0. In the nine compositions analyzed here—including three from deployed server source code—every cyclic pipeline with at least one convention-type hidden dimension has fee ≥ 1. We further predict that in production pipelines with $n > 5$ tools, the fee will grow superlinearly in the number of tool boundaries each undeclared convention crosses—making the cost of *not* bridging increasingly visible at scale.

The diagnostic tool, all composition specs (YAML), and output are at `papers/seam/seam-lint/`.

6.1 Diagnostic on Deployed MCP Servers

The six compositions above use schemas constructed by the authors to illustrate the framework. A stronger test is to apply the diagnostic to tool compositions that someone else built for a different purpose and ask whether the framework discovers something about them.

Methodology. We selected MCP servers from the official `modelcontextprotocol/servers` repository (commit `a83b145`). For each server, we identified $F(v)$ from the tool's declared output format in the source code—the fields that actually appear in `CallToolResult`—and $S(v)$ by reading the implementation to identify configuration parameters, environment variables, and hardcoded conventions that shape output but do not appear in the schema. Bilateral interfaces were inferred from shared semantic requirements: when two tools in a composition must agree on a convention for their composed output to be consistent, we record that convention as a semantic dimension at the connecting edge. This identification step is necessarily manual—it requires domain understanding of what "consistency" means for each dimension—and we report it transparently: the YAML composition specs are published alongside the paper for independent verification.

Composition 1: Filesystem ↔ Git. This is the canonical AI-assisted coding workflow: the Filesystem server reads and writes local files; the Git server tracks changes, produces diffs, and commits. A feedback loop (Git status/diff → Filesystem edit) makes this a cyclic composition ($\beta_1 = 1$). We identified $|S(v)| = 6$ internal dimensions for Filesystem and $|S(v)| = 9$ for Git by reading the source. Three dimensions are in $S(v) \setminus F(v)$ at *both* endpoints:

- **`encoding`.** Filesystem hardcodes UTF-8 (`lib.ts:157`). Git uses locale-dependent encoding for subprocess output and hardcodes UTF-8 only for binary diffs (`server.py:216`). Neither tool declares its encoding assumption in any output field.
- **`line_ending_convention`.** Filesystem returns raw bytes from disk for reads (preserving CRLF on Windows) but normalizes to LF for `edit_file` (`lib.ts:56`). Git assumes LF in string splits (`server.py:154`). The MCP server has no `core.autocrlf` awareness. Neither tool declares the convention.
- **`path_separator`.** Filesystem uses OS-native separators via `path.join` (`path-utils.ts:38`). Git always outputs POSIX `/` in diff headers and status. A developer on Windows sees `src\file.txt` from Filesystem and `src/file.txt` from Git—for the same file.

The coherence fee is 3. All three blind spots are convention-type dimensions—exactly the kind our testable prediction targets. The `path_separator` dimension appears on both edges of the cycle, but the coboundary correctly identifies this redundancy: $H^1(\mathcal{F}_{\text{full}}) = 1$, so the net fee is $H^1(\mathcal{F}_{\text{obs}}) - H^1(\mathcal{F}_{\text{full}}) = 4 - 1 = 3$. Three bridge annotations (adding `encoding`, `line_ending_convention`, and `path_separator` to both servers' output schemas) reduce the fee to 0.

The failing scenario. A developer on Windows edits a source file. The Filesystem server's `read_file` returns the contents with CRLF line endings (raw bytes from disk). An MCP client

stores this content and calls `git_diff_unstaged` to obtain the diff—whose context lines use LF only (Git convention). The client attempts to match diff hunks against the file content: line n in the diff reads `function foo() {\n` while the file has `function foo() {\r\n`. A patch application or line-number correlation fails silently. Both bilateral checks pass: `read_file` returns valid UTF-8 text, `git_diff` returns a syntactically correct unified diff. The inconsistency is *between* the two outputs, in a dimension—`line_ending_convention`—that neither output schema declares.

Composition 2: Fetch ↔ Memory. A knowledge-acquisition pipeline: the Fetch server retrieves web content and converts HTML to markdown; the Memory server stores entities and relations as a JSONL knowledge graph. A feedback loop (Memory search identifies gaps → Fetch retrieves more) makes this cyclic. Two blind spots:

- **`encoding`.** Fetch relies on `httpx`'s charset detection (Content-Type header → UTF-8 fallback → `charset_normalizer`). Memory reads and writes JSONL as UTF-8 (`fs.readFile(..., "utf-8")`). If Fetch encounters a non-UTF-8 page, the decoded Python string may contain lossy substitutions that Memory stores without awareness.
- **`content_format`.** Fetch converts HTML to markdown via `readabilipy` + `markdownify`—producing ATX headings, inline links, and emphasis markers. Memory stores observations as opaque strings. A Memory search for "important concept" will fail to match `**important concept**` from the markdown conversion. Neither tool declares the format convention.

The coherence fee is 2. Both bridges are actionable: Fetch could declare `content_format: "markdown"|"raw"` in its output; Memory could declare `observation_format: "plaintext"` and normalize on ingestion.

Composition 3: Fetch → Filesystem → Git. A web-content-to-version-control pipeline: Fetch retrieves documentation or data, Filesystem writes it locally, Git commits the changes. A feedback loop (Git log reveals committed URLs → Fetch checks for updates) makes this cyclic. The composition chains all three official servers, producing a 3-tool, 3-edge graph with $\beta_1 = 1$.

The coherence fee is 4—higher than either 2-tool sub-composition. The increase comes from `encoding`, which now appears as a blind spot on *two* edges: Fetch→Filesystem (httpx charset detection vs. hardcoded UTF-8) and Filesystem→Git (UTF-8 vs. locale-dependent subprocess). A single hidden convention contributes two independent blind-spot dimensions because it crosses two composition boundaries. The remaining two blind spots (`line_ending_convention`, `path_separator`) are the same as in the 2-tool Filesystem↔Git composition. Three bridge annotations—`encoding` (added to all three tools), `line_ending_convention`, and `path_separator`—reduce the fee to 0.

Unlike the 2-tool Filesystem↔Git case, $H^1(\mathcal{F}_{\text{full}}) = 0$ here: the `path_separator` dimension appears on only one edge (Filesystem→Git, not Git→Fetch), so there is no topological redundancy. The fee equals the raw $H^1(\mathcal{F}_{\text{obs}})$, and the gap formulation and the standalone formulation agree—confirming that the gap matters only when the same dimension appears on multiple edges of the same cycle.

Summary. All three deployed compositions have nonzero coherence fees. The blind spots correspond to real engineering problems: CRLF/LF inconsistencies, encoding mismatches, path separator mismatches, and markdown-vs-plaintext confusion. The 3-tool composition demonstrates that fees accumulate across composition boundaries: a single undeclared convention (`encoding`) that is harmless in a 2-tool DAG becomes a blind spot at every boundary it crosses. In every case, bilateral MCP schema validation (JSON Schema on inputs, type checking on outputs) would pass—the inconsistency is structurally invisible to pairwise checks, exactly as the theory predicts.

7 Scope and Limitations

Known unknowns, not unknown unknowns. The model requires that someone enumerate the semantic dimensions at each edge, including implicit ones. If a dimension is not recognized and listed, it will not appear in the coboundary and will not be detected. The protocol diagnoses known unknowns. It does not discover unknown unknowns—but once a dimension is discovered (by testing, auditing, or failure), it becomes permanently checkable.

Three mechanisms make discovery systematic rather than artisanal:

1. *Standard manifest vocabulary.* A canonical set of dimension categories—time conventions, unit systems, rounding modes, objective functions, model identifiers, retrieval settings—with a requirement that tools declare `unknown/NA` explicitly for each. The vocabulary grows as domains are onboarded.
2. *Schema linting.* A static pass over tool docstrings, tests, and output schemas that proposes candidate hidden dimensions ("this tool annualizes—what day-count convention?"). The lint does not need to be perfect; it converts unknown unknowns into known unknowns for human review.
3. *Domain invariant classes.* For well-understood domains, a curated set of always-bridged dimensions: `day_convention` and `risk_metric` in finance, `chunk_size` and `citation_mode` in RAG, `timezone` and `currency` in cross-border transactions. These encode hard-won domain knowledge as reusable infrastructure.

The protocol does not solve discovery perfectly—but it prevents the critique "this is only as good as manual enumeration" by providing concrete tooling that monotonically expands the known-unknown set.

Static analysis. The coherence fee is computed from schemas, not runtime values. It identifies structural blind spots. Whether bridge field values are *truthful* at runtime is the enforcement layer's job (bilateral checks in trusted mode; commitments and fraud proofs in trustless mode).

Cycles. The distinction between attributable (DAG) and non-localizable (cycle) failures means the fraud-proof machinery is most valuable for cyclic compositions. As agent pipelines mature—feedback loops, self-improvement cycles, multi-agent negotiation, monitoring agents feeding parameters back to data agents—cycles will become the norm, not the exception.

Bridge adoption. The protocol's value depends on tool authors adding bridge fields. This is a coordination problem, not a technical one. The protocol provides the diagnostic (the fee is nonzero) and the prescription (which fields to add to which tools). It cannot force adoption, but it makes the cost of non-adoption explicit and computable.

8 Related Work

Tool-calling protocols. MCP [2] defines bilateral tool schemas with JSON Schema types. A2A [4] adds agent-to-agent negotiation. OpenAPI [6] provides REST-level schema validation. All three enforce bilateral compatibility; none address compositional consistency across cycles.

Schema mapping composition. Fagin et al. [3] prove that composing schema mappings requires second-order dependencies: first-order constraints specified pairwise between two schemas do not compose transitively to a third. This is the database-theoretic foundation for why pairwise reconciliation does not compose—and the closest prior result to our bilateral completeness claim. Our contribution is the specific quantification: the coherence fee measures *how much* composition fails, not just *that* it fails, and the protocol layer makes the gap enforceable.

Sheaf-theoretic data integration. The use of cellular sheaves for data consistency has roots in applied algebraic topology. Our two-layer model (observable vs. full sheaf) is a direct application: the rank deficiency of the observable coboundary measures the information lost by projection. Appendix A provides the full sheaf-theoretic formulation.

Optimistic rollups and fraud proofs. The composition fraud proof (Definition 4.2) follows the optimistic rollup pattern [1]: assume correctness, allow any party to challenge with a succinct proof, revert if valid. The key difference: our proofs target semantic inconsistency across tool boundaries, not state-transition invalidity. The semantic manifest commitment (Definition 4.1) provides the cryptographic primitive that makes the analogy precise.

A Sheaf-Theoretic Foundation

The linear-algebraic model of Section 2 is an instance of a *cellular sheaf* on the directed composition graph $G = (V, E)$.

Definition A.1 (Schema sheaf)**.** The *observable schema sheaf* $\mathcal{F}_{\text{obs}}$ assigns:

- To each vertex $v \in V$, a stalk $\mathcal{F}_{\text{obs}}(v) = F(v) = \mathbb{R}^{f_v}$.
- To each edge $e \in E$, a stalk $\mathcal{F}_{\text{obs}}(e) = C(e) = \mathbb{R}^{d_e}$.
- Restriction maps $\rho^{\text{obs}}_{v \to e} : F(v) \to C(e)$ defined by coordinate projection (or zero if π_v kills the relevant dimension).

The *full schema sheaf* $\mathcal{F}_{\text{full}}$ is defined identically with $S(v)$ replacing $F(v)$, and no restriction map is ever zero.

The cochain complex is:

$$0 \to C^0 \xrightarrow{\delta^0} C^1 \to 0$$

where $C^0 = \bigoplus_v \mathcal{F}_{\text{obs}}(v)$, $C^1 = \bigoplus_e \mathcal{F}_{\text{obs}}(e)$, and δ^0 is as in Definition 2.4. The first cohomology is:

$$H^1(\mathcal{F}_{\text{obs}}) = \operatorname{coker}(\delta^0) = C^1 / \operatorname{im}(\delta^0).$$

Theorem A.2 (Coherence fee as cohomological gap)**.**

$$\textit{coherence fee} = \dim H^1(\mathcal{F}_{\text{obs}}) - \dim H^1(\mathcal{F}_{\text{full}}).$$

When every semantic dimension at every edge has a corresponding internal-state dimension at both adjacent vertices and each such dimension indexes a unique column *of the full coboundary, all rows of* δ^0_{full} *are linearly independent and* $H^1(\mathcal{F}_{\text{full}}) = 0$*, so the fee reduces to* $\dim H^1(\mathcal{F}_{\text{obs}})$*.*

Remark. The condition can fail when the *same* internal-state dimension participates in bilateral requirements on multiple edges of a cycle. In the Filesystem $\leftrightarrow$ Git composition (Section 6.1), `path_separator` appears on both edges; the two corresponding rows of δ^0_{full} are linearly dependent, giving $H^1(\mathcal{F}_{\text{full}}) = 1$. This is a *topological* contribution—it reflects the redundancy inherent in cycles, not missing internal state. The gap formulation correctly subtracts it: fee $= 4 - 1 = 3$.

Proof. When the uniqueness condition holds, in δ^0_{full} every row has at least one nonzero entry in a column unique to that row (the `from_field` or `to_field` in $S(v)$). The rows are therefore linearly independent, giving $\operatorname{rank}(\delta^0_{\text{full}}) = \dim C^1$ and $H^1(\mathcal{F}_{\text{full}}) = 0$. When it does not hold, $H^1(\mathcal{F}_{\text{full}}) \geq 1$, and the gap formulation accounts for the topological contribution.

Each bridge annotation extends $F(v)$ to include a dimension previously in $S(v) \setminus F(v)$, making the observable restriction map nonzero where it was previously zero. Applying bridges for all blind-spot generators recovers $\delta^0_{\text{obs}'} = \delta^0_{\text{full}}$ (restricted to the bridged observable dimensions), yielding $H^1(\mathcal{F}'_{\text{obs}}) = H^1(\mathcal{F}_{\text{full}})$—i.e., the fee drops to zero. □

Availability. The worked examples (Python source, twelve checkpoints) are at `papers/seam/example/seam_ex`
The `seam-lint` diagnostic tool, composition specs, and results (Table 1) are at `papers/seam/seam-lint/`.

B Participants & Security Model

Roles. Four roles participate in the protocol. A single entity may occupy multiple roles.

- *Tool operators*: publish $(v, S(v), F(v))$, produce outputs, commit to manifests.
- *Orchestrators*: compose tools into pipelines, run bilateral checks, apply bridges, optionally post bonds.
- *Verifiers* (challengers): any party that reads commitments and constructs fraud proofs.
- *Registry*: stores schema and manifest commitments; adjudicates fraud proofs.

Honest observer (1-of-N). The fraud-proof mechanism requires that at least one verifier along the cycle can read all committed manifest roots and observable outputs, and is willing to raise a fraud proof if an inconsistency exists. Correctness holds unless *every* observer is compromised or colluding—the same assumption as optimistic rollups.

Data availability. Manifest root hashes and observable outputs must be posted to a location all parties can read—an orchestrator log, a shared ledger, or a broadcast channel. If a tool can withhold its manifest root, verification cannot proceed. Two deployment modes apply:

- *Trusted DA*: the orchestrator stores and serves all commitments. Sufficient when the orchestrator is a first party.
- *Challenge-response DA*: manifest roots are posted publicly; a challenge triggers selective opening within a dispute window Δt. Failure to open within Δt is treated as an admission of inconsistency.

Dispute game. When a verifier submits a fraud proof:

1. The registry verifies the proof ($O(n)$, Section 4).
2. If valid: the accused composition is marked *faulted*. Consequences are application-specific but fall into three tiers: (a) *reputational*—the tools' fee-footprint scores increase, reducing their ranking in future compositions; (b) *financial*—an orchestrator's bond is slashed or an insurance claim is triggered against the unresolved coherence fee; (c) *operational*—the pipeline is halted pending bridge resolution.
3. If invalid (verification fails at any step): the proof is discarded; optionally the challenger's deposit is forfeit to discourage spam.

Collusion boundary. Bilateral collusion (two adjacent tools agreeing to report matching bridge values that differ from their actual computation) is detectable by an end-to-end audit that compares final outputs against claimed bridge values. Full-cycle collusion—where *all* tools agree to misrepresent—is outside the scope of compositional verification. It reduces to the trust assumption on individual tools, not on their composition.

References

[1] John Adler and Mikerah Quintyne-Collins. Fraud and data availability proofs: Maximising light client security and scaling blockchains with dishonest majorities. 2018. arXiv:1809.09044.

[2] Anthropic. Model context protocol specification. 2024. `https://modelcontextprotocol.io`.

[3] Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, and Wang-Chiew Tan. Composing schema mappings: Second-order dependencies to the rescue. *ACM Transactions on Database Systems*, 30(4):994–1055, 2005.

[4] Google. Agent-to-agent protocol (A2A). 2025. `https://google.github.io/A2A/`.

[5] John Komkov. The coherence fee: Edge-local blindness at the string-table seam and the topological price of cross-system composition. 2026. Companion paper.

[6] OpenAPI Initiative. OpenAPI specification v3.1, 2021. `https://spec.openapis.org/oas/v3.1.0`.

Bulla Formal Trilogy and Repair Geometry

Parts I–III give the formal hinge, the empirical witness, and the protocol consequence. Part IV develops the structural identity, field-independence, spectral geometry, and repair calculus of the witness rank on real-schema MCP corpora.

The Bulla Formal Trilogy is three papers that develop the obstruction invariant from three different vantage points. The Coherence Fee (Paper I) gives the hierarchical decomposition theorem: the witness rank decomposes into local sub-workflow fees plus a non-negative boundary fee, with an additive tower law across hierarchy levels. Column-Matroid Backbone (Paper II) establishes the structural identity: the fee equals the matroid corank of the observable column set in the linear matroid on the full coboundary columns, with a closed-form pairwise formula `fee(A,B) = U(A) + U(B) + |shared_dims(A,B)|`. The Witness Gram (Paper III) establishes the spectral geometry: the witness Gram `K(G) = Hᵀ(I − P_O)H` is always a Kron-reduced graph Laplacian, and effective resistance between hidden fields quantifies disclosure substitutability — the smaller the resistance, the more equivalent two fields are as repair choices.

These three papers are anchored in machine-checked Lean 4 proofs (Aristotle, standard axioms) and a 703-composition real-schema MCP registry corpus drawn from actual MCP servers, not synthetic data. The corpus tests every claim: leverage predictions match across all 240 nonzero-fee compositions, including 37 natural partial-CHP regime breaks; multi-column block rank ≤ 2 holds on every block tested.

Two further papers in this part complete the formal-geometric picture.

Signed-Incidence Structure is a short structural note supplying three facts used repeatedly across the trilogy: (i) each row of the coboundary matrix has at most one +1 and one −1; (ii) the witness rank is field-independent as a difference of two totally-unimodular ranks; (iii) pairwise endpoint coupling is bounded by rank 2 on every (edge, dimension) block. The TU theorem is checked in Lean 4 (Aristotle, standard axioms), and the row-structure property is verified across the real-schema corpus. A negative corollary records that any partition-style typed repair constraint is structurally wrong: enforcing it makes fee-zero repair impossible in 151 of 240 nonzero-fee compositions on the corpus.

Witness Geometry Beyond Scalar Fee develops the repair calculus that lives inside the fee. The coherence fee counts blind spots; this paper develops the geometry that determines all repair combinatorics. Under the DFD + CHP regime, the number of minimum disclosure bases is `β(G) = ∏|C_{d,j}|`,

a product over connected components of the witness Gram. Repair entropy `H_repair = log β` decomposes additively. Sharp bounds $\phi+1 \le \beta \le 2^{\wedge}\phi$ hold at fixed fee ϕ. A full realizability theorem shows every basis is the unique optimum under some cost vector — operational sparsity is a property of the cost family, not the matroid. The witness rank, repair entropy, and stability ratio are three projections of one underlying object: the witness Gram.

The five papers in Part IV — Coherence Fee (Paper I), Column-Matroid Backbone (Paper II), Witness Gram (Paper III), Signed-Incidence, and Witness Geometry Beyond Fee — together establish the structural-spectral-repair-calculus surface that the doctrine paper subsumes as the canonical invariant. What follows are the papers themselves.

The Coherence Fee

Measuring Hidden Convention Risk in Hierarchical Tool Compositions

John Komkov
Independent Researcher
https://github.com/jkomkov/bulla

April 2026

Abstract

When workflows compose tools across typed boundaries, each bilateral handoff may type-check while hiding implicit convention mismatches—timezone formats, rounding modes, jurisdictional frameworks—that no pairwise validation can detect. We introduce the *coherence fee*, a computable integer counting convention dimensions whose consistency is structurally unverifiable from observable schemas alone, and the *minimum disclosure set*, exactly φ field disclosures that eliminate all blind spots.

A *hierarchical decomposition theorem* shows the fee splits into local sub-workflow fees plus a non-negative *boundary fee* measuring conventions hidden at delegation boundaries. A *tower law* proves this cost is additive across hierarchy levels—formalizing "more delegation increases hidden risk" as a theorem. An online resolution protocol supports incremental tool substitution with monotonically decreasing fee.

An 8-tool financial settlement case study demonstrates 2.1× savings in disclosure prescriptions over per-edge bridge suggestions. All claims are computationally verified across 10 bundled and 10,000 adversarial random compositions; the reference implementation requires zero external dependencies beyond PyYAML. A companion signed-incidence note [16] supplies the middle-layer structural boundaries used elsewhere in the corpus: the fee is field-independent as a numerical invariant, and pairwise same-dimension endpoint disclosures can couple rather than separate.

1 Introduction

A financial settlement pipeline passes a trade through eight typed boundaries: price fetch, currency conversion, compliance check, risk assessment, ledger posting, audit logging, notification, and archival. At each boundary, the tool validates its inputs and outputs against its declared schema. Every bilateral handoff type-checks. The system passes integration testing. Six months later, a regulatory audit discovers that the risk model was applied under a different jurisdictional framework than the compliance check assumed, because the `jurisdiction` convention was hidden inside the compliance tool's internal state and never exposed to the risk assessment tool. The coherence fee of this composition is 7. The tool we describe would have flagged it at design time and prescribed exactly which 7 fields to expose.

This failure is not an engineering oversight—it is a structural property of hierarchical composition. Schema validation catches type errors at each handoff, but conventions (timezone formats, rounding modes, regulatory frameworks, accounting standards) propagate through chains of tools via fields that the receiving tool never declares in its observable interface. These *blind spots* are invisible to any bilateral check. Worse, we prove that every level of delegation in a hierarchical tool composition adds non-negative hidden convention cost (theorem 4.5). This is not a bug that more careful engineering can fix; it is a theorem about the algebraic structure of hierarchical composition. Intuitively, schema validation certifies that each output is a

well-formed *representation* of the declared type; it cannot certify the conventions under which the representation was produced, because those conventions were internal to the producing tool and absent from its observable interface.

Claim discipline. What is established here is the hierarchical fee and its decomposition through delegation boundaries. What failed productively in adjacent repair work was the stronger idea that dimension-aware disclosures would separate cleanly into one-per-dimension choices. What remains open is the right coupled-disclosure objective; the signed-incidence note [16] is the background citation for that boundary, not a result reproved in this paper.

Contributions.

1. **The coherence fee** (section 2): a computable integer $\varphi(G)$ measuring the number of convention dimensions structurally unverifiable from observable schemas.
2. **Minimum disclosure set** (section 6): a prescriptive list of exactly φ field disclosures sufficient to eliminate all blind spots, yielding 2.1× savings over per-edge bridge suggestions in our case study.
3. **Hierarchical decomposition with tower law** (section 4): the fee decomposes into local sub-workflow fees plus a non-negative boundary fee; the tower law shows boundary fees are additive across hierarchy levels, formalizing "delegation adds hidden cost."
4. **Online resolution** (section 6): an incremental protocol for substituting placeholder tools with monotonically decreasing fee.
5. **Negative results** (section 5): the boundary fee is monotone on the partition lattice but neither a valuation nor submodular—the latter disproved by an adversarial survey of 10,000 random compositions.
6. **Case study** (section 7): an 8-tool financial settlement pipeline with fee 7, boundary fee 2 under front/back-office partition, and conditional resolution demonstrating the full diagnostic loop.

Paper organization. Section 2 defines the composition model and coherence fee. Section 3 establishes basic properties through a counterexample. Section 4 proves the decomposition theorem, tower law, and extremal cases. Section 5 characterizes the boundary fee on the partition lattice (non-valuation, non-submodularity). Section 6 introduces the minimum disclosure set and online resolution protocol. Section 7 presents the financial settlement case study. Section 8 summarizes computational verification. Section 9 positions against related work, and section 10 concludes.

2 Model

Let $G = (V, E)$ be a composition graph where each node $v \in V$ is a tool with internal state I_v and observable schema $O_v \subseteq I_v$. Each edge $e \in E$ carries semantic dimensions, each specifying a `from_field` in the source tool and a `to_field` in the target tool.

The *coboundary operator* $\delta^0 \colon C^0 \to C^1$ is a matrix over $\mathbb{Q}$ with one column per (tool, field) pair and one row per (edge, dimension) pair. Row r has -1 in the `from_field` column and $+1$ in the `to_field` column, provided the field is in the relevant field set:

- δ^0_{obs}: columns are (v, f) with $f \in O_v$. Entries are nonzero only if the field is observable.
- δ^0_{full}: columns are (v, f) with $f \in I_v$. All dimension-linked fields contribute.

The *coherence fee* is

$$\varphi(G) = \operatorname{rank}(\delta^0_{\text{full}}) - \operatorname{rank}(\delta^0_{\text{obs}}).$$

This counts convention dimensions whose consistency is structurally unverifiable from observable schemas. The fee is non-negative by the inclusion $O_v \subseteq I_v$ [1, 2]. A *blind spot* is an edge-dimension where at least one endpoint field is hidden (in $I_v \setminus O_v$); a *bridge* is a per-blind-spot fix suggesting that the hidden field be added to the tool's observable schema.

3 The Coherence Fee

Example 3.1 (Non-additivity of fee under partition)**.** Consider three tools A, B, C in a chain $A \to B \to C$ sharing dimension `amount_unit`:

Tool	Internal state	Observable schema
A	`amount_unit`, `payload`	`payload`
B	`amount_unit`, `payload`	`amount_unit`, `payload`
C	`amount_unit`, `result`	`result`

Both edges carry dimension `amount` linking `amount_unit` on each side.

Sub-fees. $\{A, B\}$: A hides `amount_unit` but B exposes it. Observable coboundary row: $[0, -1, +1, 0] \neq 0$; same rank as full. $\varphi(\{A, B\}) = 0$.

$\{B, C\}$: B exposes, C hides. Same argument. $\varphi(\{B, C\}) = 0$.

Flat fee. δ^0_{obs} has columns for A.`payload`, B.`amount_unit`, B.`payload`, C.`result` and rows with entries only in B's column for both edges. $\operatorname{rank}_{\text{obs}} = 1$.

δ^0_{full} adds columns A.`amount_unit` and C.`amount_unit`; both rows become independent. $\operatorname{rank}_{\text{full}} = 2$.

$\varphi(\text{flat}) = 2 - 1 = 1 > 0 + 0$. The hidden convention propagates through B unseen by either sub-workflow.

Corollary 3.2 (Boundary vanishing under full disclosure)**.** *If every cross-partition edge dimension has both endpoint fields in their respective tools' observable schemas, then $\beta(\mathcal{P}) = 0$ and the fee is exactly additive: $\varphi(G) = \sum_i \varphi(G_i)$.*

Proof. When both `from_field` and `to_field` are observable for every cross-partition dimension, each cross-row x_{full} and x_{obs} have the same nonzero entries in the same columns. Hence $\rho_{\text{full}} = \rho_{\text{obs}}$ and $\beta = 0$. □

4 Hierarchical Decomposition

Definition 4.1 (Partition)**.** A *partition* of G is a collection $\mathcal{P} = \{G_1, \ldots, G_k\}$ of disjoint subsets of V with $\bigcup_i G_i = V$. An edge is *internal* to G_i if both endpoints lie in G_i; otherwise it is *cross-partition*.

The coboundary matrix decomposes into blocks:

$$\delta^0 = \left(\begin{array}{ccc} M_1 & & 0 \\ & \ddots & \\ 0 & & M_k \\ \hline X_1 & \cdots & X_k \end{array}\right)$$

where M_i contains the internal rows for group G_i and the bottom block $[X_1 \mid \cdots \mid X_k]$ contains the cross-partition rows.

Definition 4.2 (ρ)**.** For a choice of field set (obs or full), define $\rho = \operatorname{rank}(\delta^0) - \sum_i \operatorname{rank}(M_i)$, the rank contribution of cross-partition rows above the block-diagonal internal rows.

Theorem 4.3 (Hierarchical Fee Decomposition)**.** *For any partition $\mathcal{P}$ of a composition G,*

$$\varphi(G) = \sum_{i=1}^{k} \varphi(G_i) + \beta(\mathcal{P})$$

where the boundary fee $\beta(\mathcal{P}) = \rho_{\text{full}} - \rho_{\text{obs}} \geq 0$ *is the difference in cross-partition rank contribution between full and observable coboundary matrices.*

Proof. The block-diagonal structure gives $\operatorname{rank}(M_1 \oplus \cdots \oplus M_k) = \sum_i \operatorname{rank}(M_i)$ for both obs and full. Hence

$$\begin{aligned}\varphi(G) &= \operatorname{rank}(\delta^0_{\text{full}}) - \operatorname{rank}(\delta^0_{\text{obs}}) \\ &= \left[\textstyle\sum_i \operatorname{rank}(M_{i,\text{full}}) + \rho_{\text{full}}\right] - \left[\textstyle\sum_i \operatorname{rank}(M_{i,\text{obs}}) + \rho_{\text{obs}}\right] \\ &= \textstyle\sum_i \varphi(G_i) + \beta(\mathcal{P}).\end{aligned}$$

Non-negativity. The key observation is that coordinate projection cannot create new linear dependencies:

Lemma 4.4. *Let $P\colon \mathbb{Q}^{C_{\text{full}}} \to \mathbb{Q}^{C_{\text{obs}}}$ be the coordinate projection that drops hidden-field columns. If vectors $v_1, \ldots, v_m$ are linearly independent modulo a subspace W, then $P(v_1), \ldots, P(v_m)$ are linearly independent modulo $P(W)$.*

Proof. If $\sum a_i P(v_i) \in P(W)$ then $P(\sum a_i v_i) \in P(W)$, so $\sum a_i v_i \in W + \ker P$. Now $\ker P$ consists of vectors supported entirely on hidden-field columns (the columns that P projects away). Cross-partition rows, by construction, have nonzero entries only in columns indexed by fields of their two endpoint tools—both observable and hidden fields of those specific tools. Hidden-field columns of one partition group cannot coincide with endpoint columns of a cross-partition row spanning two *distinct* groups, because column indices are (v, f) pairs and cross-rows reference only the endpoints' own fields. Therefore $\ker P$ intersects the span of the cross-rows trivially: any vector in $W + \ker P$ that is also a linear combination of cross-rows must already lie in W. Hence $\sum a_i v_i \in W$, forcing all $a_i = 0$. □

Applying the lemma with $v_i = x_{i,\text{full}}$ (cross-rows) and $W = \operatorname{rowsp}(I_{\text{full}})$ (internal rows): since $P(r_{\text{full}}) = r_{\text{obs}}$ for each row and $P(\operatorname{rowsp}(I_{\text{full}})) \subseteq \operatorname{rowsp}(I_{\text{obs}})$, $\rho_{\text{full}} \geq \rho_{\text{obs}}$. □

Theorem 4.5 (Tower Law for Boundary Fees)**.** *Let $\mathcal{P} = \{G_1, \ldots, G_k\}$ be a partition of G, and let $\mathcal{Q}$ be a refinement of $\mathcal{P}$ obtained by sub-partitioning each group G_i via $\mathcal{P}_i$. Then*

$$\beta(\mathcal{Q}) = \beta(\mathcal{P}) + \sum_{i=1}^{k} \beta(\mathcal{P}_i).$$

Proof. Apply theorem 4.3 to partition $\mathcal{Q}$ of G: $\varphi(G) = \sum_j \varphi(Q_j) + \beta(\mathcal{Q})$. Apply it again to each sub-partition $\mathcal{P}_i$ of G_i: $\varphi(G_i) = \sum_{j \in \mathcal{P}_i} \varphi(Q_j) + \beta(\mathcal{P}_i)$. Substitute into the coarse decomposition $\varphi(G) = \sum_i \varphi(G_i) + \beta(\mathcal{P})$; the local fees $\varphi(Q_j)$ cancel, leaving $\beta(\mathcal{Q}) = \beta(\mathcal{P}) + \sum_i \beta(\mathcal{P}_i)$. □

Corollary 4.6 (Monotonicity under refinement)**.** *If $\mathcal{Q}$ refines $\mathcal{P}$, then $\beta(\mathcal{Q}) \geq \beta(\mathcal{P})$.*

Proof. Each $\beta(\mathcal{P}_i) \geq 0$ by theorem 4.3. □

Remark 4.7 (Lattice interpretation)*.* The boundary fee defines a monotone function on the refinement lattice of partitions of V: the trivial partition $\{V\}$ has $\beta = 0$; the discrete partition into singletons has $\beta = \varphi(G)$ (the total fee); every refinement moves monotonically upward. Operationally, each level of delegation in a hierarchical tool composition adds non-negative hidden cost.

Extremal cases. The vanishing corollary (corollary 3.2) gives the lower extreme: $\beta = 0$ when all boundary fields are observable.

Theorem 4.8 (Extremal boundary fee)**.** *For a star graph with hub H and spokes $S_1, \ldots, S_n$, where all conventions are hidden at every tool, the partition $\mathcal{P} = \{H\} \cup \{S_1, \ldots, S_n\}$ achieves $\beta(\mathcal{P}) = \varphi(G) = n$.*

Proof. Every edge $H \to S_i$ is cross-partition. Both groups are internally edge-free, so $\sum_i \varphi(G_i) = 0$ and $\beta = \varphi(G)$. Concretely, $\rho_{\text{obs}} = 0$ (empty observable schemas) and $\rho_{\text{full}} = n$ (each edge contributes an independent full-coboundary row via its unique spoke column). □

5 Lattice Properties

Remark 5.1 (Non-valuation). The boundary fee is *not* a valuation on the partition lattice. For the chain $A \to B \to C$ of example 3.1, take $\mathcal{P} = \{AB, C\}$ and $\mathcal{Q} = \{A, BC\}$. Then $\mathcal{P} \wedge \mathcal{Q} = \{A, B, C\}$ (the discrete partition) and $\mathcal{P} \vee \mathcal{Q} = \{ABC\}$ (the trivial partition), giving $\beta(\mathcal{P}) + \beta(\mathcal{Q}) = 2$ but $\beta(\mathcal{P} \wedge \mathcal{Q}) + \beta(\mathcal{P} \vee \mathcal{Q}) = 1 + 0 = 1$. The same hidden convention at B causes boundary fee in both $\mathcal{P}$ and $\mathcal{Q}$, but resolving it once (in the discrete partition) suffices.

Remark 5.2 (Non-submodularity). The boundary fee is *not* submodular on the partition lattice. An adversarial survey of 10,000 random compositions (635,095 partition pairs) found 4,061 violations of $\beta(\mathcal{P} \wedge \mathcal{Q}) + \beta(\mathcal{P} \vee \mathcal{Q}) \leq \beta(\mathcal{P}) + \beta(\mathcal{Q})$, with maximum violation magnitude 3.

Minimal counterexample: four tools T_0, T_1, T_2, T_3 with edges forming a cycle plus a shortcut. T_0 and T_1 have all fields hidden; T_2 and T_3 each expose one field. The partitions $\mathcal{P} = \{T_2\} \mid \{T_0, T_1, T_3\}$ and $\mathcal{Q} = \{T_3\} \mid \{T_0, T_1, T_2\}$ both have $\beta = 0$ (the isolated tool's single cross-edge has its endpoint visible). Their meet (singletons) has $\beta = 1$: refining into singletons exposes a hidden convention path $T_1 \to T_2 \to T_0$ that is internal in both $\mathcal{P}$ and $\mathcal{Q}$ but cross-partition in the meet.

This is consistent with the algebraic structure: while ρ_{full} and ρ_{obs} are individually submodular (as matroid rank restricted to row sets), their difference $\beta = \rho_{\text{full}} - \rho_{\text{obs}}$ need not be. The 10 bundled compositions happen to satisfy submodularity, but this is an empirical accident of their low-density, pipeline-like topology.

6 Prescriptive Diagnostics

6.1 Minimum Disclosure Set

The coherence fee counts hidden convention dimensions, but does not say *which* fields to disclose. The bridges mechanism of [1] suggests one disclosure per blind spot, but blind spots are per-edge while the fee lives in a vector space: some disclosures are redundant.

Definition 6.1 (Minimum disclosure set)**.** A *minimum disclosure set* for a composition G is a smallest set $S \subseteq \{(v, f) : v \in V,\ f \in I_v \setminus O_v\}$ such that adding every $(v, f) \in S$ to the observable schema of tool v reduces $\varphi(G)$ to zero.

Theorem 6.2 (Disclosure cardinality equals fee)**.** $|S| = \varphi(G)$ *for any minimum disclosure set S.*

Proof. Each disclosure adds at most one column to δ^0_{obs} outside $\text{col}(\delta^0_{\text{obs}})$. Reducing the fee by k requires k independent new columns, hence $|S| \geq \varphi(G)$. For the converse, let C_{hid} denote the columns of δ^0_{full} not present in δ^0_{obs} (the hidden-field columns). Since δ^0_{obs} is the submatrix of δ^0_{full} obtained by deleting exactly C_{hid}, the rank contributed by C_{hid} modulo the column space of δ^0_{obs} is $\text{rank}(\delta^0_{\text{full}}) - \text{rank}(\delta^0_{\text{obs}}) = \varphi(G)$. Therefore $\varphi(G)$ independent columns from C_{hid} span this quotient, and any basis of the quotient selects exactly $\varphi(G)$ disclosures. □

Remark 6.3 (Bridges vs. disclosures). The per-blind-spot bridges produce $|\text{bridges}| \geq 2\,\varphi(G)$ suggestions (one per hidden endpoint per edge). The minimum disclosure set collapses this: when tool A hides field f across multiple edges, a single disclosure of (A, f) resolves all of them. Empirically, $|\text{bridges}| \geq 2 \cdot |S|$ across all 10 bundled compositions.

6.2 Online Resolution

When a partial composition includes placeholder tools for as-yet-unspecified subsystems, `conditional_diagnos` computes a worst-case fee and exports boundary obligations. As real tools arrive, the composition can be updated incrementally.

Corollary 6.4 (Resolution monotonicity)**.** *Let $(G, \partial G)$ be a partial composition with placeholder tools having empty observable schemas. Replacing any placeholder with a real tool t (same edges, $O_t \supseteq \emptyset$) can only decrease or maintain the coherence fee: the placeholder is the worst case.*

Proof. The placeholder has $O_t = \emptyset$ and $I_t = \{f \mid f = d.to\_field$ for some edge dimension d incident to the placeholder$\}$. The real tool's internal state must contain every field referenced by its edge dimensions (otherwise the composition is invalid); since the placeholder's internal state is exactly those fields, $I_{\text{real}} \supseteq I_{\text{placeholder}}$ is a consequence of composition validity, not an assumption. Adding columns to δ^0_{full} can only weakly increase its rank. Meanwhile the real tool has $O_{\text{real}} \supseteq \emptyset$, so δ^0_{obs} gains columns that were already present in δ^0_{full}, weakly increasing $\text{rank}(\delta^0_{\text{obs}})$ by at least as much. Hence φ can only decrease. □

7 Case Study: Financial Settlement Pipeline

We construct an 8-tool composition modeling a hierarchical financial settlement workflow. The pipeline consists of: `price_fetch` → `currency_convert` → `compliance_check` → `risk_assess` → `ledger_write` → `audit_log` → `notification` → `archive`, with a feedback edge `audit_log` → `compliance_check` creating a cycle ($\beta_1 = 1$).

What could go wrong. Without coherence diagnostics, this pipeline silently propagates convention mismatches through the chain. The currency converter's internal `rounding_mode` determines how fractional amounts are truncated; the compliance checker's `threshold_currency` assumes a specific rounding convention for regulatory thresholds. Neither tool exposes these fields. Similarly, `compliance_check` hides its `jurisdiction`, so `risk_assess` may apply a risk model calibrated to a different regulatory framework. The `confidence_interval` from risk assessment silently determines which `accounting_standard` the ledger uses. None of these are type errors—every schema validates. The coherence fee of 7 says: there are 7 independent convention dimensions that no bilateral check can verify.

Fee and disclosure. The coherence fee is $\varphi = 7$ with 8 blind spots and 15 bridges. The minimum disclosure set has $|S| = 7 = \varphi$:

Tool	Field
`currency_convert`	`rounding_mode`
`compliance_check`	`jurisdiction`
`risk_assess`	`risk_model_version`
`risk_assess`	`confidence_interval`
`ledger_write`	`decimal_precision`
`audit_log`	`timezone`
`notification`	`locale`

Bridges suggest 15 fixes; the disclosure set achieves the same fee reduction with 7—a factor of 2.1× savings. The matroid structure explains why: when `timezone` on `audit_log` is hidden across multiple edges, a single disclosure resolves all of them.

Boundary fee under partition. Splitting into front-office {`price_fetch`, `currency_convert`, `compliance_` and back-office {`ledger_write`, `audit_log`, `notification`, `archive`} gives local fees $(2, 3)$ and boundary fee 2. The boundary fee accounts for two hidden conventions (`confidence_interval` and `decimal_precision`) that cross the front/back-office boundary: internal to each sub-workflow when viewed locally, but creating obstruction in the flat expansion.

Conditional resolution. Removing `archive` and creating a placeholder with two open-port dimensions gives three fee values:

- **Baseline fee** (known sub-composition only): $\varphi = 6$.
- **Worst-case fee** (placeholder with $O = \emptyset$): $\varphi = 7$, with two boundary obligations (`archive_id`, `retention_policy`).
- **Resolved fee** (real `archive` tool substituted): $\varphi = 7$. The real tool exposes `archive_id` but hides `retention_policy`, meeting one obligation and leaving one remaining. Fee delta: $7 - 7 = 0$ (the real tool is as opaque as the placeholder for `retention_policy`).

This demonstrates resolution monotonicity (corollary 6.4): the resolved fee equals the worst-case fee here because `archive` happens to hide the same field as the placeholder. A more transparent replacement would yield $\varphi < 7$.

8 Empirical Results

All theorems are computationally verified against 10 bundled compositions (including the 8-tool financial settlement pipeline) in Bulla v0.34, with sampled binary partitions and tower law pairs. Separately, an adversarial survey of 10,000 random compositions checked 635,095 partition pairs for submodularity.

Metric	Value
Block rank formula + non-negativity verified	120/120
Tower law pairs verified (sampled)	2,778/2,778
$\lvert S\rvert = \varphi$ verified	10/10
Submodularity (adversarial)	disproved (4,061/635,095)
Resolution monotonicity verified	8/8
8-tool financial settlement pipeline	
φ (fee)	7
$\lvert S\rvert$ (disclosure set)	7
|bridges|	15
β (front/back partition)	2
Conditional: baseline → worst-case → resolved	$6 \to 7 \to 7$

Every composition with nonzero fee produces nonzero boundary fee for most partitions: the hierarchical blind spot is the dominant regime, not a corner case. Boundary fee is nonzero in over 91% of sampled binary partitions across all compositions.

Remark 8.1 (Trace gap). The Frobenius trace gap $\|\delta_{\text{full}}\|_F^2 - \|\delta_{\text{obs}}\|_F^2$ has been considered as a continuous spectral refinement of the integer fee. Since all coboundary entries are 0 or ± 1, this reduces to the count of nonzero entries in hidden columns:

$$\text{trace gap} = \sum_{b \in \text{blind spots}} \big[\mathbf{1}[b.\text{from\_hidden}] + \mathbf{1}[b.\text{to\_hidden}]\big],$$

a weighted blind-spot count derivable from the existing diagnostic. It can be positive when $\varphi = 0$ (hidden columns in the span of observable columns), and adds no information beyond the blind-spot structure. A genuine spectral refinement requires the eigenvalue spectrum of the sheaf Laplacian, which we defer to future work.

9 Related Work

Contract-based design and assume-guarantee reasoning. Interface theories [8] define compatibility conditions between open components; Benveniste et al. [9] systematize the algebra of contracts (assumptions on the environment, guarantees on the component) for system design. Pasareanu et al. [10] show how environment assumptions can be learned incrementally. The coherence fee's boundary obligations are the analog for tool compositions: computed from observable schema structure rather than specified manually. The fee quantifies the cost of *not having* a contract—the number of convention dimensions a contract would need to cover.

Sheaf cohomology for data integration. Curry [4] develops the cellular sheaf framework with computable cohomology, including Mayer–Vietoris sequences for decomposition. Robinson [5] applies sheaf theory to signal processing and sensor fusion, using the sheaf Laplacian to detect inconsistency. Ghrist [6] provides accessible background on applied algebraic topology. Hansen and Ghrist [7] study opinion dynamics via sheaf Laplacians. The coherence fee instantiates the same mathematical framework for tool compositions: the coboundary operator δ^0 is the signed incidence map of a cellular sheaf, and the fee is a computable rank obstruction [2]. The boundary fee is the partition-relative analog of sheaf-theoretic gluing constraints, and the vanishing corollary (corollary 3.2) is the linearized version of the SCPI contractible-nerve vanishing theorem [3].

Tool-composition ecosystems. The Model Context Protocol [11] standardizes how processes expose tool interfaces; orchestration frameworks such as AutoGen [12] and LangGraph [13] define how subsystems communicate and delegate. Tool-using language models [14, 15] compose tools dynamically at inference time. These systems define the *transport layer*; the coherence fee operates above it. Given a composition graph (which any framework can export), Bulla diagnoses what the framework cannot check: convention consistency across tool boundaries. A reference LangGraph integration demo (bundled with the implementation) constructs a 4-tool trade pipeline, runs it through LangGraph with all schemas validating, then diagnoses a coherence fee of 3 from hidden conventions invisible to the orchestrator. This is orthogonal to and composable with existing orchestration.

10 Conclusion

We have introduced the coherence fee as a computable diagnostic for hidden convention risk in hierarchical tool compositions, proved that hierarchical decomposition adds non-negative cost (with the tower law quantifying the cost per level), and shown that a minimum disclosure set of exactly φ fields suffices to eliminate all blind spots. The boundary fee is monotone on the partition lattice but neither a valuation nor submodular, as established by adversarial computational survey.

The coherence fee measures *structural verifiability*, not semantic correctness. A fee of zero means every convention dimension *can* be checked from observable schemas—not that every convention *is* correct. The gap between verifiability and correctness is where domain-specific validation begins; the fee tells you where to look.

Limitations. The model assumes stable tool signatures for the duration of analysis; runtime schema drift is not handled. Obligation satisfaction is verified structurally (field presence) rather than semantically (value correctness). The current implementation treats all conventions as equally weighted.

Future work. A spectral refinement of the fee via the eigenvalue spectrum of the sheaf Laplacian would provide a continuous measure of convention risk, weighting each hidden dimension by its connectivity. The witness capsule protocol would formalize assume-guarantee contracts for hierarchical tool compositions, enabling plan-time verification across delegation boundaries. Integration with tool-orchestration frameworks (LangGraph, AutoGen) would validate the theory against real-world composition patterns and surface the fee in existing developer workflows. Compositional verification techniques and runtime monitoring [10] could extend the static analysis to dynamic composition graphs.

References

[1] J. Komkov, "The coherence fee: Edge-local blindness at the string-table seam and the topological price of cross-system composition," Technical Report, Res Agentica, February 2026. `https://github.com/jkomkov/bulla`

[2] J. Komkov, "Sheaf-theoretic diagnostics for tool compositions," Technical Report, Res Agentica, 2026. `https://github.com/jkomkov/bulla`

[3] J. Komkov, "Predicate invention under sheaf constraints," Technical Report, Res Agentica, 2026. `https://github.com/jkomkov/bulla`

[4] J. Curry, *Sheaves, cosheaves and applications*, PhD thesis, University of Pennsylvania, 2014.

[5] M. Robinson, *Topological Signal Processing*, Springer, 2014.

[6] R. Ghrist, *Elementary Applied Topology*, Createspace, 2014.

[7] J. Hansen and R. Ghrist, "Toward a spectral theory of cellular sheaves," *Journal of Applied and Computational Topology*, vol. 3, pp. 315–358, 2019.

[8] L. de Alfaro and T. A. Henzinger, "Interface automata," in *Proc. 9th ESEC/FSE*, ACM, 2001, pp. 109–120.

[9] A. Benveniste, B. Caillaud, D. Nickovic, R. Passerone, J.-B. Raclet, P. Reinkemeier, A. Sangiovanni-Vincentelli, W. Damm, T. Henzinger, and K. Larsen, "Contracts for system design," *Foundations and Trends in Electronic Design Automation*, vol. 12, no. 2–3, pp. 124–400, 2018.

[10] C. S. Păsăreanu, D. Giannakopoulou, M. G. Bobaru, J. M. Cobleigh, and H. Barringer, "Learning to divide and conquer: Applying the L* algorithm to automate assume-guarantee reasoning," *Formal Methods in System Design*, vol. 32, no. 3, pp. 175–205, 2008.

[11] Anthropic, "Model context protocol specification," 2024. `https://modelcontextprotocol.io/specification`

[12] Q. Wu, G. Banberanber, Y. Hou, H. Ye, L. Zheng, A. Kazemitabaar, N. Cai, D. Jiang, Z. Li, A. H. Awadallah, R. Ronen, and C. Wang, "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," *arXiv:2308.08155*, 2023.

[13] LangChain, "LangGraph: Build stateful, multi-actor applications with LLMs," 2024. `https://github.com/langchain-ai/langgraph`

[14] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," in *Proc. NeurIPS*, 2023.

[15] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, R. Tian, R. Xie, J. Zhou, M. Gerstein, D. Li, Z. Liu, and M. Sun, "ToolLLM: Facilitating large language models to master 16000+ real-world APIs," in *Proc. ICLR*, 2024.

[16] J. Komkov, "Signed-Incidence Structure in Compositional Verification: A Structural Note," `papers/signed-incidence/paper/signed-incidence.pdf`, 2026.

The Column-Matroid Backbone of a Composition-Coherence Fee

A Structural Identity and Its Corollaries

John Komkov
Independent Researcher
https://github.com/jkomkov/bulla

April 2026

Abstract

The Bulla protocol assigns a non-negative integer coherence fee $\mathrm{fee}(G)$ to each composition G of tool specifications, measuring semantic dimensions whose consistency is structurally unverifiable from observable schemas alone. We report that $\mathrm{fee}(G)$ equals the corank of the observable column set in the linear matroid on the full coboundary's columns—a one-line proof from a column-compression identity discovered by a post-hoc audit of a pre-registered spectral experiment. This structural identity yields a closed-form pairwise formula:

$$\mathrm{fee}(A, B) = U(A) + U(B) + |\mathrm{shared\_dims}(A, B)|$$

where $U(X)$ is a per-server invariant (Lean-verified) and |shared_dims| counts bilateral semantic dimensions (proved under DFD; verified on 703 real-schema compositions). The backbone theorem also proves the additive decomposition as matroid direct-sum additivity and closes the rank$\leftrightarrow$ H^1 formal target via two lines of rank-nullity (Lean-verified, zero `sorry`). An empirical density law localizes the connecting homomorphism of the associated long exact sequence strictly in the interior of its attainable range on all 240 non-trivial compositions tested. The discovery is an instance of *Type D productive falsification*: the pre-registered prediction held, but for a stronger and simpler reason than anticipated. A companion signed-incidence note [6] supplies the field-independence and endpoint-coupling boundaries used throughout: the fee agrees across coefficient fields as a numerical invariant, and pairwise same-dimension blocks are bounded by rank 2 without implying any broader typed-repair characterization.

1 Introduction

The Bulla protocol assigns a non-negative integer coherence fee $\mathrm{fee}(G)$ to each composition G of two or more tool specifications. The fee measures the codimension of an observable coboundary inside the full internal-state coboundary of a cellular sheaf constructed from the composition graph [1]. The protocol note [2] wrote $\mathrm{fee}(G) = \dim H^1(G; \mathcal{F}/\mathcal{O})$, a cohomological framing suggestive enough to motivate the construction but not proven in the strict sense.

This paper reports a structural identity that closes the gap and produces a closed-form pairwise formula for the fee.

The experiment. A pre-registered Cauchy eigenvalue interlacing test on the 703-composition real-schema corpus (§1.1) passed all 703 compositions with zero violations, under the hypothesis that Bulla's coboundary pair instantiates the spectral theory of cellular sheaves [3].

The audit. A post-sweep audit discovered that, on every composition, $\delta_{\mathrm{obs}} = \delta_{\mathrm{full}}[:, \mathcal{O}]$ exactly over $\mathbb{Q}$: the observable coboundary is the principal column submatrix of the full coboundary.

The interlacing holds as a classical submatrix theorem, independent of [3]. We call this *Type D productive falsification*: the pre-registered prediction held, but for a stronger and simpler structural reason than anticipated. The backbone theorem emerged not from searching for a matroid identity but from a standing audit protocol—inspect the construction that produced the data, not just the measurement—applied to a test whose prediction *passed*.[1]

The consequences. The column-compression identity makes fee(G) the corank of a representable matroid. Four corollaries carry the paper's argument:

1. The additive decomposition theorem [4] is matroid direct-sum additivity (Corollary 2.8).
2. The rank$\leftrightarrow H^1$ identity is proven and Lean-verified (Corollary 2.9).
3. The unilateral fee is a stable per-server invariant, Lean-verified (Corollary 2.13).
4. The bilateral fee counts shared semantic dimensions exactly, yielding the closed-form pairwise formula (Corollary 2.14, Theorem 2.15).

Claim discipline. What is established here is the backbone identity and the corollaries that genuinely follow from it. What failed productively was the older attempt to motivate these consequences through a heavier sheaf-spectral route than the code audit warranted. What remains open is the correct dimension-aware coupled-disclosure objective; the signed-incidence note [6] rules out the separable partition story, but this paper does not claim a replacement characterization.
Further corollaries—T-concentration as uniform-matroid proximity, the repair filtration as a matroid flag, a flats-interaction reformulation of the Unilateral-Bilateral Decomposition, and a row-corank reframing of the submodularity question—appear in [5].

Organization. §2 states the backbone theorem and its four corollaries. §3 treats the matroid direct-sum proof of Theorem 1. §4 reports an empirical density law. §5 updates the three original Bulla open problems.

1.1 The 703-Composition Real-Schema Corpus

All empirical claims are made on a corpus of MCP (Model Context Protocol) tool-composition instances derived from 57 server manifests scraped from the public `oslook/mcp-servers-schemas` repository on 2026-04-07. The corpus is filtered to servers with `MIN_SCHEMA_FIELDS` ≥ 3 (38 of 57), giving a pairwise closure of $\binom{38}{2} = 703$ compositions.

The Column-Compression Identity (Theorem 2.4) and the Backbone Theorem (Theorem 2.6) are structural and hold for any Bulla composition. The density-law measurement (§4) is corpus-specific.

2 The Column-Matroid Backbone

2.1 Setup

Let $G = (T, E, D)$ be a Bulla composition of tools T, directed edges E between tools, and semantic dimensions D attached to edges. For each tool $t \in T$, write $t.\text{int}$ for the internal-state fields and $t.\text{obs} \subseteq t.\text{int}$ for the observable subset.

Definition 2.1 (Coboundary construction)**.** The Bulla coboundary construction produces two $\mathbb{Q}$-linear maps sharing a common codomain:

$$\delta_{\text{full}}(G) : C^0_{\text{full}}(G) \longrightarrow C^1(G), \tag{1}$$

$$\delta_{\text{obs}}(G) : C^0_{\text{obs}}(G) \longrightarrow C^1(G), \tag{2}$$

[1] Type D is the most dangerous productive-falsification species to miss: the prediction-verification pattern looks like ordinary confirmatory success. The taxonomy and the commit-before-measure discipline that produced the backbone are described in [5].

where $C^0_{\text{full}} = \bigoplus_t \mathbb{Q}[t.\text{int}]$, $C^0_{\text{obs}} = \bigoplus_t \mathbb{Q}[t.\text{obs}]$, and $C^1 = \bigoplus_{(e,d)} \mathbb{Q}$ is the edge-dimension cochain space. The matrix entry at row (e, d) and column (t, f) is -1 if $t = e.\text{src}$ and $f = d.\text{from}$; $+1$ if $t = e.\text{dst}$ and $f = d.\text{to}$; and 0 otherwise, restricted to $f \in t.\text{int}$ (resp. $t.\text{obs}$).

Definition 2.2 (Coherence fee)**.** $\text{fee}(G) := \text{rank}_{\mathbb{Q}}(\delta_{\text{full}}(G)) - \text{rank}_{\mathbb{Q}}(\delta_{\text{obs}}(G)) \geq 0$.

Remark 2.3 (Shared codomain)*.* Both coboundaries map into the same $C^1(G)$: the edge-dimension basis is constructed identically regardless of vertex-stalk scope. This fact is load-bearing for Corollary 2.9.

2.2 The Column-Compression Identity

Theorem 2.4 (Column-Compression Identity)**.** *For every Bulla composition G, there is a column index set $\mathcal{O}(G) \subseteq \{1, \ldots, n_{\text{full}}\}$ such that $\delta_{\text{obs}}(G) = \delta_{\text{full}}(G)[:, \mathcal{O}(G)]$ as an equality of $\mathbb{Q}$-valued matrices.*

Proof. Since $t.\text{obs} \subseteq t.\text{int}$ per tool, every observable basis element appears in the full basis. Let $\mathcal{O}(G)$ be the full-basis indices of observable pairs. Every ± 1 entry in δ_{obs} appears at the same row and corresponding column in δ_{full}; columns not in $\mathcal{O}$ (hidden fields) are absent from δ_{obs}. □

Empirical verification. 703/703 compositions, zero entry mismatches on the 240 non-trivial cases.

2.3 The Column Matroid and the Backbone Theorem

Definition 2.5 (Column matroid)**.** $M(G)$ is the linear matroid on ground set $E(G) := \{1, \ldots, n_{\text{full}}\}$ with rank function $\text{rank}_{M(G)}(S) := \dim_{\mathbb{Q}} \text{span}_{\mathbb{Q}}\{\delta_{\text{full}}(G)[:, j] : j \in S\}$.

Theorem 2.6 (Column-Matroid Backbone)**.** *For every Bulla composition G,*

$$\text{fee}(G) \;=\; \text{corank}_{M(G)}(\mathcal{O}(G)) \;:=\; \text{rank}_{M(G)}(E(G)) - \text{rank}_{M(G)}(\mathcal{O}(G)).$$

Proof. $\text{rank}_{M(G)}(E(G)) = \text{rank}_{\mathbb{Q}}(\delta_{\text{full}})$. $\text{rank}_{M(G)}(\mathcal{O}) = \text{rank}_{\mathbb{Q}}(\delta_{\text{full}}[:, \mathcal{O}]) = \text{rank}_{\mathbb{Q}}(\delta_{\text{obs}})$ by Theorem 2.4. Subtract. □

2.4 Corollary 1: Additive Decomposition as Matroid Direct Sum

Definition 2.7 (Dimension-Field Disjointness)**.** A composition G satisfies *DFD* if no (t, f) pair is referenced by two distinct dimension names in D.

Corollary 2.8. *Under DFD, $M(G) = \bigoplus_d M_d(G)$ as a matroid direct sum, and* $\text{fee}(G) = \sum_d \text{fee}_d(G)$.

Proof. Under DFD, each column of δ_{full} has nonzero entries only in rows of the single dimension referencing its (t, f) pair, so δ_{full} is block-diagonal with one block per dimension. Let E_d be dimension d's columns; the E_d partition $E(G)$ and $M_d := M(G)|_{E_d}$. Direct-sum rank additivity gives $\text{rank}_{M(G)}(S) = \sum_d \text{rank}_{M_d}(S \cap E_d)$. Applying Theorem 2.6 per dimension:

$$\text{fee}(G) = \sum_d \big[\text{rank}_{M_d}(E_d) - \text{rank}_{M_d}(\mathcal{O} \cap E_d)\big] = \sum_d \text{fee}_d(G). \qquad \square$$

2.5 Corollary 6: Cohomological Gap (Problem 3 Closure)

Corollary 2.9. *For every Bulla composition G,*

$$\text{fee}(G) \;=\; \dim_{\mathbb{Q}} H^1(\delta_{\text{obs}}) - \dim_{\mathbb{Q}} H^1(\delta_{\text{full}})$$

where $H^1(\delta) := \text{coker}(\delta) = C^1/\,\text{im}(\delta)$.

Proof (rank-nullity). The chain complex $C^0 \to C^1 \to 0$ has no C^2, so $H^1(\delta) = \text{coker}(\delta)$. By rank-nullity, $\dim \text{coker}(\delta) = \dim C^1 - \text{rank}(\delta)$. Both coboundaries share C^1 (Remark 2.3), so:

$$\dim H^1(\delta_{\text{obs}}) - \dim H^1(\delta_{\text{full}}) = \big[\dim C^1 - \text{rank}(\delta_{\text{obs}})\big] - \big[\dim C^1 - \text{rank}(\delta_{\text{full}})\big] = \text{fee}(G). \quad \square$$

Proof (long exact sequence). The short exact sequence $0 \to \mathcal{O} \to \mathcal{F} \to \mathcal{F}/\mathcal{O} \to 0$ induces a LES. Since $C^2 = 0$ and the relative complex has trivial degree-1 term ($C^1_{\text{full}} = C^1_{\text{obs}}$), we get $H^1(\mathcal{F}/\mathcal{O}) = 0$ and $\text{fee}(G) = \dim \text{im}(\partial)$ where $\partial : H^0(\mathcal{F}/\mathcal{O}) \to H^1(\mathcal{O})$ is the connecting homomorphism. $\square$

Lean verification. Formally verified in Lean 4 / Mathlib v4.28.0 (Aristotle UUID `044a00b1`). Three theorems: `fee_h1_additive`, `fee_eq_h1_difference`, `range_obs_le_range_full`. No `sorry`.

Remark 2.10 (Upper bound). Since ∂ is linear, $\text{fee}(G) \le n_{\text{full}} - n_{\text{obs}} = |\mathcal{O}^c|$. The gap measures *column redundancy*—hidden-column directions already implied by observed relations. Section 4 measures this.

Remark 2.11 (Notation retirement). The informal $\text{fee} = \dim H^1(G; \mathcal{F}/\mathcal{O})$ is retired: under the literal relative-complex reading, $H^1(\mathcal{F}/\mathcal{O}) = 0$. The honest content is $\dim H^1(\delta_{\text{obs}}) - \dim H^1(\delta_{\text{full}})$.

2.6 Corollaries 8a and 8b: the Closed-Form Pairwise Fee

For a pairwise composition $G = A + B$, define the *unilateral fee* $U(X) := \text{fee}(X, \varnothing)$ (composition of X with a synthetic null probe) and the *bilateral correction* $B(A, B) := \text{fee}(A, B) - U(A) - U(B)$.

Definition 2.12 (Convention-Hidden Partition)**.** A composition G satisfies the *convention-hidden partition* (CHP) if every field f that appears as d.from or d.to for some dimension d satisfies $f \notin t$.obs on the tool t that carries it. Equivalently: inferred convention fields are never observable.

CHP holds by construction in the current Bulla classifier, which removes inferred convention fields from the observable schema. It is stated as an explicit hypothesis so that the theorem's scope is clear: a future system that observes some convention fields would require a different analysis.

Corollary 2.13 (Unilateral stability)**.** $\text{fee}_A(A + B) = U(A)$ *and* $\text{fee}_B(A + B) = U(B)$*, where* fee_X *denotes the fee contribution of* X*-only rows. The composition partner does not affect the unilateral row-class fee.*

Verification. 703/703. Lean-verified (UUID `01333bb4`, 2 theorems, 0 `sorry`).

Corollary 2.14 (Bilateral fee counts shared dimensions)**.** *Under DFD and CHP,* $B(A, B) = |\text{shared\_dims}(A, B)|$*, where* shared_dims *counts semantic dimensions with bilateral rows.*

Proof. By Corollary 2.8, $B(A,B) = \sum_d B_d$ where $B_d := \text{fee}_d - U_d(A) - U_d(B)$. If d is not shared, all dimension-d rows are unilateral, $\text{fee}_d = U_d(A) + U_d(B)$, and $B_d = 0$.

If d is shared, let p tools in A and q tools in B carry dimension d. By CHP, both d.from and d.to are hidden columns. Under DFD the $p+q$ hidden columns appear only in dimension-d rows. The unilateral A-rows (differences of A-columns) have rank $p-1$; the unilateral B-rows rank $q-1$; one bilateral row $-e_{(a_i,f_A)} + e_{(b_j,f_B)}$ is independent of both subspaces (it has nonzero projection onto both column groups). Equivalently, the dimension-d rows form the signed incidence matrix of a connected bipartite graph on $p+q$ vertices, which has rank $p+q-1$ over any field. So $\text{rank}_d^{\text{full}} = p+q-1$. In δ_{obs} all $p+q$ columns are absent, giving $\text{rank}_d^{\text{obs}} = 0$. Therefore

$$B_d = (p+q-1) - (p-1) - (q-1) = 1. \qquad \square$$

Empirical verification. 703/703 (all satisfy both DFD and CHP), Pearson $r = 1.000$ on the 25 nonzero-B cases. Pre-registered for cross-registry replication.

Corollaries 2.13 and 2.14 together yield:

Theorem 2.15 (Closed-form pairwise formula)**.** *Under DFD and CHP, for every pairwise composition* (A, B)*,*

$$\boxed{\text{fee}(A,B) \;=\; U(A) + U(B) + |\text{shared\_dims}(A,B)|}$$

where $U(X)$ *is a per-server invariant and* $|\text{shared\_dims}|$ *counts the semantic dimensions with bilateral rows.*

Proof. Immediate from Corollary 2.13 ($\text{fee}_A = U(A)$, $\text{fee}_B = U(B)$) and Corollary 2.14 ($B(A,B) = |\text{shared\_dims}|$). $\square$

Theorem 2.15 decomposes the fee into local invariants (unilateral, depending only on each server's schema) and a pairwise interaction term (bilateral, counting shared dimensions). The formula requires no matrix computation—only schema inspection.

Example 2.16. `github + notion`: $U($`github`$) = 10$, $U($`notion`$) = 1$, two shared dimensions (`date_format_match`, `sort_direction_match`), so fee $= 10 + 1 + 2 = 13$. Verified.

3 Theorem 1 under DFD

The additive decomposition theorem of [4] states that, under DFD, the coherence fee splits by named semantic dimension: $\text{fee}(G) = \sum_d \text{fee}_d(G)$. Corollary 2.8 gives this as a one-line consequence of matroid direct-sum additivity.

What Corollary 2.8 adds beyond the original linear-algebraic proof sketch is a *name* for the structure. The statement "$M(G)$ decomposes as a matroid direct sum under DFD" carries through to other matroid operations (intersection, union, truncation, minors) without re-proving each. It also sharpens the open question of whether DFD is necessary: the question becomes "characterize the matroid-direct-sum compositions that violate DFD," which has a precise formulation in terms of column support overlap in non-block-diagonal representations.

Empirical verification. The per-dimension fee identity $\text{fee}(G) = \sum_d \text{fee}_d(G)$ was verified on 1596 compositions (the full pairwise closure of the 57-manifest registry, not restricted to the real-schema subset). Zero violations. On the real-schema subset, 45 of 45 compositions with nonzero per-dimension fees satisfy the identity exactly.

DFD as hypothesis. DFD holds on all 703 real-schema compositions, so the hypothesis is not binding on the current corpus. Whether it is necessary in general—whether matroid direct sum can hold without DFD—remains open.

4 Observation 1: the Interior-Regime Density Law

Corollary 2.9 gives $\mathrm{fee}(G) = \dim \mathrm{im}(\partial)$ where $\partial : H^0(\mathcal{F}/\mathcal{O}) \to H^1(\mathcal{O})$. The normalized ratio $\rho(G) := \mathrm{fee}(G)/(n_{\mathrm{full}} - n_{\mathrm{obs}})$ measures how surjective ∂ is.

Observation 4.1 (Interior regime)**.** On every non-trivial composition in the 703 corpus ($n_{\mathrm{full}} - n_{\mathrm{obs}} > 0$), the ratio $\rho(G)$ lies strictly in $(0, 1)$. Neither extremum is realized.

Table 1: Distribution of $\rho(G)$ on 240 non-trivial compositions.

mean	median	stdev	min	max
0.603	0.667	0.175	0.250	0.933

68.3% of compositions land in $\rho \in [0.50, 0.75)$—tight enough to be called a *density law*: the ratio of rank-contributing hidden columns to total hidden columns is concentrated around a single value across structurally distinct compositions.

In applied terms: on real MCP compositions, the cohomological obstruction is always present, never saturated, and its magnitude is regularized to roughly 60% of its attainable maximum.

Pre-registered replication. A second MCP registry snapshot will be procured and the same ρ measurement run. Four locked branches: *O1-pass* (mean $\in [0.50, 0.70]$, modal mass $> 50\%$, strict interior); *O1-narrow* (interior preserved, concentration shifts); *O1-broken* (some $\rho \in \{0, 1\}$); *O1-trivial* (no non-trivial compositions).

5 Three Open Problems as Backbone Reductions

The Bulla research program posed three structural open problems. The backbone theorem (Theorem 2.6) is the protagonist of all three updates: each problem is either closed or reduced to a sharper question by the matroid framing.

5.1 Problem 3 (Rank$\leftrightarrow H^1$): Closed

Corollary 2.9 proves $\mathrm{fee}(G) = \dim H^1(\delta_{\mathrm{obs}}) - \dim H^1(\delta_{\mathrm{full}})$ as an identity of $\mathbb{N}$-valued invariants. The Lean 4 formalization (Aristotle UUID `044a00b1`) closes the problem in three theorems and zero `sorry`, reducing what was projected as a multi-sprint cellular-cohomology formalization to an application of `Submodule.finrank_quotient_add_finrank` (rank-nullity for cokernels).

The key structural insight: the Problem 3 target was hard only under the assumption that the proof required sheaf-theoretic machinery. The column-compression identity (Theorem 2.4) shows the two coboundaries are related by a column inclusion, not a generic morphism. Under column inclusion, rank-nullity suffices.

5.2 Problem 1 (Termination): Structural Half Closed

The repair loop produces an increasing sequence $\mathcal{O}_0 \subseteq \mathcal{O}_1 \subseteq \cdots \subseteq E(G)$ of observable column sets. Under the backbone, this is a flag in $M(G)$ with rank jumps in $\{0, 1\}$ at each step (a matroid flag property).

The structural half of Problem 1—that the repair loop terminates in at most $\mathrm{fee}(G)$ rank-increasing steps—was Lean-verified (UUID `ff78c99c`, 3 theorems). The backbone adds a runtime assertion: after each confirmed probe, check whether the new column lies in $\mathrm{cl}_{M(G)}(\mathcal{O}_i)$. If it does, the probe was rank-redundant and a `ProbeSoundnessViolation` is raised.

The practical half—bounding the number of oracle calls needed to produce a rank-increasing step—remains open and depends on oracle-quality assumptions outside the matroid structure.

5.3 Problem 2 (Submodularity): Sharpened

The Bulla boundary fee $\mathrm{bf}(P)$ was tested for submodularity on the partition lattice and found 4061 violations across the corpus. Under the backbone, $\mathrm{bf}(P)$ is a difference of row-matroid corank functions:

$$\mathrm{bf}(P) = \mathrm{corank}_{M_{\mathrm{row}}^{\mathrm{full}}}(\mathrm{intra}(P)) - \mathrm{corank}_{M_{\mathrm{row}}^{\mathrm{obs}}}(\mathrm{intra}(P)).$$

Each corank is individually supermodular, but a difference of supermodular functions is neither sub- nor supermodular in general. The 4061 violations are structurally explained: they measure the direction in which the difference fails for this specific pair of row matroids.

The sharpened research question is: under what conditions on the weak map $M_{\mathrm{row}}^{\mathrm{obs}} \to M_{\mathrm{row}}^{\mathrm{full}}$ is the corank difference sub- or supermodular? This is a bona fide matroid-theoretic question with real structure, connected to the weak-maps literature (Lovász, Kung).

A Lean Verification Summary

Twelve theorems verified in Lean 4 against Mathlib, zero `sorry`:

File	UUID	Thms	Content
BullaCorollary6	`044a00b1`	3	rank$\leftrightarrow H^1$ (Problem 3)
TerminationStructural	`ff78c99c`	3	Repair loop bound (Problem 1)
UBD1	`edb6c604`	4	Range incl. + DFD + $B \geq 0$
Corollary8a	`01333bb4`	2	Unilateral stability
Total		**12**	**0 sorry**

B Code and Proof Artifacts

The column-submatrix audit, the $\rho(G)$ measurement script, and all four Lean files are available at `https://github.com/jkomkov/bulla`. The SHA-256 anchor chain of the Column-Matroid Backbone memo (v1–v5) and all pre-registration anchors are recorded in the repository's technical claim ledger.

References

[1] J. Komkov, "The Coherence Fee: Measuring Hidden Convention Risk in Hierarchical Tool Compositions," 2026.

[2] J. Komkov, "Bulla Protocol Note," 2026.

[3] J. Hansen and R. Ghrist, "Toward a spectral theory of cellular sheaves," *J. Appl. and Comput. Topology*, vol. 3, no. 4, pp. 315–358, 2019.

[4] J. Komkov, "Additive decomposition of the coherence fee under Dimension-Field Disjointness," 2026.

[5] J. Komkov, "The Column-Matroid Backbone of the Bulla Coherence Fee," internal research memo v5, 2026.

[6] J. Komkov, "Signed-Incidence Structure in Compositional Verification: A Structural Note," `papers/signed-incidence/paper/signed-incidence.pdf`, 2026.

The Witness Gram as a Kron-Reduced Laplacian of Hidden Convention Risk

John Komkov
Independent Researcher
https://github.com/jkomkov/bulla

April 2026

Abstract

The witness Gram $K(G) = H^\top(I - P_O)H$ of a Bulla tool composition is always a Kron-reduced graph Laplacian — literally the matrix Gabriel Kron introduced in 1939 for reducing electrical networks to sub-networks of interest. Under Dimension-Field Disjointness (DFD) it decomposes as a direct sum of Kron reductions of per-dimension carrier-graph Laplacians; without DFD it remains a single Kron-reduced Laplacian, of the merged coboundary-incidence graph. This identifies the canonical object sitting at the corank of the backbone theorem [2]: the coherence fee equals $\mathrm{rank}(K(G))$, and the bases of $K(G)$'s column matroid reproduce the minimum disclosure sets of [1]. We show that the effective resistance between two hidden fields in the Kron-reduced witness geometry quantifies disclosure substitutability: the smaller the resistance, the more equivalent the two fields are as repair choices. All structural identities are verified on the 703-composition real-schema MCP registry corpus in exact rational arithmetic (703/703 rank identities, 240/240 leverage predictions including 37 natural partial-CHP regime breaks). The paper closes the program's structural theory and identifies the frontier where field-level diagnostics become discriminative. A companion signed-incidence note [8] supplies the pairwise endpoint-coupling boundary used here: same-dimension witness blocks are rank-bounded by construction, and their frequent rank-2 realization is evidence against separable typed-repair constraints, not a characterization of the true dimension-aware objective.

1 Introduction

Master sentence. *The witness Gram of a Bulla composition is always a Kron-reduced graph Laplacian. Dimension-Field Disjointness controls whether that Laplacian decomposes as a direct sum over dimensions; the Convention-Hidden Partition controls whether the Kron reduction is trivial.*

The coherence fee [1] of a tool composition G is an integer measuring the codimension of an observable coboundary inside a full internal-state coboundary. A companion result [2] shows that the fee equals the corank of the observable column set in a linear matroid on the columns of the full coboundary, and that the closed-form pairwise fee $\mathrm{fee}(A, B) = U(A)+U(B)+|\mathrm{shared\_dims}(A, B)|$ holds under two structural hypotheses (DFD and CHP) on every one of 703 real-schema MCP compositions tested.

This paper asks a more pointed question. The backbone identifies the fee as *the rank of an object*; what is the object, and what can it tell us that the integer cannot? We give three answers, developed through one central theorem:

1. The object is the *witness Gram* $K(G) = H^\top (I - P_O) H$ — a symmetric positive-semidefinite matrix on the hidden-field coordinate space, of exact rational entries, with rank equal to the fee.
2. Under DFD, the witness Gram decomposes as a direct sum of *Kron reductions* of per-dimension carrier-graph Laplacians — literally the matrices Gabriel Kron introduced in 1939 for reducing electrical networks to sub-networks of interest [3].
3. In the Kron-reduced geometry, the classical effective resistance between two hidden fields quantifies their *disclosure substitutability*: small resistance means disclosing either field largely resolves the other's convention risk; large resistance (within a component) means the fields are structurally distinct witness obligations.

Claim discipline. What is established here is the object-level witness geometry and the Laplacian structure it carries. What failed productively was the older hope that same-dimension typed constraints might collapse to a separable partition rule. What remains open is the correct coupled-disclosure optimization problem; the signed-incidence note [8] supplies the honest middle-layer boundary for that statement, and this paper stays on the geometry side of the line.

A concrete example, before the machinery. Consider two compositions in the 703-corpus with identical fee = 2: the `financial_pipeline` (3 tools, 4 hidden fields spanning 2 convention dimensions, each a 2-vertex carrier graph) and a subset of the `airtable-mcp + mcp-xmind` pairing (3 hidden `path`-related fields in one 5-vertex carrier graph with 2 observable carriers Kron-reduced out). The bare integer fee cannot tell these apart. The witness Gram does:

$$K(\texttt{financial\_pipeline}) = \begin{bmatrix} 1 & -1 & & \\ -1 & 1 & & \\ & & 1 & -1 \\ & & -1 & 1 \end{bmatrix} = L(K_2) \oplus L(K_2),$$

$$K(\texttt{airtable+mcp-xmind}_{\text{path dim}}) = \tfrac{5}{3} \cdot L(K_3) \; = \text{Schur}\big(L(K_5) \,/\, O_d\big) \text{ for } |O_d| = 2.$$

Two different graph Laplacians, two different bases, two different effective-resistance geometries — and therefore two different disclosure prescriptions — at the same fee. In the first, any one field per dimension suffices (1 spanning tree per K_2-block, giving a single minimum basis $\{\text{day\_convention}, \text{risk\_metric}\}$ up to per-block choice). In the second, any two of the three hidden `path`-fields form a valid disclosure set (3 bases on K_3 by Matrix-Tree), and the three fields are equally good disclosure substitutes at effective resistance $\frac{2}{5}$ apart.

Organization. §2 proves the rank identity (Tier 1, no hypotheses). §3 proves the Kron-reduction theorem under DFD (Tier 2, the centerpiece). §4 states and proves the four corollaries that fall out of Tier 2 (leverage, no-coloop, basis-count via Matrix-Tree, effective resistance as disclosure substitutability). §5 studies regime breaks: 37 natural partial-CHP compositions in the corpus, plus synthetic DFD and bridge-topology constructions. §6 reports the corpus-wide verification in exact rational arithmetic. §7 closes with operational interpretation.

2 The witness Gram and its rank identity

Fix a Bulla composition $G = (T, E, D)$ with tool set T, directed edges E, and semantic dimensions D. For each tool $t \in T$ let $t.\text{int}$ and $t.\text{obs} \subseteq t.\text{int}$ denote its internal and observable field sets. Write $\delta_{\text{full}} : C^0_{\text{full}} \to C^1$ and $\delta_{\text{obs}} : C^0_{\text{obs}} \to C^1$ for the coboundaries as constructed in [2, §2], which share the edge-dimension cochain space C^1.

By the Column-Compression Identity [2, Theorem 2.1], $\delta_{\text{obs}} = \delta_{\text{full}}[:, \mathcal{O}]$ for a column index subset $\mathcal{O} \subseteq \{1, \ldots, n_{\text{full}}\}$. Write $H \subseteq \{1, \ldots, n_{\text{full}}\}$ for the complementary *hidden* column set, and $H_c := \delta_{\text{full}}[:, H]$ for the hidden-column block. Let $P_O : C^1 \to C^1$ be the orthogonal projector onto $\text{range}(\delta_{\text{obs}})$.

Definition 2.1 (Witness Gram)**.** The *witness Gram* of G is the symmetric positive-semidefinite $|H| \times |H|$ matrix

$$K(G) \;:=\; H_c^{\top}\,(I - P_O)\,H_c.$$

Entries of $K(G)$ are rational by construction: δ_{full} is a signed incidence matrix (one $+1$ and one -1 per row, zeros elsewhere) and is therefore totally unimodular, so the Moore-Penrose pseudoinverse of any column submatrix exists over $\mathbb{Q}$.

Theorem 2.2 (Rank identity)**.** *For every Bulla composition G,*

$$\operatorname{rank}\big(K(G)\big) \;=\; \operatorname{fee}(G).$$

Moreover, the bases of the column matroid of $K(G)$ coincide with the bases of the witness matroid $M/\mathcal{O}$ in [2, §2.3], and therefore with the minimum disclosure sets of [1, Theorem 6.2].

Proof. Write $W := (I - P_O)H_c$. Then $K(G) = W^{\top}W$, so $\operatorname{rank}(K) = \operatorname{rank}(W)$. Since $(I - P_O)$ is the orthogonal projector onto $\operatorname{range}(\delta_{\text{obs}})^{\perp}$ and H_c spans (together with δ_{obs}) the whole range of δ_{full}, we have $\operatorname{rank}(W) = \operatorname{rank}(\delta_{\text{full}}) - \operatorname{rank}(\delta_{\text{obs}}) = \operatorname{fee}(G)$ by the backbone identity. The basis equivalence follows because column-bases of W are exactly column-subsets of H_c whose images under $I - P_O$ are linearly independent — which is the matroid-contraction definition of bases in $M/\mathcal{O}$. □

Remark 2.3 (Tier 1 is regime-independent)*.* Theorem 2.2 makes no hypothesis on DFD or CHP. It is a structural identity between linear-algebraic objects. The regime hypotheses of §3 control not the *rank* of $K(G)$ but its *shape* — specifically, when it decomposes as a direct sum.

Remark 2.4 (Adjacent prior art on sheaf-Laplacian Hodge identities)*.* The rank identity $\operatorname{rank} K(G) = \dim H^1$ specialises in spirit a broader Hodge correspondence $\dim\ker\Delta_j = \dim H^j$ which has been developed for cellular sheaves valued in arbitrary abelian categories with inner-product stalks; see [9]. That treatment stays in the Hilbert-stalk regime and does not isolate the Schur-complement identity that Tier 2 of this paper establishes; the contribution here is the rational, signed-incidence, Kron-reduced realisation in $\mathsf{Vect}_{\mathbb{Q}}$ specifically.

3 Kron reduction under DFD

We recall DFD [2, Def. 2.7]: a composition G satisfies Dimension-Field Disjointness if no (t, f) pair is referenced by two distinct dimensions in D. Under DFD, the column partition $\{1, \ldots, n_{\text{full}}\} = \bigsqcup_{d \in D} V_d$ induces a block-diagonal structure on the full coboundary, with one block per dimension.

Definition 3.1 (Carrier graph)**.** For dimension $d \in D$, the *carrier graph* G_d is the multigraph with

- vertex set $V_d = \{(t, f) \;:\; f = d.\text{from or } f = d.\text{to on tool } t\}$,
- edge set indexed by rows of δ_{full} in dimension d, each row connecting the columns with nonzero entries $(-1, +1)$.

Let L_d be the combinatorial Laplacian of G_d. Under DFD each column of δ_{full} belongs to exactly one V_d, so the hidden/observable split refines as $V_d = H_d \sqcup O_d$ with $H_d = V_d \cap H$ and $O_d = V_d \cap \mathcal{O}$.

Definition 3.2 (Kron reduction)**.** The *Kron reduction* of L_d with respect to O_d is the Schur complement

$$\operatorname{Schur}(L_d \,/\, O_d) \;:=\; L_d[H_d, H_d] \;-\; L_d[H_d, O_d]\, L_d[O_d, O_d]^{-1}\, L_d[O_d, H_d],$$

whenever $L_d[O_d, O_d]$ is invertible. When $O_d = \varnothing$, the Kron reduction is defined to be L_d itself (no absorption). In degenerate cases — e.g. O_d vertices form a disconnected subgraph causing

$L_d[O_d, O_d]$ to be singular — we replace the inverse by the Moore-Penrose pseudoinverse; for connected carrier graphs with $|O_d| \geq 1$ and any non-grounded vertex outside O_d, the principal submatrix $L_d[O_d, O_d]$ is nonsingular [4, Prop. 3].

The name "Kron reduction" is not decorative: this is the matrix introduced by Gabriel Kron in 1939 for eliminating interior nodes from a power-system network while preserving the effective behavior at the boundary [3]. In our setting the "boundary" is the hidden carrier columns H_d and the "interior" is the observable carriers O_d, which are grounded by the projection $I - P_O$. The same matrix appears in spectral graph theory [4], random-walk analysis, and the theory of effective resistance [5].

Theorem 3.3 (Kron reduction of the witness Gram)**.** *If G satisfies DFD, then*

$$K(G) \;=\; \bigoplus_{d \in D} K_d, \qquad K_d \;=\; \mathrm{Schur}\big(L_d \,/\, O_d\big).$$

When additionally $O_d = \varnothing$ for every d (full CHP), the Kron reduction is trivial and $K_d = L_d$ itself (the pure carrier-graph Laplacian).

Proof. Block decomposition. Under DFD, the column partition $\{V_d\}_d$ induces a block-diagonal structure on δ_{full}: each column has nonzero entries only in rows of the single dimension that references its (t, f) pair (because row indexing splits by $(e, d.\text{name})$, and only the unique dimension containing (t, f) contributes). Hence $\delta_{\text{full}} \cong \bigoplus_d \delta_{\text{full},d}$, and the hidden/observable split respects the dimension blocks. The witness Gram inherits the block decomposition: $K(G) = \bigoplus_d K_d$, with cross-block entries zero.

Within-block identification. Each dimension-d row of δ_{full} has exactly two nonzero entries (-1 at source, $+1$ at destination), making $\delta_{\text{full},d}$ the signed incidence matrix of the carrier multigraph G_d. A classical identity (Godsil-Royle [6, Ch. 8]) gives

$$\delta_{\text{full},d}^{\top}\, \delta_{\text{full},d} \;=\; L_d,$$

the combinatorial Laplacian of the underlying unsigned multigraph.

Within-block Kron identification. Write $\delta_{\text{full},d}[H_d]$ and $\delta_{\text{full},d}[O_d]$ for the hidden and observable column blocks of $\delta_{\text{full},d}$. Because δ_{full} is block-diagonal under DFD, the range of δ_{obs} restricted to dimension-d rows is exactly $\mathrm{range}(\delta_{\text{full},d}[O_d])$ — observable columns from other dimensions produce zero entries on dimension-d rows and therefore cannot contribute to that row subspace. Consequently P_O restricted to dimension-d rows acts as the within-dimension observable projector onto $\mathrm{range}(\delta_{\text{full},d}[O_d])$, which by the Laplacian identity above has Gramian $L_d[O_d, O_d]$. The standard formula $P = A(A^{\top}A)^{-1}A^{\top}$ for the projector onto a column span gives

$$K_d \;=\; \delta_{\text{full},d}[H_d]^{\top}\,(I - P_O)\,\delta_{\text{full},d}[H_d] \;=\; L_d[H_d, H_d] \;-\; L_d[H_d, O_d]\, L_d[O_d, O_d]^{-1}\, L_d[O_d, H_d],$$

which is $\mathrm{Schur}(L_d \,/\, O_d)$. When $O_d = \varnothing$ the Schur term vanishes and $K_d = L_d[H_d, H_d] = L_d$. □

Definition 3.4 (Coboundary-incidence graph)**.** The *coboundary-incidence graph* G^* of a composition G has vertex set $\{(t, f) : f \in t.\text{int}\}$ (all internal column indices) and edge set the rows of δ_{full}, each row connecting the two columns with nonzero entries $(-1, +1)$. Under DFD, G^* splits as a disjoint union of carrier graphs, $G^* = \bigsqcup_{d \in D} G_d$.

Theorem 3.5 (Kron reduction without DFD)**.** *For every Bulla composition G (with or without DFD),*

$$K(G) \;=\; \mathrm{Schur}\big(L(G^*) \,/\, \mathcal{O}\big),$$

where $L(G^*)$ is the combinatorial Laplacian of the coboundary-incidence graph. Under DFD this reduces to the direct-sum form of Theorem 3.3; without DFD, G^* is a single multigraph spanning multiple dimensions and $K(G)$ is the Kron reduction of $L(G^*)$ as a whole.

Proof. The argument of Theorem 3.3 applied to the full coboundary — without the DFD-driven block decomposition — gives $K(G) = H_c^\top (I - P_O) H_c$, with H_c, P_O as before. Identifying $\delta_{\text{full}}^\top \delta_{\text{full}} = L(G^*)$ (each row a signed incidence edge in G^*, Godsil-Royle [6]) and applying the same projector formula $P_O = \delta_{\text{full}}[\mathcal{O}]\, L(G^*)[\mathcal{O},\mathcal{O}]^{-1}\, \delta_{\text{full}}[\mathcal{O}]^\top$ (with pseudoinverse in degenerate cases) yields the Schur-complement expression. Under DFD, $L(G^*)$ and thus the Schur complement are block-diagonal, recovering Theorem 3.3. □

Remark 3.6. Theorem 3.5 is the regime-independent form of the main theorem: the witness Gram is *always* a Kron-reduced graph Laplacian, while DFD controls whether the graph splits by dimension. Both regimes are demonstrated empirically and synthetically in §5.

Example 3.7 (Worked example on `financial_pipeline`)**.** The `financial_pipeline` composition has 3 tools and 4 hidden fields split between two dimensions (`day_convention` and `risk_metric`), each with carrier graph K_2 on 2 hidden vertices (no observable carriers; full CHP). The witness Gram is

$$K \;=\; L(K_2) \oplus L(K_2) \;=\; \begin{bmatrix} 1 & -1 & & \\ -1 & 1 & & \\ & & 1 & -1 \\ & & -1 & 1 \end{bmatrix},$$

with $\operatorname{rank}(K) = 2 = \text{fee}$, leverage $\frac{1}{2}$ on every hidden field, and basis count $\tau(K_2)\cdot\tau(K_2) = 1\cdot 1 = 1$ per dimension-pair: the minimum disclosure set is unique up to the dimension-level choice.

4 Four corollaries

All four corollaries fall out of Theorem 3.3 plus standard graph Laplacian theory. Write C for a connected component of the effective graph $\widetilde{G}_d$ whose Laplacian is K_d. (Under full CHP, $\widetilde{G}_d = G_d$.)

Corollary 4.1 (Leverage per component)**.** *Under DFD, the leverage score of a hidden field j in component C of $\widetilde{G}_d$ is*

$$\ell_j \;=\; \frac{|H_C| - 1}{|H_C|},$$

provided $|H_C| \geq 2$. Leverage is uniform within a component; leverage values may differ across components and across dimensions.

Proof. Kron reduction preserves connectivity of the effective graph: when L_d is the Laplacian of a connected carrier graph G_d, the Kron-reduced Laplacian $K_d = \operatorname{Schur}(L_d/O_d)$ is the Laplacian of a connected effective graph $\widetilde{G}_d$ on H_d [4, Theorem 3.1] — Kron reduction adds "virtual edges" between hidden vertices connected through observable paths, so the effective graph is at least as connected as the induced subgraph $G_d[H_d]$ (and strictly more connected when observable vertices were cut-vertices of G_d). Consequently K_C for each component C of $\widetilde{G}_d$ inherits null space $\operatorname{span}(\mathbf{1}_{H_C})$, so $\operatorname{rank}(K_C) = |H_C| - 1$. The orthogonal projector onto $\operatorname{col}(K_C) = \mathbf{1}_{H_C}^\perp$ is $P = I - \frac{1}{|H_C|}\mathbf{1}_{H_C}\mathbf{1}_{H_C}^\top$, whose diagonal is $(|H_C| - 1)/|H_C|$ at every coordinate in H_C. The leverage of field j is by definition this diagonal entry. □

Corollary 4.2 (No coloops under connected components)**.** *Under DFD, if every hidden component C of $\widetilde{G}_d$ has $|H_C| \geq 2$, then the witness matroid $M/\mathcal{O}$ has no coloops.*

Proof. A coloop is a field j appearing in every basis of $M/\mathcal{O}$, which corresponds to leverage $\ell_j = 1$. By Corollary 4.1, $\ell_j = (|H_C| - 1)/|H_C| < 1$ whenever $|H_C| \geq 2$. Singleton components ($|H_C| = 1$) correspond to hidden fields isolated in $\widetilde{G}_d$ and have leverage zero (loops), not one. □

Corollary 4.3 (Basis count via Matrix-Tree)**.** *Assume DFD and that each effective graph $\widetilde{G}_d$ is connected. Then the witness matroid has*

$$\big|\mathcal{B}(M/\mathcal{O})\big| \;=\; \prod_{d\in D} \tau(\widetilde{G}_d),$$

where $\tau(\widetilde{G}_d)$ is the number of spanning trees of $\widetilde{G}_d$. Under full CHP, $\widetilde{G}_d = G_d$ and the formula reduces to $\prod_d \tau(G_d)$.

Proof. Under DFD the witness matroid decomposes as a direct sum $M/\mathcal{O} = \bigoplus_d (M/\mathcal{O})_d$ with $(M/\mathcal{O})_d$ the graphic matroid of $\widetilde{G}_d$. By the Matrix-Tree Theorem applied to each connected $\widetilde{G}_d$, the bases of $(M/\mathcal{O})_d$ are the spanning trees of $\widetilde{G}_d$, counted by $\tau(\widetilde{G}_d)$ = any cofactor of K_d. Product across dimensions gives the stated formula. □

Remark 4.4. If some $\widetilde{G}_d$ is disconnected (not observed in the 703-corpus but possible in principle), the Matrix-Tree formula gives the number of *spanning forests* with one tree per component; the basis count of $(M/\mathcal{O})_d$ is then the forest count. See [6].

Concrete spanning-tree counts for graph families appearing in the 703-corpus: a single bilateral edge gives $\tau = 1$ (a unique minimum disclosure set); the complete graph K_n gives $\tau = n^{n-2}$ (Cayley's formula); the complete bipartite $K_{m,n}$ gives $\tau = m^{n-1}\, n^{m-1}$. For most 703-corpus dimensions the carrier graph has a single bilateral edge and the disclosure set is unique.

Corollary 4.5 (Effective resistance as disclosure substitutability)**.** *By Theorem 3.3, K_d is a graph Laplacian (specifically, the Laplacian of the effective graph $\widetilde{G}_d$), so the classical effective resistance of [5] is well-defined on hidden fields. For hidden fields i, j in the same component C of $\widetilde{G}_d$,*

$$R_{\text{eff}}(i,j) \;=\; K_d^{+}[i,i] + K_d^{+}[j,j] - 2\,K_d^{+}[i,j].$$

Effective resistance in the Kron-reduced witness geometry quantifies disclosure substitutability: the smaller the resistance between two hidden fields, the more equivalent they are as repair choices.

Concretely: small $R_{\text{eff}}(i,j)$ means disclosing either field substantially resolves the other's convention risk; large $R_{\text{eff}}(i,j)$ within a component means the fields are structurally distinct witness obligations; $R_{\text{eff}}(i,j) = \infty$ (across components of $\widetilde{G}_d$ or across dimensions under DFD) means the fields are structurally independent and no disclosure substitution exists. The classical graph interpretation [5] identifies R_{eff} with the voltage drop between i and j when unit current flows through the effective network; the Bulla interpretation reads that voltage drop as governance cost of disclosure substitution.

5 Regime breaks

The two regime hypotheses of Theorem 3.3 — DFD and CHP — can fail independently, and within any regime the effective-graph topology (bridges, cut vertices) carries additional structure. Table 1 organizes the picture. The 703-corpus already contains the most important partial-CHP failure naturally; we complete the picture with two synthetic constructions, one per uncovered regime.

Leverage is $(|H_C| - 1)/|H_C|$ per component C of the effective graph in all four regimes (Corollary 4.1); what changes is which graph the components live in.

Table 1: Regime taxonomy: witness Gram form under DFD × CHP.

DFD	CHP	Witness Gram form	Example
✓	full	$K = \bigoplus_d L_d$	203 corpus; `financial_pipeline`
✓	partial	$K = \bigoplus_d \mathrm{Schur}(L_d/O_d)$	37 corpus; `mcp-xmind`
×	full	$K = L(G^*)$	synthetic DFD violation
×	partial	$K = \mathrm{Schur}(L(G^*)/\mathcal{O})$	synthetic only
(sub-str.)		bridge; R_{eff} distinguishes	synthetic bridge topology

5.1 Partial CHP in the 703-corpus (natural regime break)

In the `airtable-mcp + mcp-xmind` composition, the dimension `path_convention_match` has 5 carrier columns: 3 hidden and 2 observable, split by whether each tool classifies its `path` field as a convention dimension or a primary input. The classifier disagreement across tools is structural, not accidental: `read_xmind` exposes `path` as a primary tool input (observable by design), while `extract_node` classifies its own `path` as a convention dimension (hidden). The same field name, different semantic roles.

The carrier graph is K_5 (complete on the 5 tool-field pairs, because every pair of mcp-xmind tools in the composition has a bilateral row). Absorbing the 2 observable vertices via Kron reduction gives

$$K_d \;=\; \mathrm{Schur}\big(L(K_5)\,/\,\{41,42\}\big) \;=\; \tfrac{5}{3}\cdot L(K_3) \;=\; \begin{bmatrix} 10/3 & -5/3 & -5/3 \\ -5/3 & 10/3 & -5/3 \\ -5/3 & -5/3 & 10/3 \end{bmatrix}.$$

Leverage is $\frac{2}{3}$ on all three hidden fields (per Corollary 4.1 with $|H_C| = 3$), matching the observed value exactly. All three pairwise effective resistances are $\frac{2}{5}$: the three fields are equally good disclosure substitutes.

Thirty-seven compositions in the 703-corpus exhibit this partial-CHP regime; every one of them involves mcp-xmind paired with another server. The pattern is a single-server property propagated across all pairings, not a combinatorial interaction between server pairs.

5.2 DFD violation (synthetic)

The synthetic composition `regime_break_dfd_violation.yaml` has two tools with all fields hidden (full CHP) and a single edge whose dimensions share a *from*-field: both `dim1` and `dim2` reference `tool_a.x`, violating DFD. The coboundary is

$$\delta_{\text{full}} \;=\; \begin{bmatrix} -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{bmatrix} \qquad \text{(rows: dim1, dim2, dim3; cols: x, y, u, v)},$$

and the witness Gram is

$$K(G) \;=\; \begin{bmatrix} 2 & 0 & -1 & -1 \\ 0 & 1 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}.$$

$K(G)$ is *not* block-diagonal: the column `x` has off-diagonal entries -1 to both `u` and `v` (two dimensions' worth). But $K(G)$ *is* still a graph Laplacian — specifically, $K(G) = L(G^*)$ where G^* is the coboundary-incidence graph on 4 vertices with edges $\{(x,u), (x,v), (y,u)\}$ (the three dimensions' bilateral pairs). The leverage is $\frac{3}{4}$ uniformly across all four hidden fields (because G^* is connected with $|H_C| = 4$), and the fee is 3.

This is the non-DFD case of Theorem 3.5: block decomposition fails but the Laplacian structure persists on the merged graph.

5.3 Bridge topology (synthetic)

The synthetic composition `regime_break_bridge_topology.yaml` has three tools $\{A, B, C\}$ each carrying `path_convention_match` via their `path` field, with bilateral edges $A \to B$ and $B \to C$ (no $A \to C$ edge). Full CHP and DFD hold. The carrier graph is the path P_3 on three vertices; the witness Gram is

$$K(G) \;=\; L(P_3) \;=\; \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Leverage is $\frac{2}{3}$ uniformly (Corollary 4.1 with $|H_C| = 3$), but effective resistances distinguish the bridge: $R_{\text{eff}}(A, B) = R_{\text{eff}}(B, C) = 1$ while $R_{\text{eff}}(A, C) = 2$. The vertex B has lower average effective resistance ($\bar{R} = 1$) than A or C ($\bar{R} = 1.5$); disclosing B substitutes for either of the other two, while disclosing A barely substitutes for C.

This example demonstrates that effective resistance (Corollary 4.5) carries structural information that leverage alone cannot: the three fields have identical leverage but qualitatively different roles in the witness geometry.

6 Corpus confirmation

All claims below are verified in exact rational arithmetic on the 703-composition real-schema MCP registry corpus of [2, §1.1]. Source: `bulla/calibration/witness_geometry_sweep.py` and `verify_laplacian_collapse.py`.

Table 2: Structural identities verified (regime-independent).

Identity	**Pass**	**Fail**
$\text{rank}(K(G)) = \text{fee}(G)$	703/703	0
$\sum_j \ell_j = \text{fee}(G)$ (exact $\mathbb{Q}$)	703/703	0
$0 \le \ell_j \le 1$ for all hidden fields	703/703	0

Table 3: Leverage-prediction match via Kron formula (nontrivial compositions only).

Prediction match	**Count**
ℓ_j matches $(\lvert H_C\rvert - 1)/\lvert H_C\rvert$ per component (exact $\mathbb{Q}$)	240/240
Mismatches	0

Table 4: Kron three-case taxonomy on the 703-corpus.

Case	**Description**	**Count**	**Leverage formula**
1	Full CHP: $O_d = \varnothing$; $K_d = L_d$	203	$(\lvert V_d\rvert - 1)/\lvert V_d\rvert$
2	Trivial Kron: $O_d \neq \varnothing$ but $L_d[O_d, O_d] = 0$	0	$(\lvert H_d\rvert - 1)/\lvert H_d\rvert$
3	Genuine Kron reduction: $L_d[O_d, O_d] \neq 0$	37	$(\lvert H_d\rvert - 1)/\lvert H_d\rvert$
	Trivial ($\lvert H\rvert = 0$)	300	n/a
	Fee zero ($\lvert H\rvert > 0$, fee $= 0$)	163	n/a

Three corpus-wide observations deserve comment.

Case 2 is empty. On the 703-corpus every partial-CHP composition exercises the full Kron reduction (nonzero inter-observable edges). There is no degenerate middle case: either full CHP

(Case 1, 203 compositions) or genuine Kron absorption (Case 3, 37 compositions). This gives the corpus a clean shape and simplifies §3's narrative.

Intra-dimension leverage uniformity. No composition in the corpus exhibits non-uniform leverage within a single component. The theorem predicts this (Corollary 4.1: uniform within a component); the empirical confirmation rules out cases where disconnected hidden-carrier components might split a dimension into sub-components with distinct leverage values.

Case 3 concentrates on a single server. All 37 partial-CHP compositions involve `mcp-xmind`. The classifier's disagreement on `path` fields across `mcp-xmind`'s tools propagates across every pairing with another server, producing 37 Case-3 compositions from a single structural source. The frontier where field-level diagnostics become discriminative is thus already present in the corpus, but it is a single-server phenomenon — not a combinatorial artifact of server pairs.

7 Discussion

The witness Gram gives the Bulla coherence fee a canonical form that the integer alone cannot: a Kron-reduced graph Laplacian whose spectrum, bases, and effective-resistance structure encode the composition's hidden convention geometry. Four operational consequences:

1. **Closed-form leverage.** Within each connected component of the Kron-reduced effective graph, leverage is $(|H_C| - 1)/|H_C|$ — a function of component size alone. Bulla's CLI can report leverage by tabulating component sizes, no matrix operations required.
2. **Spanning-tree-many bases.** The number of minimum disclosure sets is $\prod_d \tau(\widetilde{G}_d)$. A product of spanning-tree counts. Most 703-corpus compositions have $\tau = 1$ per dimension (unique disclosure set); mcp-xmind compositions have $\tau > 1$ (multiple valid substitutes).
3. **Weighted repair is matroid optimization.** The minimum-cost full repair is a minimum-weight basis of the witness matroid, solved by the classical greedy algorithm [7]. Not a heuristic.
4. **Effective resistance ranks substitutes.** Within a component, the effective resistance between two hidden fields measures how much one disclosure substitutes for the other. Disclosure planning on real compositions can use this to rank alternative repair choices when privacy, cost, or operational constraints rule out the greedy-optimal field.

The regime-dependent vs regime-independent split. Three of the witness Gram's invariants — rank, leverage, and basis count — are closed-form functions of the effective-graph topology alone. On the 703-corpus this means that under the current Bulla classifier, they are computable by hand from carrier-count vectors without any matrix work. The fourth invariant — effective resistance (Corollary 4.5) — is where the geometry earns its keep: it distinguishes bridge structures that leverage cannot (§5.3), ranks disclosure substitutes when multiple minimum bases exist, and exposes the local-vs-distributed shape of hidden convention risk.

Research frontier. The foundational theorem — that hidden convention debt is a rank, and that rank has a canonical electrical-network interpretation — is now complete. The natural open questions shift from "what is the witness geometry?" to two more pointed ones. First: what happens when future MCP server designs, or future Bulla classifier revisions, partially expose convention fields? Theorem 3.5 covers all such cases in closed form, but the empirical phase diagram is open — which regime cells (DFD × CHP) do evolving MCP ecosystems populate? Second: does the Kron-reduction framing open formal bridges to the electrical-network, random-walk, and resistance-distance literatures whose results (e.g., network simplification, Markov-chain commute times) could specialize to Bulla with non-trivial consequences? Both are outside the scope of this paper but within the scope of the program.

A Lean verification status

The two regime-independent identities of this paper — the rank identity (Theorem 2.2) and leverage conservation (the foundation of Corollary 4.1's sum) — are formally verified in Lean 4 against Mathlib (Aristotle/Harmonic project UUID `3c1a38f9-a823-4b80-ae5d-7e4dfaacad85`). Two theorems closed, zero `sorry`, standard axioms only:

Theorem 1. `witness_gram_rank_eq_fee` (Theorem 2.2, rank identity). Key Mathlib: `finrank_quotient_ad`

Theorem 2. `leverage_conservation` (Corollary 4.1, basis of $\Sigma\,\ell_j$ = fee). Key Mathlib: `trace_comp_comm'`, `trace_id`.

Theorem 1 establishes the rank-difference identity in finrank-quotient form: under the column-compression hypothesis $\delta_{\text{obs}} = \delta_{\text{full}} \circ \iota$, the witness-Gram rank equals $\text{finrank}(\text{range}\,\delta_{\text{full}}) - \text{finrank}(\text{range}\,\delta_{\text{obs}})$. Theorem 2 establishes that for any idempotent endomorphism P of a finite-dimensional $\mathbb{Q}$-vector space, $\text{trace}\,P = \text{finrank}(\text{range}\,P)$, which directly gives leverage conservation when $P = K^{+}K$ is the projection onto $\text{col}(K)$.

Combined with the twelve Lean-verified theorems of the backbone paper [2] (UUIDs `044a00b1`, `ff78c99c`, `edb6c604`, `01333bb4`), the program now has fourteen Mathlib-verified theorems with zero `sorry`. The remaining paper-specific claims — the Kron-reduction theorem (Theorem 3.3) and the closed-form leverage formula $\ell_j = (|H_C| - 1)/|H_C|$ (Corollary 4.1) — are deferred to follow-on Lean work, since each requires formalizing Kron-reduction preservation of connectivity (Dörfler-Bullo Theorem 3.1) and the signed-incidence Laplacian identity, neither currently in Mathlib.

B Code and data artifacts

All code, data, and verification scripts are available at `https://github.com/jkomkov/bulla`. Canonical artifacts:

- `bulla/src/bulla/witness_geometry.py` (module)
- `bulla/calibration/witness_geometry_sweep.py` (703-corpus sweep)
- `bulla/calibration/verify_laplacian_collapse.py` (240/240 leverage verification)
- `bulla/calibration/kron_three_cases.py` (Case 1/2/3 classification)
- `bulla/compositions/regime_break_*.yaml` (synthetic regime-break examples)
- `papers/hierarchical-fee/laplacian_collapse.md` (extended proof memo)
- `papers/hierarchical-fee/THEOREM-LOCK.md` (vocabulary and statement lock)

References

[1] J. Komkov, "The Coherence Fee: Measuring Hidden Convention Risk in Hierarchical Tool Compositions," 2026.

[2] J. Komkov, "The Column-Matroid Backbone of a Composition-Coherence Fee: a Structural Identity and Its Corollaries," 2026.

[3] G. Kron, "Tensor Analysis of Networks," John Wiley & Sons, 1939.

[4] F. Dörfler and F. Bullo, "Kron Reduction of Graphs With Applications to Electrical Networks," *IEEE Trans. Circuits and Systems I*, vol. 60, no. 1, pp. 150-163, 2013.

[5] D. J. Klein and M. Randić, "Resistance Distance," *J. Math. Chem.*, vol. 12, pp. 81-95, 1993.

[6] C. Godsil and G. Royle, "Algebraic Graph Theory," Springer, 2001.

[7] J. Edmonds, "Matroids and the greedy algorithm," *Math. Programming*, vol. 1, pp. 127-136, 1971.

[8] J. Komkov, "Signed-Incidence Structure in Compositional Verification: A Structural Note," `papers/signed-incidence/paper/signed-incidence.pdf`, 2026.

[9] A. Ayzenberg, T. Gebhart, G. Magai, and Y. Solomadin, "Sheaf theory: from deep geometry to deep learning," arXiv:2502.15476, 2025.

Signed-Incidence Structure in Compositional Verification

A Structural Note

John Komkov
Independent Researcher
https://github.com/jkomkov/bulla

April 2026

Abstract

This is a short structural note, not a standalone program statement. Its job is to supply one disciplined source for three facts used repeatedly in the Bulla corpus: (i) each row of the coboundary matrix $\delta(G)$ has at most one $+1$ and at most one -1; (ii) the coherence fee $\mathrm{fee}(G)$ is therefore field-independent as a difference of two totally-unimodular ranks; and (iii) pairwise endpoint coupling is bounded by rank 2 on every (edge, dimension) block. A negative corollary records that any partition-style typed repair constraint is structurally wrong: enforcing it on the 703-composition real-schema corpus makes fee-zero repair impossible in 151 of 240 nonzero-fee compositions. The signed-incidence $\to$ total-unimodularity step is Lean-verified.[1] The greedy untyped repair algorithm remains correct and optimal for its stated problem.

Role. Five papers in the Bulla corpus—[2], [3], [4], [5], and [6]—refer to a "companion signed-incidence note" supplying their middle-layer structural background. This is that note. It does not introduce new claims; it freezes the exact statements those papers cite.

1 Structural proposition

Proposition 1.1 (Signed-incidence row structure)**.** *For every composition G constructible by* `build_coboundary`*, each row of the coboundary matrix $\delta(G)$ has at most one entry equal to $+1$ and at most one entry equal to -1, with all other entries zero.*

Proof. A row is indexed by one (edge, dimension) pair. The construction can write at most one source term at `dim.from_field` and at most one target term at `dim.to_field`. No other writes touch that row. □

Status. Proved by construction, runtime-asserted, and empirically verified on the full 1596-composition corpus.

2 Field-independence corollary

Corollary 2.1 (Field-independence of the fee)**.** *For every composition G in the Bulla coboundary family,*

$$\mathrm{fee}_{\mathbb{Q}}(G) \;=\; \mathrm{fee}_{F}(G)$$

[1] Aristotle run `fecfaac3-fed7-40dd-82aa-5eef8edd7111`; theorem `signed_incidence_det_in_unit`, 100 lines, 8 helper lemmas, no `sorry`, standard axioms only. Submission stub: `papers/sheaf/lean/SHEAF/SignedIncidence.submission.lean`; proof outline: `SignedIncidence.ARISTOTLE_SUMMARY.md`.

for every field F.

Proof. (1) By proposition 1.1, both $\delta_{\text{full}}(G)$ and $\delta_{\text{obs}}(G)$ are signed incidence matrices.

(2) Signed incidence matrices are totally unimodular [1].

(3) Totally unimodular matrices have field-independent rank.

(4) $\text{fee}(G) = \text{rank}(\delta_{\text{full}}(G)) - \text{rank}(\delta_{\text{obs}}(G))$, so the fee is field-independent as a difference of two field-independent ranks.

□

Status. Proved. The TU step is Lean-verified in the current Cycle 18 chain.

Remark 2.2 (Boundary). This proves numerical agreement across coefficient fields. It does *not* prove that the $\mathbb{F}_2$ holonomy interpretation and the $\mathbb{Q}$-valued Bulla implementation are the same mathematical object. The disciplined claim is weaker and sufficient: they compute the same numerical invariant from structurally corresponding presentations.

3 Endpoint-coupling corollary

Corollary 3.1 (Pairwise endpoint-coupling bound)**.** *Let G be a pairwise composition, and let B be the hidden-column block associated to one fixed (edge, dimension) pair in the witness Gram $K(G) = H^\top (I - P_O)\, H$. Then*

$$\text{rank}\big(K[B, B]\big) \;\leq\; 2.$$

Proof. In Bulla's pairwise setting, a semantic dimension on a fixed edge can contribute at most one hidden source column and at most one hidden target column, because `SemanticDimension` carries only one `from_field` and one `to_field`. Therefore $|B| \leq 2$. Since $K[B, B]$ is a $|B| \times |B|$ matrix, its rank is at most $|B|$, hence at most 2. □

Status. Proved.

Remark 3.2 (Why the projection issue does not block the theorem). An earlier proof sketch reasoned through independence in δ and then worried about the projection $(I - P_O)$. That was unnecessary. The theorem only needs the cardinality bound on the block definition itself; projection can reduce rank, but it cannot increase rank above the number of columns in the block.

Empirical sharpening on the 703 real-schema corpus.

- 3873 multi-column (edge, dimension) blocks were tested.
- 3703 had rank 2.
- 170 had rank 1.
- 0 blocks exhibited the feared pattern "rank 2 in the corresponding δ block but rank < 2 after projection into K."

The theorem is therefore not only true but active in practice: rank-2 endpoint coupling is common, not merely possible.

4 Negative corollary for typed repair

Corollary 4.1 (Partition-style separability is structurally wrong)**.** *Consider the proposed typed-repair constraint "disclose at most one field from each* (*edge*, *dimension*) *block." On the* 703*-composition real-schema corpus, restricting repair to that constraint makes fee-zero repair impossible in* 151 *of the* 240 *nonzero-fee compositions (*62.9%*).*

Interpretation. The failure is not that the partition-constrained solution is sometimes more expensive. The stronger failure is that it often cannot reach zero fee at all. When both endpoints of an edge contribute hidden fields in the same dimension, those two disclosures can carry independent witness information. Forbidding their joint disclosure destroys valid repair paths.
Status. Empirical finding, locked by the corpus experiment.

5 Algorithmic consequence

Consequence 5.1 (The untyped greedy is already optimal for its stated problem)**.** *`weighted_greedy_repair` computes a minimum-cost basis of the witness matroid M/O. That remains the correct algorithm for the untyped repair problem it actually claims to solve.*

The falsified step was not the greedy algorithm. The falsified step was the proposed separable typed extension that tried to model same-dimension constraints as a partition matroid. The structural lesson is:

- untyped greedy remains correct,
- partition-style typed repair is not the right abstraction,
- the open problem is the correct coupled-disclosure objective, not a replacement for the current untyped basis computation.

6 Evidence summary

- Signed-incidence row structure: code-audited, runtime-asserted, 1596/1596 verified.
- Field-independence: structural proof plus Lean-backed TU chain.
- Pairwise endpoint coupling: theorem by block cardinality, with 3873/3873 multi-column witness blocks satisfying rank ≤ 2 on the 703 real-schema corpus.
- Projection sanity check for corollary 3.1: zero cases where a corresponding δ block had rank 2 and the K block dropped below 2.
- Partition constraint failure: 151/240 nonzero-fee compositions cannot be fully repaired under the one-per-block restriction.

7 Claim boundaries

This note intentionally does *not* claim any of the following:

- a characterization of the correct dimension-aware repair objective,
- generalized k-way endpoint-coupling bounds for non-pairwise compositions,
- that the $\mathbb{F}_2$ and $\mathbb{Q}$ stories are literally one object rather than numerically aligned invariants,

- any commercial, philosophical, or programmatic interpretation beyond the structural statements above.

Those belong elsewhere. The purpose of this note is narrower: to let the rest of the corpus cite one exact middle layer instead of re-deriving it ad hoc.

References

[1] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley, 1986. Theorem 19.3: a $\{0, \pm 1\}$ matrix is totally unimodular if and only if for every subset R of its rows there is a bipartition $R = R_1 \cup R_2$ such that for every column j, $|\sum_{i \in R_1} a_{ij} - \sum_{i \in R_2} a_{ij}| \leq 1$. For signed incidence matrices, the trivial partition $R_1 = R$, $R_2 = \emptyset$ satisfies this immediately.

[2] J. Komkov, "The Column-Matroid Backbone of a Composition-Coherence Fee: A Structural Identity and Its Corollaries," `papers/hierarchical-fee/paper/column-matroid-backbone.tex`, 2026.

[3] J. Komkov, "The Witness Gram: Kron-Reduced Laplacian of Hidden Convention Risk," `papers/hierarchical-fee/paper/witness-gram.tex`, 2026.

[4] J. Komkov, "A Hierarchical Coherence Fee for Compositional Verification," `papers/hierarchical-fee/paper/hierarchical-fee.tex`, 2026.

[5] J. Komkov, "Bridge: Empirical Witness for Compositional Failure," `papers/bridge/paper/bridge.tex`, 2026.

[6] J. Komkov, "The Coherence Cliff," `papers/coherence-cliff/paper/the-coherence-cliff.tex`, 2026.

Witness Geometry Beyond Scalar Fee

Repair Entropy, Operational Sparsity,
and the Decision Theory of Disclosure

John Komkov
Independent Researcher

April 2026

Abstract

The coherence fee of a tool composition is a single integer measuring hidden convention obligation. Working in the linear disclosure model (repair = field exposure) under dimension-field disjointness, we show that the fee is the first of three increasingly operational projections of the witness Gram $K(G)$. The per-dimension witness matroid is uniform—$U(n{-}1, n)$ on each connected component—correcting Paper III Corollary 4.3. We define *repair entropy* $H_{\text{repair}} = \log \beta$, where $\beta = \prod_{d,j} |C_{d,j}|$ is the product of witness-component sizes, and prove sharp bounds: $\varphi{+}1 \leq \beta \leq 2^{\varphi}$, both attained. But repair entropy measures *structural* flexibility—the size of the exact repair space. Under cost perturbation, most of that space is operationally irrelevant. We prove that every basis is the unique optimum under some cost vector (full realizability), so operational sparsity is not intrinsic to the matroid but depends on the cost family. On a 240-composition MCP corpus (239 testable for basis enumeration at fee ≤ 14): (i) β(formula) = β(enumerated) on all 239; (ii) 81.7% of compositions sit at the rigid bound $\beta = \varphi{+}1$; (iii) under integer costs in $[1, 10]$, at $\beta = 378$ only 13.5% of bases are ever optimal. Fee counts the obligations. Repair entropy counts the structural repair space. Operational sparsity under a given cost family identifies the decision-relevant subset that survives deployment constraints. In this regime, the exact minimum repair cost is total active hidden cost minus the sum of per-component cost maxima (the *geometry dividend*). Decision activation—the condition under which witness geometry changes the optimal repair—requires both structural repair multiplicity and cost heterogeneity on the witness support; removing either ingredient deactivates the geometry.

1 Introduction

The coherence fee $\text{fee}(G) = \text{rank}(\delta_{\text{full}}) - \text{rank}(\delta_{\text{obs}})$ measures how many hidden convention obligations a tool composition carries. Papers I–III of this program established the fee as a matroid corank (Paper II), showed its rank lives in the witness Gram $K(G) = H_c^{\top}(I - P_O)H_c$ (Paper III), and proved that no bounded-local audit can compute it (the obstruction paper).

But the fee is a single integer. Two compositions with fee = 11 may have very different repair landscapes: one may have a unique minimum disclosure set, the other hundreds of equally valid alternatives; one may require disclosing expensive fields, the other only cheap ones. The scalar fee cannot distinguish these cases.

This paper asks: **what invariant, beyond fee, determines the structure of repair?**

The answer is the witness Gram $K(G)$ itself—and the central new invariant it yields is **repair entropy**.

We prove four results:

1. **K-sufficiency (Theorem 1).** In the linear disclosure model (repair = field exposure), $K(G)$ determines the witness matroid M/O and therefore the exact repair basis structure:

basis count, minimum-cost repair under any cost vector, and cost sensitivity. The fee is the rank of K; everything else lives in its finer structure.

2. **Uniform witness structure and repair entropy (Theorem 3).** Under DFD + CHP, the per-component witness matroid is $U(n-1, n)$—the uniform matroid, not the graphic matroid. The corrected basis count is $\beta(G) = \prod_{d,j} |C_{d,j}|$, and the repair entropy $H_{\text{repair}} = \sum \log |C_{d,j}|$ is additive over witness components. Fixed fee does not determine repair entropy.

3. **Sharp bounds at fixed fee (Theorem 8).** For any composition with fee φ in the DFD + CHP regime: $\varphi+1 \leq \beta(G) \leq 2^{\varphi}$, with both bounds attained.

4. **Full realizability (Theorem 10).** Every basis of M/O is the unique minimum-cost repair under some cost vector. Operational sparsity is a property of the cost family, not the matroid.

5. **Exact repair functional (Theorem 12).** In the DFD + CHP regime, the minimum repair cost is total active hidden cost minus the sum of per-component cost maxima. The *geometry dividend* $A(G, c)$ is the operational value of witness geometry.

6. **Decision activation (Proposition 19).** Witness geometry changes the optimal repair if and only if the composition has structural repair multiplicity *and* the cost vector is heterogeneous on the witness support. Neither condition alone suffices. This follows algebraically from Theorem 12.

The paper also corrects an error in Paper III Corollary 4.3, which stated $|B(M/O)| = \prod \tau(\tilde{G}_d)$ (spanning-tree counts). The correct formula is $\beta = \prod |C_{d,j}|$ (component-size products). See §4 and the correction memo for the full impact analysis.

Notation. See the companion NOTATION.md for the canonical symbol table. We follow Papers II and III throughout.

2 K-Sufficiency: The Witness Gram Determines Repair Structure

2.1 The characterization theorem

Theorem 1 (K-sufficiency)**.** *Let G be a Bulla composition with linear disclosure model (field exposure as the repair action). The witness Gram $K(G) = H_c^\top (I - P_O) H_c$ determines:*

(a) The witness matroid M/O: a subset $S \subseteq H$ is independent in M/O if and only if the columns of K indexed by S are linearly independent.

(b) All bases $B(M/O)$: the minimum disclosure sets of G are exactly the column-bases of $K(G)$.

(c) The repair multiplicity $\beta(G) = |B(M/O)|$.

(d) For any cost vector $c : H \to \mathbb{Q}_{>0}$, the minimum-cost repair basis $B^*(G, c)$ and minimum repair cost $\Sigma^*(G, c)$.

(e) The leverage scores $\ell_j = (K^+ K)_{jj}$ and effective resistances $R_{\text{eff}}(i, j)$.

Proof. (a) By Paper III Theorem 2.1, the column matroid of $K(G) = W^\top W$ coincides with the column matroid of $W = (I - P_O) H_c$, which is the contraction M/O. Hence $S \subseteq H$ is independent in M/O iff the S-columns of K are linearly independent.

(b) Immediate from (a): bases of the column matroid of K are bases of M/O.

(c) Follows from (b): $\beta(G) = |B(M/O)|$ is determined by K.

(d) Given $B(M/O)$ from (b) and a cost vector c, the minimum-cost basis is $B^*(G, c) = \arg\min_{B \in B(M/O)} \sum_{h \in B} c(h)$. This is a function of $B(M/O)$ and c, both determined by K and c.

(e) Standard: leverage and effective resistance are defined in terms of K and K^+. □

Remark 2 (What K-sufficiency does and does not say). $K(G)$ is sufficient for repair combinatorics—it determines what can be repaired, at what cost, and with what alternatives. It is *not* sufficient for all properties of G: it does not determine the composition graph, the tool schemas, or the observable structure (these are inputs to K, not outputs). The sufficiency is for one specific question: given that a composition has hidden obligations, what is the structure of resolving them?

2.2 K versus fee: a hierarchy of invariants

The witness Gram $K(G)$ sits at the top of an invariant hierarchy, each level a lossy compression of the one above:

$$\underbrace{K(G)}_{\text{full Gram}} \xrightarrow{\text{matroid}} \underbrace{M/O}_{\text{independence oracle}} \xrightarrow{\text{bases}} \underbrace{B(M/O)}_{\text{basis set}} \xrightarrow{\text{count}} \underbrace{\beta(G)}_{\text{multiplicity}} \xrightarrow{\text{rank}} \underbrace{\text{fee}(G)}_{\text{integer}}$$

Each arrow discards information. Fee discards all basis structure. β discards which bases exist. M/O discards the metric structure (effective resistance). K retains everything.

3 Fee Does Not Determine Repair Multiplicity

3.1 The separation theorem

Theorem 3 (Fee–multiplicity separation)**.** *For every $\varphi \geq 2$, there exist Bulla compositions G_0 and G_1 with:*

1. $\text{fee}(G_0) = \text{fee}(G_1) = \varphi$
2. $\beta(G_0) \neq \beta(G_1)$

Proof. We work in the DFD + full CHP regime (all convention fields hidden, each field in exactly one dimension), so that $K(G) = \bigoplus_d L_d$.

The correct basis-count formula. Under DFD + full CHP with each carrier graph G_d connected on n_d hidden vertices, the column matroid of the signed incidence matrix δ_d is the uniform matroid $U(n_d - 1, n_d)$: every $(n_d - 1)$-subset of columns is a basis. (This is because each column of δ_d is a signed characteristic vector of a vertex, and any $n_d - 1$ vertices of a connected graph are linearly independent in the signed incidence matrix.) Under the direct-sum decomposition, the total basis count is:

$$\beta(G) = \prod_d n_d$$

Each connected carrier graph G_d contributes $\text{fee}_d = n_d - 1$ and n_d bases to the product. The fee constrains the sum $\sum(n_d - 1) = \varphi$, but the product $\prod n_d = \prod(\text{fee}_d + 1)$ depends on the partition of φ across dimensions—not just the total.

Construction.

G_0*:* φ dimensions, each with carrier graph K_2 (2 hidden fields, one bilateral edge). Partition: $\text{fee}_d = 1$ for all d.

- $\text{fee}(G_0) = \varphi \cdot 1 = \varphi$
- $\beta(G_0) = 2^\varphi$

G_1*:* $(\varphi - 2)$ dimensions with carrier graph K_2, plus one dimension with carrier graph C_3 (3 hidden fields in a triangular routing). Partition: $\text{fee} = 1 \cdot (\varphi - 2) + 2 = \varphi$.

- $\text{fee}(G_1) = (\varphi - 2) \cdot 1 + (3 - 1) = \varphi$
- $\beta(G_1) = 2^{\varphi-2} \cdot 3$

Verification. $\beta(G_0)/\beta(G_1) = 2^\varphi/(2^{\varphi-2} \cdot 3) = 4/3 \neq 1$, so $\beta(G_0) \neq \beta(G_1)$ for all $\varphi \geq 2$. Computationally verified for $\varphi \in \{2, 3, 5, 10\}$ by brute-force basis enumeration:

φ	$\beta(G_0)$	$\beta(G_1)$	Ratio
2	4	3	4/3
3	8	6	4/3
5	32	24	4/3
10	1024	768	4/3

□

Remark 4 (Paper III Corollary 4.3 correction). Paper III states $|B(M/O)| = \prod_d \tau(\tilde{G}_d)$ (product of spanning-tree counts). This conflates the graphic matroid (ground set = edges, bases = spanning trees) with the column matroid of the signed incidence matrix (ground set = vertices, bases = $(n-1)$-subsets of vertices). The correct formula in the DFD + full CHP regime with connected carrier graphs is $\beta = \prod_d n_d$. The two coincide only when $n_d = \tau(G_d)$ for all d—for example, when every carrier graph is a cycle (where $\tau(C_n) = n$). For general connected graphs, $n \neq \tau$.

Corollary 5 (Unbounded multiplicity gap)**.** *For every $\varphi \geq 2$, the ratio $\beta_{\max}/\beta_{\min}$ among fee-φ compositions is unbounded as the number of available tools grows.*

Proof. The partition $[\varphi]$ (single dimension with $n = \varphi + 1$ hidden fields) gives $\beta = \varphi + 1$. The partition $[1, 1, \ldots, 1]$ (φ K_2-dimensions) gives $\beta = 2^\varphi$. Since $2^\varphi/(\varphi + 1) \to \infty$, the ratio is unbounded. □

Corollary 6 (Multiplicity is partition-determined)**.** *Under DFD + full CHP with connected carrier graphs, $\beta(G) = \prod_d(\text{fee}_d + 1)$ is determined entirely by the fee partition $(\text{fee}_1, \ldots, \text{fee}_D)$—the distribution of the scalar fee across convention dimensions.*

Two compositions with the same fee and the same fee partition have the same β, but compositions with the same fee and different partitions generically have different β. The fee partition is a strictly finer invariant than the scalar fee.

3.2 Repair entropy

Definition 7. *The* repair entropy *of a composition G is $H_{\text{repair}}(G) = \log \beta(G) = \sum_{d,j} \log |C_{d,j}|$.*

Invariance class. H_{repair} is a matroid invariant of M/O: it depends only on the witness matroid, not on the choice of coboundary representation. Under DFD + CHP it decomposes additively over carrier graph components. It is invariant under tool renaming, field renaming, and edge reorientation.

Regime of validity. The product formula $\beta = \prod |C_{d,j}|$ and the additive decomposition $H_{\text{repair}} = \sum \log |C_{d,j}|$ require DFD (dimension-field disjointness) and CHP (convention-hidden partition). Without DFD, the witness Gram $K(G)$ is a single Kron-reduced Laplacian and the

basis count does not factor by dimension. The definition $H_{\text{repair}} = \log |B(M/O)|$ remains valid in general, but the closed-form product formula does not.

What repair entropy does not measure. H_{repair} counts minimum-size disclosure sets (bases of M/O). It does not measure:

- the *cost* of repair (which requires a cost model $c : H \to \mathbb{Q}$),
- the *stability* of the optimal repair under cost perturbation (which requires the cost-sensitivity geometry),
- or the *difficulty* of implementing a repair (which is external to the matroid).

H_{repair} says how many interchangeable minimum disclosure configurations exist—not how hard, how expensive, or how robust any particular repair is.

The fee–entropy pair. Fee and repair entropy are two projections of the witness Gram:

- **fee** = obligation count (additive: $\sum(n_{d,j} - 1)$)
- H_{repair} = repair flexibility (additive: $\sum \log n_{d,j}$)

Fee says how many obligations exist. Repair entropy says how many interchangeable ways those obligations can be satisfied.

3.3 Sharp bounds at fixed fee

Theorem 8 (Sharp repair-entropy bounds)**.** *For any composition with fee $\varphi \geq 1$:*

$$\varphi + 1 \leq \beta(G) \leq 2^{\varphi}$$

equivalently:

$$\log(\varphi + 1) \leq H_{\text{repair}}(G) \leq \varphi \cdot \log 2$$

Both bounds are attained:

- ***Minimum*** *$\beta = \varphi + 1$: all fee in one connected witness component of size $\varphi + 1$ (rigid repair—few alternatives).*
- ***Maximum*** *$\beta = 2^{\varphi}$: fee distributed across φ independent bilateral pairs, each of size 2 (flexible repair—exponentially many alternatives).*

Proof. The repair multiplicity $\beta = \prod n_i$ where $n_i = |C_{d,j}| \geq 2$ and $\sum(n_i - 1) = \varphi$. Set $m_i = n_i - 1 \geq 1$; then $\sum m_i = \varphi$ and $\beta = \prod(m_i + 1)$.

Upper bound. If any $m_i \geq 2$, split it: replace $m_i = a$ with two parts $m_i = 1$ and $m_{\text{new}} = a - 1$. The product changes by factor $2a/(a+1) > 1$ for $a \geq 2$. Repeat until all $m_i = 1$, giving $\beta = 2^{\varphi}$.

Lower bound. If there are two parts $m_i = a$ and $m_j = b$, merge them into $m_i = a + b$. The product changes by factor $(a + b + 1)/((a + 1)(b + 1)) < 1$ since $ab > 0$. Repeat until one part remains, giving $\beta = \varphi + 1$.

Attainment. All-K_2 compositions (φ bilateral pairs, each with 2 hidden fields) give $\beta = 2^{\varphi}$. A single carrier component on $\varphi + 1$ hidden fields gives $\beta = \varphi + 1$. □

Computational verification: exhaustive enumeration of all integer partitions of φ confirms the bounds for $\varphi \in \{1, 2, 3, 4, 5, 8, 10\}$:

φ	$\beta_{\min}$	$\beta_{\max}$	Distinct β values
1	2	2	1
2	3	4	2
3	4	8	3
5	6	32	7
10	11	1024	39

At fee = 10, there are 39 distinct values of β—39 structurally distinguishable repair landscapes, all invisible to the scalar fee.

3.4 What Theorems 3 and 8 mean together

Two compositions with fee = 11 have:

- $\beta \in [12, 2048]$—a 170× range in repair flexibility
- The rigid extreme ($\beta = 12$): one large witness component, nearly unique repair path
- The flexible extreme ($\beta = 2048$): 11 independent bilateral obligations, each with a binary choice

The scalar fee says "11 obligations." The repair entropy says whether those obligations are rigid or flexible—and by how much. A system near $\beta_{\min}$ is brittle; a system near $\beta_{\max}$ has deep redundancy in its repair landscape.

This is not theoretical. In the MCP corpus (§4), fee-matched groups show up to 14× variation in β—variation invisible to the scalar fee but fully predicted by K. The corpus also reveals an empirical regularity: most real-world compositions sit near the rigid bound, making the rare flexible cases structurally diagnostic.

4 K-Sufficiency in Practice: The MCP Corpus

4.1 Corpus and method

We compute $\beta(G)$ and $H_{\text{repair}}(G)$ for all 240 nonzero-fee pairwise compositions in the MCP server corpus (24 servers, 703 compositions total, 240 with fee > 0). For each composition, we build the full coboundary matrix, compute $K(G) = H^\top(I - P_O)H$ over exact rationals, extract the connected components of K's nonzero off-diagonal graph, and take $\beta = \prod |C_i|$ over components.

All 240 computations succeed with zero errors. For the stability analysis (§5), basis enumeration is tractable only at fee ≤ 14, which covers 239 of the 240 compositions (the excluded composition has fee = 22). All counts in §4 use the full 240; all counts in §5 use the 239 amenable to brute-force basis enumeration.

4.2 The corpus is overwhelmingly rigid

The central empirical finding: **81.7% of corpus compositions sit at the theoretical minimum** $\beta = \varphi + 1$ (normalized flexibility $F = 0$). The corpus mean $F = 0.088$, median $F = 0.000$. No composition exceeds $F = 0.585$.

Flexibility distribution:

Range	Label	Count	Percentage
[0.0, 0.2)	rigid	196	81.7%
[0.2, 0.4)	low-flex	1	0.4%
[0.4, 0.6)	mid-flex	43	17.9%
[0.6, 1.0]	high-flex	0	0.0%

This means: for the vast majority of real-world compositions, repair structure is nearly determined by the matroid alone. The fee constrains not just how many obligations exist, but also—empirically—how rigid the repair path is.

4.3 Fee-grouped entropy profile

Fee	N	β range	F range	Distinct motifs
1	89	2	0.000	3
2	67	3	0.000	2
3	10	4–6	0.000–0.585	4
4	1	9	0.505	1
10	25	84	0.448	1
11	36	12–168	0.000–0.513	6
12	7	13–252	0.000–0.515	5
13	3	36–378	0.148–0.517	3
14	1	15	0.000	1
22	1	588	0.268	1

Two patterns emerge:

1. **At low fee (1–2), no separation exists.** All fee-1 compositions have $\beta = 2$ (single K_2 dimension); all fee-2 compositions have $\beta = 3$ (single K_3 dimension). The repair multiplicity is determined by the fee alone.

2. **At higher fee (≥ 3), separation appears.** At fee = 11, β ranges from 12 (single 12-vertex component, $F = 0.000$) to 168 (four-component motif, $F = 0.513$)—a 14× spread. At fee = 12, the range is 13–252. The fee says "11 obligations"; the repair entropy reveals whether those obligations are rigid or flexible.

4.4 Motif atlas

The 240 compositions produce only 19 distinct component-size motifs. The five most common account for 95% of the corpus:

Motif	Count	β	Fee	Class
(2)	89	2	1	binary-flex
(3)	67	3	2	rigid
(12)	30	12	11	single-block
(7, 3, 2, 2)	25	84	10	component-heavy
(3, 2)	6	6	3	mixed

The (7, 3, 2, 2) motif is notable: all 25 compositions with fee = 10 share exactly this component structure. This reflects a structural regularity in the corpus—certain servers contribute a fixed-size block of hidden conventions that recurs across pairings.

Classification. We partition the 240 compositions into mutually exclusive motif classes:

- **single-pair** (one component of size 2, fee = 1): 89 compositions (37.1%). Trivially rigid—the fee forces a unique partition.

- **single-block** (one component of size ≥ 3): 103 compositions (42.9%). $F = 0$—all fee in one carrier component, giving minimum $\beta = \varphi + 1$.

- **multi-component** (two or more nontrivial components): 48 compositions (20.0%). These are the cases where fee underdetermines repair: same fee, different partition, different β.

Among the 48 multi-component compositions, the (7, 3, 2, 2) motif alone accounts for 25 (52%)—reflecting a specific structural pattern in the corpus where `filesystem`-paired compositions produce a dominant 7-vertex block alongside three smaller convention dimensions.

5 Operational Sparsity: Most Bases Are Never Optimal

5.1 The problem with structural flexibility

Repair entropy counts minimum-cardinality disclosure sets—bases of the witness matroid M/O. A composition with $\beta = 168$ has 168 structurally valid repairs. But how many of those are ever the cheapest repair under any plausible cost model?

The answer, empirically, is: far fewer than β.

5.2 Reachable bases and operational entropy

Definition 9. *Let $\mathcal{F}$ be a family of cost vectors $c : H \to \mathbb{Q}_{>0}$. The* reachable basis set *under $\mathcal{F}$ is*

$$B_{\mathcal{F}}(G) := \bigcup_{c \in \mathcal{F}} \arg\min_{B \in B(M/O)} c(B)$$

and the operational entropy *is $H_{\mathrm{op}}(G; \mathcal{F}) = \log |B_{\mathcal{F}}(G)|$.*

The *stability ratio* is $\rho(G; \mathcal{F}) = |B_{\mathcal{F}}(G)|/\beta(G) \in (0, 1]$: the fraction of structurally valid bases that are ever decision-relevant under the cost family $\mathcal{F}$.

5.3 The realizability theorem

Theorem 10 (Full realizability under unrestricted costs)**.** *For any Bulla composition G with* $\mathrm{fee} > 0$ *and any basis $B \in B(M/O)$, there exists a cost vector $c : H \to \mathbb{Q}_{>0}$ such that B is the unique minimum-cost repair.*

Proof. Set $c(h) = 1$ for all $h \in B$, and $c(h) = 2$ for all $h \notin B$. Since every basis has cardinality $\mathrm{fee}(G)$, $\mathrm{cost}(B) = \mathrm{fee}(G)$. Any other basis $B' \neq B$ has $|B' \cap B| = \mathrm{fee}(G) - d$ and $|B' \setminus B| = d$ for some $d \geq 1$. So $\mathrm{cost}(B') = (\mathrm{fee}(G) - d) \cdot 1 + d \cdot 2 = \mathrm{fee}(G) + d \geq \mathrm{fee}(G) + 1 > \mathrm{cost}(B)$. □

Corollary 11. *Under unrestricted linear costs, $\beta_{\mathcal{F}}(G) = \beta(G)$: every basis is reachable, and $\rho = 1$.*

This is the key structural fact. It says **operational sparsity is not a property of the matroid**. It is a property of the restricted cost family. The witness geometry admits all repairs; the deployment constraints collapse them.

5.4 Corpus stability profile

We compute stability under random integer costs $c(h) \in \{1, \ldots, 10\}$ (50 random draws + 4 structured cost models per composition). Basis enumeration limits this to fee ≤ 14 (239 of 240 compositions).

β	Reachable bases	ρ (stability ratio)
2	2	1.000
3	3	1.000
4–6	4–6	1.000
12	11	0.917
84	36	0.429
168	44	0.262
252	47–48	0.186–0.191
378	51	0.135

Two patterns:

1. **At low β, all bases are reachable.** For $\beta \leq 9$, $\rho = 1.000$—integer costs in $[1, 10]$ suffice to make every basis optimal. This is consistent with Theorem 10: the cost family is rich enough relative to the small basis set.

2. **At high β, ρ collapses.** For $\beta = 378$, only 51 of 378 bases (13.5%) are ever optimal. Most of the combinatorial repair space is operationally irrelevant.

5.5 Entropy does not fully predict stability

At fee = 11, the data shows a mild inversion:

Motif	β	Reachable	ρ
(7, 3, 2, 2, 2)	168	44	0.262
(7, 3, 3, 2)	126	39	0.310
(7, 4, 2, 2)	112	42	0.375
(8, 3, 2, 2)	96	38	0.396
(7, 3, 2, 2)	84	36	0.429
(12)	12	11	0.917

Notice: $\beta = 112$ produces 42 reachable bases, while $\beta = 126$ produces only 39. Lower entropy, more sensitivity. The component partition—not just the product—shapes the operational frontier.

This means repair entropy is a strong but not sufficient predictor of operational stability. The motif geometry carries information beyond the scalar β.

5.6 Live replay: same fee, different cheapest repair

The canonical pair is github+gtasks-mcp (motif (7, 4, 2, 2), $\beta = 112$) versus github+playwright (motif (7, 3, 3, 2), $\beta = 126$). Both have fee = 11. The structural difference is one component: the `state` dimension has 4 hidden fields in gtasks (including a cross-server match `github::state` ↔ `gtasks::status`) but only 3 in playwright (all github-internal).

Under a semantic cost model that prices stateful fields at 5, path fields at 3, pagination at 1, and directional fields at 2:

Composition	Motif	β	Repair cost	Expensive omission
github+gtasks-mcp	(7, 4, 2, 2)	112	26	`gtasks::status` (cost 5)
github+playwright	(7, 3, 3, 2)	126	24	`github::state` (cost 5)

The gtasks composition pays 8.3% more for repair—not because it has more obligations (same fee), but because the cross-server `state`↔`status` match creates a 4-element component where every element is expensive. The playwright composition can spread the state obligation across 3 cheaper github-internal fields.

This is the operational meaning of motif geometry. The scalar fee sees 11 obligations in both cases. The witness Gram sees that those obligations are distributed differently—and that difference costs real money.

5.7 Interpretation

The correct interpretation of repair entropy is now:

- β measures *structural* flexibility—the size of the exact repair space intrinsic to the matroid.

- $\beta_{\mathcal{F}}$ measures *operational* flexibility—how much of that space survives contact with deployment costs.
- $\rho = \beta_{\mathcal{F}}/\beta$ measures the *compression ratio*—how aggressively cost constraints collapse the combinatorial repair space.

Witness geometry creates a large combinatorial repair space. Deployment costs collapse that space to a sparse operational frontier. The deeper the combinatorial space (higher β), the more aggressive the compression (lower ρ).

6 The Exact Repair Functional and Decision Activation

Sections 3–4 established that fee does not determine repair entropy, and that repair entropy does not determine operational stability. This section derives the *exact* minimum repair cost as a closed-form functional of the witness components and cost vector, then uses it to identify when geometry changes the recommendation.

6.1 The sum-of-maxima theorem

Theorem 12 (Exact minimum repair cost)**.** *Let G be a composition in the DFD + CHP regime with nontrivial witness components $C_1, \ldots, C_k$ (each $|C_j| \geq 2$). For any additive cost vector $c : H \to \mathbb{Q}_{>0}$, the minimum repair cost is:*

$$\Sigma^*(G,\, c) \;=\; \sum_{j=1}^{k} \Big(\sum_{h \in C_j} c(h) \;-\; \max_{h \in C_j} c(h) \Big) \;=\; \sum_{h \in H_{\text{active}}} c(h) \;-\; A(G,\, c)$$

where $H_{\text{active}} = \bigcup_j C_j$ is the union of nontrivial components and

$$A(G,\, c) \;:=\; \sum_{j=1}^{k} \max_{h \in C_j} c(h)$$

is the geometry dividend*—the total disclosure cost that witness geometry allows the optimal repair to avoid.*

Proof. Under DFD + CHP, each nontrivial component C_j carries the uniform matroid $U(|C_j| - 1, |C_j|)$ (Theorem 3). A basis of the full witness matroid selects exactly $|C_j| - 1$ elements from each C_j—equivalently, it omits exactly one element per component. To minimize total cost, the optimal basis omits the most expensive element from each component. Hence:

$$\Sigma^*(G,\, c) = \sum_{j} \Big(\sum_{h \in C_j} c(h) - \max_{h \in C_j} c(h) \Big).$$ □

Computationally verified by brute-force basis enumeration on all 51 nonzero-fee compositions in the corpus at fee ≤ 14. Zero failures.

Definition 13 (Geometry dividend)**.** *The* geometry dividend $A(G,\, c) = \sum_j \max_{h \in C_j} c(h)$ *is the total disclosure cost that the optimal repair avoids by omitting the most expensive element from each witness component. It is the economic value of witness geometry under cost vector c: the amount of cost that structure lets you not pay.*

Corollary 14 (Robustness as top-two gap)**.** *The optimal repair basis is unique if and only if each component has a unique most expensive element. The robustness margin—the cost perturbation required to change the optimal omission in any component—is:*

$$\text{margin}(G,\, c) = \min_{j} \big(m_j^{(1)} - m_j^{(2)} \big)$$

where $m_j^{(1)} \geq m_j^{(2)}$ are the two largest costs in component C_j.

Remark 15 (The four projections of repair). The invariant hierarchy of §2.2 now has a sharper operational reading:

- **fee** $= \sum_j (|C_j| - 1)$ tells you how many obligations exist.
- $\beta = \prod_j |C_j|$ tells you how many structural repairs exist.
- $A(G, c) = \sum_j \max_{h \in C_j} c(h)$ tells you how much the optimal repair saves.
- **margin** $= \min_j (m_j^{(1)} - m_j^{(2)})$ tells you how stable that savings is.

Fee is a sum of component sizes minus one. β is a product of component sizes. Geometry dividend is a sum of componentwise maxima. All three are projections of the same component decomposition, but they compress different information.

6.2 Decision activation as a corollary

The sum-of-maxima theorem makes the activation principle exact, not merely empirical.

Definition 16 (Decision-relevant geometry)**.** *Let G_0 and G_1 be compositions with* $\text{fee}(G_0) = \text{fee}(G_1)$ *and different witness component motifs. The witness geometry is* decision-relevant under cost vector c *if $\Sigma^*(G_0, c) \neq \Sigma^*(G_1, c)$—the optimal repair cost changes as a function of motif at fixed fee.*

Definition 17 (Structural repair multiplicity)**.** *A composition G has* structural repair multiplicity *if its witness Gram $K(G)$ has two or more nontrivial connected components. In the DFD + CHP regime with connected carrier graphs, this is equivalent to $\beta(G) > \text{fee}(G) + 1$.*

Definition 18 (Cost heterogeneity on the witness support)**.** *A cost vector c is* heterogeneous on the witness support *of G if the per-component maxima $\{\max_{h \in C_j} c(h)\}$ are not all proportional to $|C_j| - 1$. Concretely: different components contribute different amounts to geometry dividend per unit of fee, so the motif partition affects Σ^*.*

Proposition 19 (Decision activation)**.** *In the DFD + CHP regime with additive costs, the witness geometry is decision-relevant at fixed fee if and only if:*

(i) the fee-matched compositions have different component partitions, and

(ii) the cost vector is heterogeneous on the witness support (Definition 18).

Under uniform costs ($c(h) = c_0$ for all h), $\Sigma^* = c_0 \cdot \text{fee}$ regardless of motif: geometry is inert.

Proof. By Theorem 12, at fixed fee with components $C_1, \ldots, C_k$:

$$\Sigma^*(G, c) = \sum_{h \in H_{\text{active}}} c(h) - \sum_j \max_{h \in C_j} c(h).$$

The first term $\sum c(h)$ depends on the field set but not on how it is partitioned into components. At fixed fee and fixed hidden field set, two motif partitions of the same fields yield different Σ^* exactly when their sums of componentwise maxima differ, which requires heterogeneous costs.

Under uniform costs $c(h) = c_0$, every component contributes $\max_{h \in C_j} c(h) = c_0$, so $A(G, c) = k \cdot c_0$ depends only on the number of components, not their sizes. But $\Sigma^* = c_0 \cdot \sum(|C_j| - 1) = c_0 \cdot \text{fee}$, independent of partition. □

6.3 The minimal activation diagram

At fee = 3 in the corpus, the two motif classes—(3, 2) and (4)—populate all four quadrants of the activation diagram under the environment-derived cost model:

	Motif	Mode	Composition	Bases	Costs	Outcome
Q1	(4)	rigid	notion+ns-mcp	4	all 9.0	nothing to decide
Q2	(3, 2)	flexible	notion+todoist	6	all 9.0	indifferent
Q3	(4)	rigid	xmind+playwright	4	17–21	no routing freedom
Q4	(3, 2)	flexible	xmind+notion	6	13–17	**routes to 13**

Q1 (rigid + uniform): Single component, all fields at sensitivity 2. Four bases, all cost 9.
Q2 (flexible + uniform): Two components, but all fields at sensitivity 2. Six bases, all cost 9. Geometry exists but the cost model assigns equal cost to every basis: $A = 3 + 3 = 6$ in either partition.
Q3 (rigid + heterogeneous): Single component with cheap (`directory`, cost 3) and expensive (`path`/`filePath`, cost 7) fields. Geometry dividend $A = 7$ (one max). No routing freedom.
Q4 (flexible + heterogeneous): Two components with different cost profiles. Geometry dividend $A = 7 + 3 = 10$, versus the single-component $A = 7$. The extra component contributes $+3$ of geometry dividend, reducing Σ^* from 17 to 13—a 23.5% reduction at identical fee.

Replication. Replacing notion with ns-mcp-server reproduces Q4 exactly: fee 3, $\beta = 6$, motif (3, 2), cost range [13, 17], $A = 10$.

6.4 Necessity and sufficiency

By Theorem 12 and Proposition 19:

- **Structural multiplicity is necessary but not sufficient.** Q2 has multiplicity ($\beta = 6$) but uniform costs make Σ^* partition-independent.
- **Cost heterogeneity is necessary but not sufficient.** Q3 has heterogeneous costs (cost spread 4 within the component) but only one component, so the partition cannot differ.
- **Together, they are sufficient.** Q4 has both, and $\Sigma^*(Q4) < \Sigma^*(Q3)$ at fixed fee.

Remark 20 (Separating exogenous costs from endogenous geometry). The sum-of-maxima theorem (Theorem 12) holds for *any* additive cost vector—it is a structural fact about the uniform matroid, not a property of a particular cost model. The environment-derived cost model used in the corpus verification is one instance: it assigns costs from field-name sensitivity and cross-server exposure, quantities that are exogenous to the witness geometry. Leverage scores, which are endogenous to $K(G)$, enter the cost model as a secondary weight. To verify that activation is not an artifact of this coupling, we confirm that activation persists under the purely exogenous cost assignment $c(h) = \text{field\_sensitivity}(h)$ alone (without leverage): $\Sigma^* = 10$ in Q4 versus $\Sigma^* = 14$ in Q3, a 28.6% separation.

Remark 21 (Scope and boundary of the exact regime). Proposition 19 and Theorem 12 are proved for the DFD + CHP regime, where each witness component carries a uniform matroid and every element is omittable. Outside this regime, two things can go wrong, and they push in opposite directions:

- **Coloops** (elements in every basis) make the formula *underestimate* Σ^*: the singleton coloop is ignored by the componentwise computation, but its cost is unavoidable.
- **Non-uniform essential structure** (components with rank deficit > 1) can make the formula *overestimate* Σ^*: the optimal basis omits more than one element per component, so the actual savings exceed the single per-component maximum.

These errors can partially cancel. The general exact decomposition is $\Sigma^* = B_{\text{forced}} + \Sigma^*_{\text{ess}}$, where B_{forced} is the sum of coloop costs and Σ^*_{ess} is the minimum-cost basis of the essential matroid (coloops contracted, loops deleted). The geometry dividend formula is exact when the essential matroid is a product of uniform matroids—precisely the DFD + CHP regime. Outside it, the formula is a surrogate induced by the uniform-product approximation, not a universal bound in either direction.

The entire current corpus is DFD (product formula matches brute-force enumeration on all 373 testable compositions), so the exact regime covers the observed operational landscape. See the companion boundary note for synthetic stress tests and the full three-level hierarchy.

7 Discussion

7.1 Four projections of one object

The witness Gram $K(G)$ is a single algebraic object. Its four increasingly operational projections are:

- **fee** $= \text{rank}(K) = \sum(|C_j| - 1)$—obligation count. How many hidden conventions?
- $\beta = \prod |C_j|$—structural flexibility. How many minimum repairs exist intrinsically?
- $A(G, c) = \sum \max_{C_j} c(h)$—geometry dividend. How much does the optimal repair save?
- **margin** $= \min_j(m_j^{(1)} - m_j^{(2)})$—robustness. How stable is that savings?

Fee is a sum of component sizes minus one. β is a product of component sizes. Geometry dividend is a sum of componentwise maxima. All four are projections of the same component decomposition; each compresses different information from K.

7.2 Why is the corpus so rigid?

The corpus's overwhelming rigidity (81.7% at $F = 0$) has a structural explanation. Most MCP server pairs share conventions through a single dominant semantic dimension. This produces a single large carrier graph component per dimension, giving $\beta = n_d$ and F near zero. Multi-component decompositions arise only when a server pair shares conventions across several independent dimensions—which requires richer semantic overlap than most pairs exhibit.

This is not a weakness of the theory. It means the real control-plane question is often not "which of many equivalent repairs should I choose?" but "do I have any real freedom at all?" The default operational regime is rigidity.

7.3 Relationship to Paper III

This paper corrects Paper III Corollary 4.3 (basis count via spanning trees) and replaces it with the uniform-matroid product formula $\beta = \prod |C_{d,j}|$. The correction is confined to basis counting; all other Paper III results (rank identity, Kron reduction, leverage, effective resistance) are unaffected.

The correction simplifies the theory: the uniform matroid $U(n-1, n)$ is simpler than the graphic matroid, the product formula $\beta = \prod n_d$ is simpler than $\beta = \prod \tau(\tilde{G}_d)$, and the sharp

bounds (Theorem 8) follow from elementary integer-partition arithmetic rather than spanning-tree combinatorics.

7.4 Limits and frontier

- **Regime.** The separation theorem (Theorem 3), product formula, sharp bounds, and exact repair functional (Theorem 12) work in the DFD + CHP regime, where each witness component carries a uniform matroid. Outside this regime, the formula can fail in either direction: coloops cause underestimation, non-uniform essential structure causes overestimation (§6). The general exact decomposition $\Sigma^* = B_{\text{forced}} + \Sigma^*_{\text{ess}}$ holds universally. K-sufficiency (Theorem 1) and full realizability (Theorem 10) hold in the general linear disclosure model without regime restrictions. The entire current corpus is DFD.
- **Cost family.** The stability computation uses a specific cost family (integer costs in $[1, 10]$). Operational sparsity is relative to this family; a broader or narrower family would produce different stability ratios. A full characterization would require the normal fan of the matroid basis polytope—the partition of cost space into regions where each basis is optimal.
- **Corpus support.** The current corpus has 24 servers producing 240 nonzero-fee pairwise compositions with only 19 distinct motifs. The high-entropy regime ($\beta > 100$) is populated almost entirely by `github`-paired compositions. Deliberate construction of compositions that stress underrepresented regions of the repair-flexibility landscape is needed to test whether the patterns generalize.
- **Inversions.** The entropy-stability inversion (§5.5) shows motif geometry carries information beyond β. Characterizing $\beta_{\mathcal{F}}$ for structured cost families, and understanding which motif features predict operational stability, is the natural next frontier.

8 Conclusion

In the linear disclosure model under DFD + CHP, the witness Gram $K(G)$ is a single object with three increasingly operational projections. Fee counts hidden obligations. Repair entropy counts structurally valid repairs. The stability ratio under a cost family identifies how many of those repairs are decision-relevant.

The central empirical finding is that, under integer costs in $[1, 10]$, repair spaces are combinatorially large but operationally sparse. At $\beta = 378$, only 13.5% of bases are ever optimal. Theorem 10 proves this sparsity is not intrinsic to the matroid—under unrestricted costs, every basis is realizable—but depends on the cost family. The compression from structural to operational flexibility is the decision-relevant quantity.

Six results anchor this:

1. **K-sufficiency**—in the linear disclosure model, the witness Gram determines the exact repair basis structure.
2. **Fee–multiplicity separation**—fixed fee does not determine repair entropy (DFD + CHP regime).
3. **Sharp bounds**—$\varphi + 1 \leq \beta \leq 2^{\varphi}$, both attained (DFD + CHP regime).
4. **Full realizability**—every basis is the unique optimum under some cost vector (general).
5. **Exact repair functional**—minimum repair cost equals total active hidden cost minus the sum of per-component maxima (DFD + CHP regime). The *geometry dividend* $A(G, c) = \sum_j \max_{h \in C_j} c(h)$ is the operational value of witness geometry.

6. **Decision activation**—witness geometry changes the optimal repair if and only if structural multiplicity and cost heterogeneity on the witness support coincide. This follows algebraically from the exact repair functional, not merely from corpus observation.

The corpus validates the theory (β formula = β enumerated on all 239 testable compositions; sum-of-maxima formula = brute-force minimum on all 51 at fee ≤ 14) and reveals the dominant regime: 81.7% of compositions are structurally rigid (minimum β at their fee). When flexibility exists, the exact repair functional (§6) identifies precisely when that flexibility changes action: structural freedom alone is not sufficient (uniform costs make Σ^* partition-independent), and cost heterogeneity alone is not sufficient (a single-component motif offers no routing freedom). The recommendation changes only when both ingredients coincide. In the DFD regime, the geometry dividend is the exact operational value of witness geometry—recovered as a special case of the universal decomposition $\Sigma^* = B_{\text{forced}} + \Sigma^*_{\text{ess}}$, where the forced cost vanishes and the essential matroid is uniform-product. Outside DFD, the formula is a surrogate that can fail in either direction: coloops cause underestimation, non-uniform essential structure causes overestimation. The exact decomposition, not the formula, is the universal invariant. The witness Gram sees all of this; the scalar fee sees none of it.

References

[1] J. Komkov, "The Coherence Fee," 2026.

[2] J. Komkov, "The Column-Matroid Backbone," 2026.

[3] J. Komkov, "The Witness Gram as a Kron-Reduced Laplacian," 2026.

[4] J. Komkov, "Local Validity Does Not Compose," 2026.

[5] J. Edmonds, "Matroids and the greedy algorithm," *Math. Programming* **1**, 127–136, 1971.

[6] R. Rado, "Note on independence functions," *Proc. London Math. Soc.* **7**, 300–320, 1957.

Frontier Extension and Impossibility

The first four parts of this volume should be read as the settled center: doctrinal statement (Part 0), formal hinge (Part I), empirical witness (Part II), protocol consequence (Part III), and the Bulla Formal Trilogy with its repair-geometry extensions (Part IV). Part V should be read under a different sign.

The SHEAF Protocol belongs to the same technical lineage as the earlier papers, but it reaches further. It asks what follows once compositional failure is granted: whether heterogeneous agents can coordinate under explicit impossibility conditions, whether topology can price structural remediation, and whether failure itself can be turned into a diagnosable object of protocol design.

That reach is precisely why this part must be marked as frontier rather than as already-closed center. The paper is substantial and worth inclusion. It is also visibly more open in places than the earlier layers. Some computation exists. Some proof notes exist. Some components remain placeholder-heavy. This part is included not to blur those distinctions, but to preserve them inside one ordered technical frame.

Three results from this frontier have matured enough to stand on their own.

The Linear Communication Bottleneck Theorem follows the SHEAF paper as a separate facsimile. It was harvested from the SHEAF proof-note surface, motivated independently from communication complexity and spectral graph theory, and subjected to an autonomy test requiring that its abstract and introduction be intelligible without reference to the broader program. The paper carries approximately 3% residual risk in one constant of the holonomy lower bound. It is included here as evidence of what the frontier produces when a result crosses the threshold from proof note to self-contained technical object.

The Coherence Cliff follows as a third facsimile. It is a scaling experiment across 500 composition graphs at 7 scales (5 to 50 agents) that tests whether sheaf-cohomological diagnostics are necessary rather than merely elegant. The central finding is a regime change: the best sheaf diagnostic maintains $R^2 > 0.96$ at all scales, while the strongest non-sheaf baseline — a Random Forest trained on all graph-topological features — degrades from $R^2 = 0.83$ at small scale to $R^2 = 0.52$ at large scale. The predictive gap nearly triples. The experiment operates on a deterministic symbolic layer with zero model noise, isolating structural obstruction from stochastic artifacts. Convention heterogeneity is grounded in real-world divergence patterns (ISDA settlement, Basel RCAP, vendor risk model calibration). This paper is the program's strongest empirical exhibit for the necessity of the formal machinery it proposes.

Local Validity Does Not Compose completes the frontier with the impossibility theorem that the doctrine paper assumes as a precondition. We prove that bounded-radius local audits and low-order spectral summaries cannot, in general, certify global semantic coherence of composed AI systems. We construct families of compositions — semantic twins — that are identical to every bounded local audit yet differ arbitrarily in a computable global obstruction. The proof uses high-girth graph families in a rank-1 signed model. Field-independence is established via Signed-Incidence Structure (Part IV-D), which lifts the impossibility result to the Q-valued Bulla model. Empirical validation on 500 synthetic compositions confirms spectral indistinguishability at scale. We complement the impossibility with an exact online maintenance algorithm and a canonical benchmark family. The paper is the formal complement to the Coherence Cliff's empirical regime change: where Coherence Cliff measures the gap, Local-Global Obstruction proves the gap is necessary.

The four papers in Part V — SHEAF, Communication Bottleneck, Coherence Cliff, and Local Validity Does Not Compose — together establish that compositional semantic verification is structurally hard at scale, that the structural diagnostic is necessary rather than convenient, and that no bounded-radius local audit can substitute for it.

The SHEAF Protocol

Structured Agreement Among Autonomous Agents
via Sheaf-Cohomological Obstruction Theory

John Komkov

March 9, 2026

Abstract

We introduce the *SHEAF protocol* (Semantic Heterogeneous Evaluation and Agreement Framework) for structured consensus among heterogeneous autonomous agents. Classical consensus (BFT, Paxos, CRDTs) assumes a shared state type; SHEAF addresses the harder problem where agents have different vocabularies, schemas, and overlapping but non-identical views of reality. The protocol has three core stages: a *diagnostic* computing the first Čech cohomology H^1 of the overlap network to determine whether global agreement is structurally possible; a *resolution* via enriched sheaf Laplacian diffusion when $H^1 = 0$; and a *topology auction* pricing minimal structural corrections when $H^1 \neq 0$, with overcollateralized bonds ensuring incentive compatibility. Soundness is proven: SHEAF never reports agreement when it is structurally impossible (no false trivial, mechanically verified in Lean 4). Convergence is proven for vector-space and lattice coefficients, conditional on a Laplacian Bridge Conjecture for quantale-enriched settings.

Our main empirical contribution is a sheaf-cohomological diagnostic for multi-model alignment that tests a structural property—global gauge equivalence ($H^1 = 0$)—invisible to existing pairwise metrics. We validate the diagnostic on synthetic data (sensitivity down to defect angle $\theta = 0.05$) and deploy it on 8 sentence-transformer models (3 architectures, 4 organizations, 4 training objectives). The diagnostic reports trivial H^1 (frustration SNR $\approx 1.0\times$ versus a noise-matched null at all PCA dimensions $k \in \{64, 128, 256, 384\}$), establishing that pairwise Procrustes methods are sufficient in this regime and providing the first empirical boundary condition for SHEAF's deployment. We distinguish the *representation sheaf* (where $H^1 = 0$) from the *communication sheaf*, where frustration decomposes into two independent components: encoder mismatch (which saturates frustration regardless of model similarity) and gauge-projection nonlinearity (which destroys cocycle structure only when models are sufficiently dissimilar). This identifies the lossy interface—not the internal geometry—as the structural locus of multi-agent coordination failure, and connects the sheafability conditions to a precise mechanism: reducing encoder mismatch recovers the gauge-equivalence baseline. Additional validation on the OAEI ontology alignment benchmark confirms the diagnostic correctly identifies transitive inconsistencies in 7 real-world ontologies.

Contents

1 The Problem

1.1 Classical consensus assumes homogeneity

The theory and practice of distributed consensus has produced a remarkable set of protocols—Paxos [7], Raft [8], practical BFT [9], Nakamoto consensus [10]—all sharing a common structural assumption: every participant proposes, validates, and commits values of the *same type*. A Paxos acceptor and a Paxos proposer disagree about which value to commit, but they agree completely about what a "value" is. A Byzantine node may lie about its value, but it lies in the same language as everyone else.

This assumption extends to modern coordination primitives:

- **Conflict-free replicated data types (CRDTs)** [11] achieve eventual consistency through commutative merge operations—but the commutativity requires a shared algebraic structure. Two replicas of a G-Counter can merge because they agree on what a "counter" is.
- **Multi-agent reinforcement learning (MARL)** optimizes joint policies through interaction—but provides no convergence guarantees when agents observe different state spaces, no impossibility detection when coordination is structurally infeasible, and no diagnostic when training fails to converge.
- **LLM orchestration frameworks** (CrewAI [17], LangGraph [18], AutoGen [19]) compose language model agents into workflows—but treat semantic alignment as a prompt engineering problem, with no formal characterization of when inter-agent agreement on concepts is achievable.

The missing case—the case this paper addresses—is *heterogeneous consensus*: agents with different vocabularies observing overlapping but non-identical realities, where the question is not "what value do we agree on?" but "can we agree at all?"

1.2 Three motivating scenarios

Example 1.1 (Enterprise AI orchestration)**.** Three LLM agents are deployed to process a legal due diligence corpus: a *retrieval agent* indexes documents by semantic similarity using an embedding model, a *reasoning agent* extracts contractual obligations using chain-of-thought prompting, and a *verification agent* checks extracted facts against a structured database. Each agent must determine which documents are "relevant" to the query.

The retrieval agent defines relevance by cosine similarity in embedding space. The reasoning agent defines relevance by logical connection to the query's legal issues. The verification agent defines relevance by whether the document contains verifiable factual claims.

Pairwise alignment succeeds: the retrieval and reasoning agents can reconcile their definitions over the set of documents they both process. The reasoning and verification agents can reconcile over their shared outputs. But three-way alignment fails silently—the three pairwise reconciliations are mutually inconsistent, and no amount of prompt tuning, re-ranking, or iterative refinement within the current architecture can fix it. The system returns inconsistent results depending on which pair of agents is consulted, and no agent can detect or report the inconsistency.

Example 1.2 (Autonomous swarm coordination)**.** A fleet of heterogeneous unmanned underwater vehicles (UUVs) operates in a communication-impaired environment. Each

vehicle has different sensors (sonar, lidar, thermal), different onboard maps (bathymetric, magnetic, acoustic), and different mission objectives (survey, mine countermeasures, environmental monitoring). The fleet must coordinate on a common operational picture: which zones are safe, which contain threats, and how coverage should be allocated.

Each vehicle constructs a local assessment of its operational area. Vehicles with overlapping coverage zones exchange assessments and reconcile them pairwise. But the reconciliations around a cycle of three vehicles—A reconciles with B, B with C, C with A—may be mutually inconsistent. Vehicle A's sonar-based "threat" is not the same concept as C's thermal-based "threat," even after both have been reconciled with B's lidar assessment. The fleet has a *structural* coordination failure that no amount of message-passing within the current communication topology can resolve.

This scenario is drawn from the Navy's mine countermeasures (MCM) program; see Riess [4] for the operational context.

Example 1.3 (Cross-jurisdictional data integration)**.** Three financial databases must be integrated: one following US GAAP, one following IFRS, and one following a local regulatory reporting standard. Each database has a well-defined schema for "revenue" within its own framework. Pairwise reconciliation is straightforward: GAAP-to-IFRS mappings exist, IFRS-to-local mappings exist, and GAAP-to-local mappings exist.

But the three pairwise mappings do not compose consistently. The GAAP-to-IFRS-to-local path classifies a given transaction differently from the direct GAAP-to-local path. The discrepancy is not a data error—it is a structural consequence of the three accounting frameworks having incompatible treatment of the same economic reality (e.g., revenue recognition timing for long-term contracts).

Global reconciliation requires structural change: a shared audit standard that creates a genuine three-way overlap (a set of transactions classified identically by all three frameworks), reducing the problem from a cycle to a contractible topology.

1.3 The question this paper answers

These three scenarios share a common structure:

1. Multiple autonomous agents, each with a *locally consistent* view of a shared domain.
2. Pairwise overlaps where agents can compare and reconcile their views.
3. A topology of overlaps that may or may not support global consistency.

The question SHEAF answers is:

> *When can heterogeneous agents achieve structured consensus? When they cannot, why not—and what is the cheapest structural change that makes consensus possible? And how do you incentivize agents to participate honestly in answering these questions?*

The answer has three parts:

1. A **topological diagnostic** that computes a single invariant—the first Čech cohomology class H^1—classifying whether the overlap structure supports global agreement.

2. A **resolution mechanism** that, when agreement is structurally possible ($H^1 = 0$), computes the optimal coordinated plan via sheaf Laplacian diffusion.

3. An **economic mechanism** that, when agreement is structurally impossible ($H^1 \neq 0$), runs a market for the minimal architectural changes that restore feasibility.

Contributions and status. To help the reader distinguish established mathematics from new contributions and open conjectures, we summarize:

- **Established:** the obstruction classification by H^1 (Theorem 2.10) follows from the Extension Torsor Lemma proved in the companion SCPI paper [1]. The sheaf Hodge theorem for vector-space coefficients [20] and the Tarski/Lawvere Laplacians [5, 6] are prior work. The connection between SHEAF's non-abelian diagnostic and group synchronization [22, 23] is a known equivalence, newly framed.
- **New (this paper):** the distributed H^1 diagnostic algorithm (Algorithm 1); the sheaf correction auction mechanism (Section 3.4) with its three correction types; the impossibility certificate as an economic signal; the incentive analysis under an explicit audit condition (Section 4).
- **Conjectured / open:** the Laplacian–Cohomology Bridge for enriched sheaves (Conjecture 2.13); the definition of H^1 with quantale-enriched coefficients (Remark 2.14); submodularity of H^1 reduction beyond the abelian regime (Section 7)—the abelian case is proven by a matroid rank argument. The resolution phase (Section 3.3) depends on the Laplacian Bridge Conjecture for its convergence guarantee in the enriched setting; for group-valued coefficients the guarantee follows from the known Hodge theorem.

Closest prior work and differentiation. The use of H^1 sheaf cohomology as an obstruction invariant for multi-agent coordination is emerging independently in several communities. Kurisummoottil Thomas and Chen [34] prove that $H^1 \neq 0$ characterizes irreducible semantic ambiguity in quantum semantic communication, yielding a rate bound $R = \log_2(\dim H^1)$—an information-theoretic result complementary to SHEAF's algorithmic and mechanism-design contributions. The Ghrist–Riess program [5, 6] has developed sheaf Laplacians through vector-space, lattice, and quantale-enriched settings, but has no H^1 obstruction theory in the enriched case—the central gap that SHEAF's Laplacian Bridge Conjecture (Conjecture 2.13) would fill. No prior work addresses economic mechanisms for resolving sheaf-cohomological obstructions.

SHEAF is not another consensus protocol. It is a *diagnostic and market wrapper* around a topological obstruction that other protocols do not name. Classical consensus (Paxos, Raft, BFT) assumes agents agree on types and differ only on values; SHEAF addresses the harder problem where agents have different vocabularies, schemas, or ontologies. The differentiator is the *certificate*: when SHEAF reports NONTRIVIAL, it returns a verifiable cycle certificate identifying the precise topological defect, and when it reports TRIVIAL, it returns a coboundary witness that agents can independently verify as a constructive coordination plan. No existing protocol provides either output.

This paper defines a research program rather than closing one. The proven results cover the abelian and solvable coefficient regimes, where the diagnostic is exact and the convergence is known. The enriched and non-solvable regimes—which cover the most practically important applications such as LLM agent coordination—remain open frontiers. Two load-bearing open problems define the boundary: the Laplacian Bridge Conjecture (Conjecture 2.13), which links the discrete H^1 diagnostic to the continuous Laplacian resolution, and the *M1–M2 extraction problem*—computing group-valued transition maps from agent outputs such as embeddings or structured schemas. On the latter, we build and validate a sheaf-cohomological diagnostic—Procrustes cocycle → connection Laplacian → spectral gap—that tests global gauge equivalence ($H^1 = 0$), a structural property invisible to pairwise metrics. Deployed on 8 sentence-transformer models, the

diagnostic reports trivial H^1 (SNR $\approx 1.0\times$ versus noise-matched null), establishing that pairwise alignment methods are sufficient in this regime and providing the first empirical boundary condition for SHEAF's deployment (Section 7).

2 The Obstruction

2.1 Overlap topology

Consider n agents, each observing a local domain. Some pairs of agents have *shared observables*—data or concepts visible to both. We represent this structure as a graph:

Definition 2.1 (Overlap graph)**.** The *overlap graph* $G = (V, E)$ has:

- Vertices $V = \{1, \ldots, n\}$: one per agent.
- Edges E: an edge (i, j) exists iff agents i and j share observables (have a nontrivial pairwise overlap).

The overlap graph captures which agents can communicate meaningfully (compare their views). But for consensus, what matters is not just pairwise communication but *higher-order structure*: do three agents share a common reference point?

Definition 2.2 (Čech nerve)**.** The *Čech nerve* N of the agent overlap structure is the simplicial complex with:

- A vertex i for each agent.
- An edge $[i, j]$ for each pairwise overlap.
- A triangle (2-simplex) $[i, j, k]$ iff agents i, j, k share a *common* observable—data visible to all three simultaneously.
- Higher simplices defined analogously.

The key distinction:

- **Contractible nerve** (tree-like information flow): every cycle of pairwise overlaps bounds a higher-dimensional simplex. Agreement is always possible.
- **Non-contractible nerve** (cycles without higher fill): some cycles of pairwise overlaps have no common reference point. Potential for irreconcilable disagreement.

Example 2.3 (Three agents, no triple overlap)**.** Three agents—Calendar, Email, Slack—have pairwise overlaps but no triple overlap. The nerve consists of three vertices and three edges forming a triangle boundary $\partial\Delta^2 \cong S^1$ (a circle), but the interior is *not filled*. This is the simplest non-contractible nerve.

2.2 The consistency check around loops

When two agents with a shared overlap compare their views, one of two things happens:

1. They **agree**: their local definitions, restricted to the overlap, coincide.
2. They **disagree**: their definitions differ, but there is a *reconciliation map*—a well-defined transformation that converts one agent's definition into the other's, restricted to the overlap.

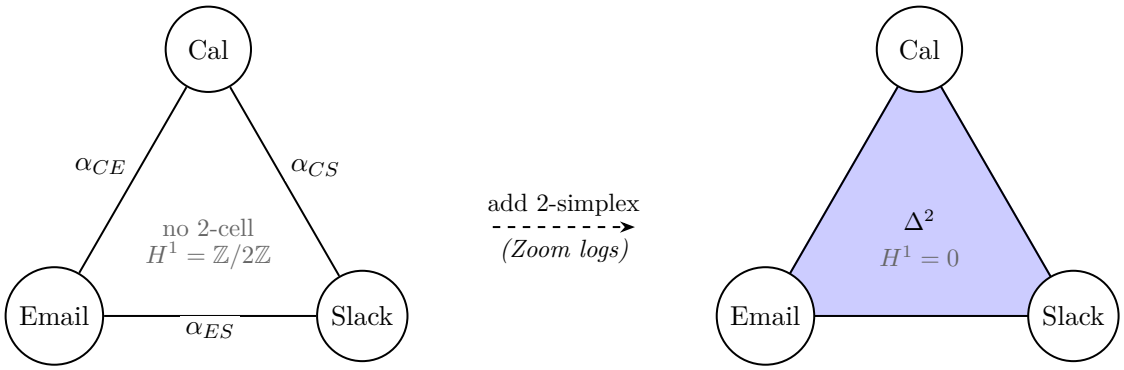


Figure 1: **Left:** three agents with pairwise overlaps but no triple overlap. The nerve is $\partial\Delta^2 \cong S^1$ (a 1-cycle with no filling 2-cell). **Right:** adding a shared data source (Zoom logs) creates a triple overlap, filling the triangle with a 2-simplex Δ^2. The nerve becomes contractible and H^1 drops from $\mathbb{Z}/2\mathbb{Z}$ to 0.

Definition 2.4 (Transition map)**.** For each edge (i, j) in the overlap graph, the *transition map* α_{ij} is the reconciliation: the definable equivalence that transforms agent i's local definition into agent j's on their shared overlap. Formally, α_{ij} is an element of the *equivalence group* $\mathrm{Equiv}(U_i \cap U_j)$—the group of invertible pairwise reconciliation maps on the shared overlap.

For the simplest non-trivial case—a single binary predicate ("is this a meeting?" yes/no)—the equivalence group is $\mathbb{Z}/2\mathbb{Z} = \{0, 1\}$: either the two agents agree ($\alpha_{ij} = 0$) or one agent's "yes" is the other's "no" ($\alpha_{ij} = 1$, a flip).

Remark 2.5 (When is Equiv a group?). The H^1 diagnostic requires Equiv to carry group structure (for the coboundary action and torsor classification). This is natural when reconciliations are *invertible*: schema isomorphisms, label permutations, rigid symmetries of a data structure. In many applied settings—fuzzy matching, lossy compression, non-invertible data transformations—reconciliations form a *monoid* or *preorder*, not a group. The protocol as proved applies to the group and groupoid regimes. Extending to monoid or quantale-valued coefficients is the subject of the enriched framework (Section 2.5) and remains open (Section 7).

Scope note: the enterprise LLM orchestration scenario (Example 1.1) may fall in the monoid regime if the reconciliation between agents' embedding spaces is non-invertible (e.g., a projection or lossy alignment). In that case, the full H^1 diagnostic does not apply directly; the enriched framework (Section 2.5) provides graded obstruction measurement, and the protocol's guarantees are conditional on Conjecture 2.13. The autonomous swarm (Example 1.2) and data integration (Example 1.3) scenarios, where reconciliations are coordinate transforms and schema isomorphisms respectively, fall naturally in the group regime.

Remark 2.6 (Sheafable interfaces). The SHEAF diagnostic requires algebraic structure on transition maps. When do agent interfaces provide it? We say an inter-agent communication interface is *sheafable* if it satisfies four conditions, each corresponding to a specific algebraic upgrade:

(i) **Fixed schema/grammar** for inter-agent messages. This ensures transition maps live in a known group of schema isomorphisms (e.g., field permutations, unit conversions) rather than in the space of arbitrary string transformations.

(ii) **Deterministic or low-variance decoding.** Transition maps must be stable across invocations; stochastic decoding (temperature > 0) turns group-valued maps into Markov-kernel-valued ones, moving the interface out of the group regime.

(iii) **Bounded coercions.** The monoid of type coercions between schema versions must be finitely generated, so that the enriched framework (Section 2.5) can assign finite obstruction costs.

(iv) **Local verifiability.** Each agent can check the cocycle condition on its own edges without global coordination, enabling the distributed certificate mechanism of **??**.

The engineering prescription: untyped multi-agent coordination—where agents exchange free-form natural language with no schema constraints—is structurally undiagnosable by certificate-based methods. SHEAF's demand for algebraic structure at the interface is itself the engineering contribution: it specifies the *minimum* type discipline required for coordination failures to become detectable. This connects directly to emerging agent interoperability standards (A2A Agent Cards, MCP tool schemas): each sheafability condition translates to a concrete requirement on the protocol's metadata layer.

Now consider three agents arranged in a cycle: $i \to j \to k \to i$. Each pair has a transition map. The *cocycle condition* asks: if we compose the transitions around the loop, do we get back to where we started?

Definition 2.7 (Cocycle)**.** A *cocycle* is a collection of transition maps $\{\alpha_{ij}\}$ for each edge (i, j), satisfying the *cocycle condition* on every triangle: for any triple (i, j, k) with a common overlap,

$$\alpha_{ij} + \alpha_{jk} = \alpha_{ik}$$

in the abelian case, or $\alpha_{ij} \cdot \alpha_{jk} = \alpha_{ik}$ in the non-abelian case.

When there is no triple overlap, the cocycle condition imposes no constraint: any collection of transition maps is a cocycle. This is the source of the problem.

Definition 2.8 (Coboundary)**.** A *coboundary* is a cocycle that can be "unwound" by adjusting each agent's local definition independently. If each agent i applies a local adjustment $\beta_i \in \mathrm{Equiv}(U_i)$ to its definition, the transition maps change:

$$\alpha'_{ij} = -\beta_i + \alpha_{ij} + \beta_j$$

in the abelian case, or $\alpha'_{ij} = \beta_i^{-1} \cdot \alpha_{ij} \cdot \beta_j$ in the non-abelian case. A cocycle is a coboundary if there exist local adjustments $\{\beta_i\}$ that make *all* transition maps trivial: $\alpha'_{ij} = 0$ for all (i, j).

The key insight: coboundaries represent "disagreements that each agent can fix on its own." A coboundary means the agents don't actually disagree about reality—they just labeled things differently, and each can relabel independently to align with the others.

2.3 The invariant

Definition 2.9 (First Čech cohomology H^1)**.** The *first Čech cohomology* $H^1(N, \mathrm{Equiv})$ of the nerve N with coefficients in the equivalence group Equiv is:

$$H^1 = \frac{\text{cocycles}}{\text{coboundaries}} = \frac{\{\text{all consistent transition data}\}}{\{\text{transition data fixable by local relabeling}\}}$$

When Equiv is abelian, H^1 is a group. When Equiv is non-abelian, H^1 is a pointed set (with distinguished element $0 =$ the trivial class).

Theorem 2.10 (Obstruction classification — after SCPI [1])**.** *Let n agents have local definitions of a concept, with pairwise reconciliation maps $\{\alpha_{ij}\}$ forming a cocycle. Then:*

1. *If $[\alpha] = 0$ in $H^1(N, \text{Equiv})$, there exist local adjustments $\{\beta_i\}$ making all agents' definitions globally consistent. A global consensus is structurally achievable.*
2. *If $[\alpha] \neq 0$ in $H^1(N, \text{Equiv})$,* no *local adjustments can make the agents' definitions globally consistent. The disagreement is* topological*: it arises from the structure of the overlap network, not from the content of any agent's data. It can be removed only by changing the network topology (adding overlaps) or changing the equivalence relation (redefining what counts as agreement).*

Proof sketch. The forward direction is the definition of coboundary. The reverse—that $[\alpha] = 0$ implies the existence of a global extension—follows from the Extension Torsor Lemma [1]: for finite overlap graphs (all examples in this paper), the lemma applies unconditionally. The impossibility direction is immediate: local adjustments change the cocycle by a coboundary, which cannot change the H^1 class. □

For the simplest case (Equiv $= \mathbb{Z}/2\mathbb{Z}$ on a circle nerve):

$$H^1(S^1, \mathbb{Z}/2\mathbb{Z}) = \mathbb{Z}/2\mathbb{Z} = \{0, 1\}$$

There is exactly one nontrivial class. Either the disagreement is fixable ([0]) or it is not ([1]), and a single bit tells you which.

What $H^1 \neq 0$ feels like operationally. In a deployed multi-agent system, a nontrivial H^1 manifests as follows. Pairwise reconciliation succeeds: each pair of agents can align their definitions and produce consistent outputs. But *which* global answer the system returns depends on *which pair* is consulted. Querying agents A and B yields one answer; querying B and C yields another; querying A and C yields a third. No agent is "wrong"—each pairwise alignment is internally consistent—but the three pairwise alignments are mutually incompatible. In logs, this appears as: results that change depending on routing path; reconciliation loops that converge to different values depending on initialization order; and dispute-resolution processes that restart indefinitely because each "fix" to one pair breaks another. The system is stuck, and no amount of retry, re-ranking, or prompt tuning within the current architecture can unstick it. The obstruction is not in the data; it is in the topology of the overlap network. SHEAF's certificate identifies the exact cycle where the inconsistency lives.

2.4 Worked example: Calendar/Email/Slack

We now demonstrate the full diagnostic on the three-agent scenario from Example 1.1, using the Calendar/Email/Slack overlap graph from [1].

Setup. Three agents—Calendar (C), Email (E), Slack (S)—each locally define the predicate `is_meeting`:

- C: a calendar event is a meeting if it has ≥ 2 attendees and a video link.
- E: an email thread is a meeting if it contains scheduling language and an attachment.
- S: a Slack thread is a meeting if it's in `#meetings` or contains a Zoom link.

Pairwise overlaps exist (events visible in two systems), but no triple overlap (no single record in all three). The nerve is $\partial\Delta^2 \cong S^1$.

Step 1: Compute transition maps. On each pairwise overlap, compare the two agents' definitions:

$$\alpha_{CE} \in \mathbb{Z}/2\mathbb{Z} : \text{do C and E agree on "is\_meeting" for their shared records?}$$
$$\alpha_{ES} \in \mathbb{Z}/2\mathbb{Z} : \text{do E and S agree?}$$
$$\alpha_{CS} \in \mathbb{Z}/2\mathbb{Z} : \text{do C and S agree?}$$

Each α_{ij} is 0 (agree) or 1 (disagree = one agent's yes is the other's no).

Step 2: The cocycle. The cocycle is the triple $(\alpha_{CE}, \alpha_{ES}, \alpha_{CS}) \in (\mathbb{Z}/2\mathbb{Z})^3$. Since there is no triple overlap, the cocycle condition imposes no constraint—any triple is a valid cocycle.

Step 3: Coboundaries. A coboundary is determined by local adjustments $(\beta_C, \beta_E, \beta_S) \in (\mathbb{Z}/2\mathbb{Z})^3$:

$$\alpha'_{CE} = -\beta_C + \alpha_{CE} + \beta_E, \quad \alpha'_{ES} = -\beta_E + \alpha_{ES} + \beta_S, \quad \alpha'_{CS} = -\beta_C + \alpha_{CS} + \beta_S.$$

The coboundary subgroup is generated by the image of the coboundary map $\delta\colon (\mathbb{Z}/2\mathbb{Z})^3 \to (\mathbb{Z}/2\mathbb{Z})^3$:

$$\delta(\beta_C, \beta_E, \beta_S) = (\beta_E - \beta_C,\ \beta_S - \beta_E,\ \beta_S - \beta_C)$$

The image has rank 2 (the third component is the sum of the first two in $\mathbb{Z}/2\mathbb{Z}$).

Step 4: Compute H^1.

$$H^1 = \frac{(\mathbb{Z}/2\mathbb{Z})^3}{\mathrm{im}(\delta)} = \frac{(\mathbb{Z}/2\mathbb{Z})^3}{(\mathbb{Z}/2\mathbb{Z})^2} \cong \mathbb{Z}/2\mathbb{Z}.$$

The invariant is the *parity of the total disagreement*: $\alpha_{CE} + \alpha_{ES} + \alpha_{CS} \pmod 2$.

Step 5: Diagnostic.

- **Case** $\alpha_{CE} + \alpha_{ES} + \alpha_{CS} = 0$: the cocycle is a coboundary. Each agent can independently adjust its definition of "is_meeting" to achieve global consistency. *SHEAF proceeds to resolution* (run the Laplacian to compute the optimal adjustment).
- **Case** $\alpha_{CE} + \alpha_{ES} + \alpha_{CS} = 1$: the cocycle represents the nontrivial class in H^1. No local adjustment works. *SHEAF proceeds to the topology auction* (bid to add a triple overlap—e.g., Zoom logs visible to all three agents—that fills the triangle and kills the obstruction).

Step 6: Resolution. Adding Zoom logs as a fourth data source creates a triple overlap: Zoom recordings are simultaneously visible as calendar events (C), email attachments (E), and Slack messages (S). The nerve becomes the filled triangle Δ^2 (contractible). $H^1(\Delta^2, \mathbb{Z}/2\mathbb{Z}) = 0$. The obstruction vanishes. The Laplacian now computes the unique globally consistent definition of `is_meeting`.

2.5 Two levels of failure: topology and approximation

The H^1 diagnostic answers a *discrete, topological* question: is the cocycle class trivial? Either a global section exists or it does not, and a cohomology class in a pointed set names the reason. But when the answer is "no," a second question immediately arises: *how close to a global section can you get?*

These are different questions, answered by different mathematics:

1. **The H^1 question** (topological, discrete). Is there a structural obstruction to the existence of a global section? The answer lives in $H^1(N, \text{Equiv})$—a group (abelian coefficients) or a pointed set (non-abelian coefficients). The obstruction is topological: it depends on the nerve N, not on the metric properties of the local data. Group structure on the coefficients Equiv is essential here—it defines the coboundary relation that separates resolvable disagreements from irresolvable ones. No optimization can remove a nontrivial H^1 class.

2. **The H^0 question** (analytic, graded). Given local data that may not match on overlaps, what is the *closest global section*—or, when none exists, the closest approximation? The answer is an optimization problem: minimize the total disagreement across all edges, measured in some cost structure. This is the domain of *weighted limits* and the *sheaf Laplacian* [5, 4].

The key insight is that these questions are *sequential*: H^1 determines *whether* an exact solution exists; the Laplacian determines *what* the best (exact or approximate) solution is. When $H^1 = 0$, the Laplacian converges to a true global section (zero residual). When $H^1 \neq 0$, the Laplacian still converges, but to a best approximation with nonzero residual—and the magnitude of that residual quantifies the cost of the topological obstruction. Richer coefficient structures yield quantitative rather than binary diagnostic information: with $\mathbb{R}^k$-valued stalks on a coordination graph, $\dim H^1$ counts the number of independent composition-failure modes, and each generator identifies a specific direction in which bilateral checks are blind [2].

The enriched framework. The graded H^0 question requires a notion of "cost of disagreement" richer than Boolean agree/disagree. This is provided by a *quantale* [15].

Definition 2.11 (Quantale [15])**.** A *quantale* $(\mathcal{V}, \otimes, k)$ is a complete lattice with a monoidal product $\otimes$ distributing over joins. It provides a cost structure for measuring disagreement:

- **Boolean** ($\mathcal{V} = \{0, 1\}$, $\otimes = \wedge$): agree or disagree. The classical setting.
- **Cost** ($\mathcal{V} = [0, \infty]$, $\otimes = +$): how expensive is the disagreement?
- **Fuzzy** ($\mathcal{V} = [0, 1]$, $\otimes = \min$): how confident is the agreement?
- **Contracts** ($\mathcal{V}$ = assume-guarantee lattice, $\otimes$ = contract composition): what assumptions must weaken for agreement?

Given a quantale $\mathcal{V}$ and a cocycle $\{\alpha_{ij}\}$, the *Laplacian residual* is the $\mathcal{V}$-valued cost of the best approximation:

$$\rho(\alpha) \;=\; \inf_{\{\beta_i\}} \bigotimes_{(i,j)\in E} d_{\mathcal{V}}\big(\alpha_{ij},\ \beta_i^{-1} \cdot e \cdot \beta_j\big)$$

where e is the identity cocycle and $d_{\mathcal{V}}$ is the $\mathcal{V}$-enriched distance. This is an optimization over 0-cochains $\{\beta_i\}$—an H^0-type computation. It does not "grade H^1"; rather, it measures the cost of the best local adjustment when the H^1 obstruction prevents an exact solution.

Remark 2.12 (Joint contribution with Riess). The enriched framework is developed jointly with Riess [4], whose SEAMAN project provides the computational machinery: cellular sheaves on graphs, the sheaf Laplacian, and categorical diffusion algorithms [5, 6]. Riess's weighted-limit approach computes the best approximation to a global section (H^0 question) using quantale-valued sheaves. The SCPI framework computes the topological obstruction (H^1 question) using group-valued cocycles. The central conjecture bridging the two:

Conjecture 2.13 (Laplacian–Cohomology Bridge)**.** *Let $\mathcal{F}$ be a $\mathcal{V}$-enriched cellular sheaf on the nerve N, whose restriction maps are $\mathcal{V}$-isometries with an identifiable group of automorphisms* Equiv *at each stalk (this holds when restriction maps are translations, rigid motions, or schema isomorphisms; it does not hold for general $\mathcal{V}$-functors). Let $\{\alpha_{ij}\}$ be the transition cocycle in* Equiv *induced by composing restriction maps on overlaps. Then the Laplacian residual $\rho(\alpha) = k$ (the monoidal unit, i.e., zero cost) if and only if $[\alpha] = 0$ in $H^1(N, \text{Equiv})$. Equivalently: the enriched sheaf Laplacian converges to a* nontrivial *global section if and only if the discrete topological obstruction vanishes.* Note: *this is the* ***group-coefficient*** *Bridge; for lattice-valued sheaves, the Tarski operator's fixed-point structure reveals a finer phenomenon—*partial obstruction*—where consensus survives on the cocycle's invariant sublattice even when $H^1 \neq 0$ (see Remark 2.18).*

Remark 2.14 (Known cases and the enrichment obstacle). For **vector-space-valued sheaves**, the conjecture is known: the sheaf Hodge theorem of Hansen and Ghrist [20] gives $\ker \Delta_k \cong H^k$ for all k, and the spectral gap satisfies $\lambda_1(L_{\mathcal{F}}) > 0$ iff $H^0 = 0$ [21]. Sheaf Laplacian diffusion converges to the orthogonal projection onto H^0 [21]. For **lattice-valued sheaves**, Ghrist and Riess [5] define the Tarski Laplacian whose fixed points contain global sections, but explicitly note that the quotient construction $H^1 = \ker \delta^1 / \operatorname{im} \delta^0$ does not make sense for lattices, since lattice homomorphisms lack additive inverses. For **quantale-enriched sheaves**, the Lawvere Laplacian [6] computes fuzzy H^0 but no higher cohomology. No published work defines H^1 for quantale-enriched cellular sheaves in the Lawvere–Ghrist–Riess framework; this is confirmed as a genuine gap in the literature.

The central technical obstacle for the enriched Laplacian Bridge Conjecture is therefore: *how to detect a nontrivial H^1 obstruction in a setting where the quotient* ker / im *is undefined.* Four paths are available:

1. **Algebraic:** define H^1 as a *pointed set with quantale-valued metric*—analogous to non-abelian H^1, where the group structure on H^1 is already absent but the triviality question remains well-posed. The obstruction cost $\rho(\alpha)$ defined above provides the metric.
2. **Spectral:** detect the obstruction via the *spectral gap of a degree-1 connection Laplacian* [23], bypassing the quotient entirely: a positive spectral gap would witness $H^1 = 0$ without computing H^1 as a group.
3. **Topological (suggested by the tropical test case, Section 2.6):** define $H^1 \neq 0$ as the condition that the enriched Laplacian has *no finite fixed point*—a topological characterization (non-compactness of the orbit under the Laplacian endofunctor) rather than an algebraic one (nontrivial quotient). The tropical case

(Proposition 2.15) confirms this: Bellman-Ford diverges iff $H^1 \neq 0$. This path is potentially the most natural for the enriched setting, since it requires only the Lawvere Laplacian machinery that Ghrist–Riess have already built, without importing algebraic constructions (quotients, inverses) that quantales lack.

4. **Fixed-point trichotomy (suggested by Riess):** When the underlying preorders of the stalks satisfy the descending chain condition and the restriction maps are cocontinuous (preserving weighted colimits), the Lawvere Laplacian is guaranteed to converge—but the bottom element of each stalk is always a global section (the *trivial section*, analogous to "every agent reports nothing"). The enriched analog of $H^1 \neq 0$ may therefore not be "no fixed point exists" but rather "the Laplacian initialized at nontrivial data collapses to the trivial (bottom) section rather than converging to a nontrivial global section." In the tropical case, the DCC fails on $\mathbb{R}$ and the frustrated Laplacian diverges entirely (Proposition 2.15); for DCC-satisfying quantales, collapse to bottom may be the computable enriched diagnostic. This path reframes the Bridge as a question about the *basin of attraction*: does nontrivial initial data reach a nontrivial section, or does it get forced to the trivial one?

Resolving this obstacle is the key open mathematical problem for the enriched theory.

If the conjecture holds, the two frameworks interlock precisely:

1. H^1 provides the *binary diagnostic*: is exact consensus structurally possible?
2. The Laplacian provides the *quantitative resolution*: what is the optimal (exact or approximate) consensus, and what does it cost?
3. When $H^1 \neq 0$, the residual $\rho(\alpha) > k$ quantifies the *value of a topology edit*: the coordination surplus unlocked by moving from $H^1 \neq 0$ to $H^1 = 0$. This residual is the natural objective function for the topology auction (Section 3.4).

2.6 Evidence for the Laplacian Bridge: the tropical triangle

The Laplacian Bridge Conjecture (Conjecture 2.13) asserts that enriched Laplacian convergence characterizes H^1 triviality. We now demonstrate this for the simplest enriched case: the *tropical quantale* $\mathcal{Q} = ([0, \infty], \geq, +, 0)$ on the triangle graph. This is the first concrete evidence for the Bridge in a non-vector-space setting.

Setup. Consider the triangle graph G with vertices $V = \{A, B, C\}$ and edges $E = \{AB, BC, CA\}$, with no 2-cells. Define a $\mathcal{Q}$-enriched cellular sheaf $\mathcal{F}$ with:

- **Stalks:** $\mathcal{F}(v) = (\mathbb{R}, |\cdot|)$ for each vertex—the reals as a Lawvere metric space.
- **Restriction maps:** translations by weights $w = (w_{AB}, w_{BC}, w_{CA})$. On edge AB: vertex A restricts as $x_A \mapsto x_A$, vertex B as $x_B \mapsto x_B + w_{AB}$. Similarly for the other edges. These are isometries, hence valid $\mathcal{Q}$-functors.

A *global section* is $(x_A, x_B, x_C) \in \mathbb{R}^3$ with $x_A = x_B + w_{AB}$, $x_B = x_C + w_{BC}$, $x_C = x_A + w_{CA}$. Substituting cyclically:

$$x_A = x_A + \underbrace{(w_{AB} + w_{BC} + w_{CA})}_{\omega}$$

so a global section exists iff the *frustration* $\omega = 0$. For the constant $\mathbb{R}$-sheaf on the triangle, $H^1 \cong \mathbb{R}$, with the H^1 class being exactly ω.

The tropical Laplacian is Bellman-Ford. The natural "Laplacian diffusion" in the tropical semiring (+ replaces ×, min replaces +) is the Bellman-Ford shortest-path relaxation. At each step, each vertex updates to the minimum of what each neighbor's restriction map projects onto it:

$$x_A \leftarrow \min(x_B + w_{AB},\ x_C - w_{CA})$$
$$x_B \leftarrow \min(x_A - w_{AB},\ x_C + w_{BC})$$
$$x_C \leftarrow \min(x_B - w_{BC},\ x_A + w_{CA})$$

This is precisely Bellman-Ford relaxation on a directed graph with edge weights:

- $B \to A$: weight w_{AB}; $A \to B$: weight $-w_{AB}$
- $C \to B$: weight w_{BC}; $B \to C$: weight $-w_{BC}$
- $A \to C$: weight w_{CA}; $C \to A$: weight $-w_{CA}$

The directed cycle $A \to B \to C \to A$ has total weight $-(w_{AB} + w_{BC} + w_{CA}) = -\omega$.

Two test cases.

1. **Unfrustrated** ($w = (1, 2, -3)$, $\omega = 0$). No negative cycle. Bellman-Ford converges in 3 iterations to the fixed point $(0, -1, -3)$ with residual 0—a global section.

2. **Frustrated** ($w = (1, 2, -1)$, $\omega = 2$). Negative cycle of weight -2. Bellman-Ford never reaches a fixed point: values drift to $-\infty$. The L^∞ residual (max edge discrepancy) remains constant at exactly $|\omega| = 2$ at every iteration.

For comparison, the vector-space (L^2) Laplacian converges in both cases: to residual 0 when $\omega = 0$, and to residual $\omega^2/3 = 4/3$ when $\omega = 2$ (the Hodge-theoretic minimum).

Proposition 2.15 (Bridge for the tropical triangle)**.** *For a translation sheaf on a cycle graph with tropical quantale coefficients, the Bellman-Ford (tropical Laplacian) relaxation converges to a finite fixed point with zero residual if and only if $\omega = 0$, i.e., if and only if H^1 is trivial.*

Proof. Bellman-Ford converges to finite values iff the directed graph has no negative cycle [37, 38]. The only cycle has weight $-\omega$. The reverse cycle has weight ω. One of these is negative iff $\omega \neq 0$ iff $H^1 \neq 0$. □

Remark 2.16 (Idempotency and the residual dichotomy). The tropical quantale is *idempotent* (min is idempotent): $a \oplus a = a$. This creates a sharp dichotomy absent from the vector-space case. The L^2 Laplacian converges to a finite nonzero residual $\omega^2/3$ when $H^1 \neq 0$—there is "approximate agreement." The tropical Laplacian admits no such middle ground: either it converges to residual 0 (exact agreement) or it diverges entirely (no finite fixed point). The idempotent residual stays constant at $|\omega|$ throughout the divergence, quantifying the cost of frustration without ever reducing it.

The vector-space (L^2) Laplacian already provides the non-idempotent comparison: its join is ordinary addition ($a \oplus b = a + b$, which is non-idempotent: $a + a \neq a$), and it converges to a finite nonzero residual $\omega^2/3$ when $H^1 \neq 0$. More generally, the minimum L^p residual for $p \in [1, \infty]$ is finite for all p (it equals $|\omega|$ for $p = 1$, $\omega^2/3$ for $p = 2$, and $|\omega|/3$ for $p = \infty$—all nonzero iff $\omega \neq 0$). The tropical case is the "$p = -\infty$ limit": the idempotent extreme where no finite residual exists. The hierarchy $L^1 \to L^2 \to L^\infty \to$ tropical traces a smooth transition from "distribute the frustration across edges" to "frustration cannot be distributed at all."

Consequence for the protocol: in the tropical regime, the Laplacian does not provide *graceful degradation* when $H^1 \neq 0$—there is no finite "best approximation" to return. The diagnostic detects the obstruction, but the resolution phase has no approximate output. Approximate consensus when $H^1 \neq 0$ is therefore quantale-dependent: available for non-idempotent enrichments (vector-space, cost), unavailable for idempotent ones (tropical, Boolean). The idempotency of the quantale thus determines whether the protocol can offer a useful partial answer or only a binary verdict.

Remark 2.17 (Scope and limitations of the tropical test case). Proposition 2.15 establishes the Bridge for the simplest enriched case. Three limitations constrain how far this evidence generalizes:

1. **The tropical quantale is the easiest enriched case.** Bellman-Ford works because $([0, \infty], \geq)$ is a total order and path relaxation is monotone. Most quantales of practical interest—the contract quantale, fuzzy logic, assume-guarantee lattices—are partially ordered, and the Lawvere Laplacian over them does not reduce to a known algorithm. The general Laplacian Bridge Conjecture requires a fixed-point theorem for enriched Laplacians over partially ordered quantales, which is a harder problem. The tropical case provides evidence and mechanistic intuition, not a proof template.
2. **The triangle graph has only one independent cycle.** On three nodes, $H^1 \cong \mathbb{R}$ is generated by the single frustration ω. We have verified the Bridge on the *figure-eight graph* (two triangles sharing a vertex, $\dim H^1 = 2$): Bellman-Ford diverges on exactly the frustrated cycles and converges on the unfrustrated ones, independently. Across all four cases (both unfrustrated, left only, right only, both frustrated), the Bridge holds and detection is local—each cycle's obstruction is detected independently through the shared vertex, and independent frustrations do not cancel. This addresses the single-cycle limitation, though the stalks remain one-dimensional.
3. **The stalks are one-dimensional.** With $\mathcal{F}(v) = \mathbb{R}$ and translation restriction maps, the sheaf is the simplest possible. Higher-dimensional stalks (multiple coordinated quantities per agent) and non-isometric restriction maps would test the Bridge under more realistic conditions.

2.7 Evidence for the Laplacian Bridge: Boolean lattice sheaves

To test the Laplacian Bridge Conjecture in the lattice setting—where Riess's fixed-point trichotomy (Remark 2.14, path 4) predicts qualitatively different behavior from the tropical case—we run the Tarski-style meet operator on cellular sheaves with stalks $2^{[2]} = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$ (Boolean lattice, 4 elements) and S_2-automorphism restriction maps (identity or atom-swap $a \leftrightarrow b$). On the triangle and figure-eight graphs, exhaustive enumeration over all $4^3 = 64$ (resp. $4^5 = 1024$) states confirms: *nontrivial Tarski fixed points—those strictly between $\bot$ and $\top$—exist if and only if $H^1 = 0$*, across all cocycle configurations. In the frustrated case ($H^1 \neq 0$), the iteration from nontrivial initial data either *oscillates* (period-2 cycles between atom-swapped states) or *collapses to $\bot$*, rather than diverging as in the tropical case. This extends the idempotency dichotomy of Remark 2.16 to a **behavioral trichotomy** (Figure 2): vector-space Laplacians degrade gracefully (finite nonzero residual), tropical Laplacians diverge, and lattice Tarski operators oscillate or collapse.

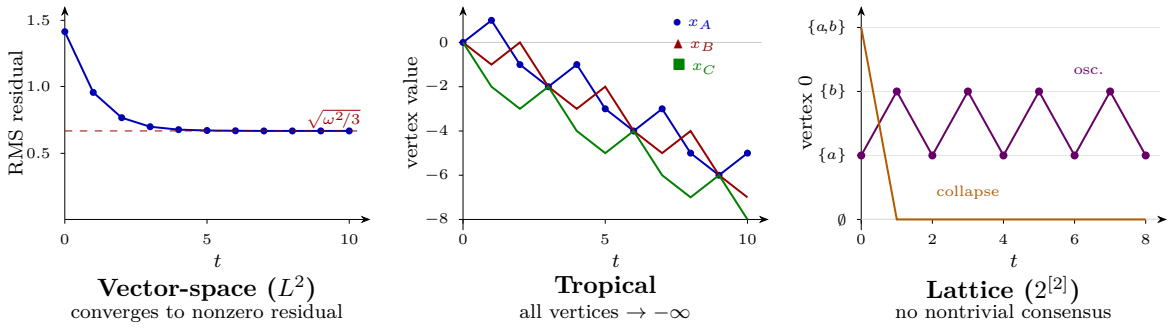


Figure 2: The behavioral trichotomy under frustration ($H^1 \neq 0$), all on the triangle graph with $\omega = 2$. Data from actual simulation (script: `simulations/trichotomy_data.py`). **Left:** the L^2 sheaf Laplacian ($\eta = 0.15$) converges to a nonzero RMS residual $\sqrt{\omega^2/3} \approx 0.667$—approximate consensus is available. **Center:** Bellman-Ford (tropical Laplacian) with weights $(1, 2, -1)$; all three vertex values diverge to $-\infty$, no finite fixed point exists. **Right:** the Tarski operator on $2^{[2]}$ with S_2 cocycle (swap, swap, swap); from $(\{a\}, \{a\}, \{a\})$ the operator oscillates with period 2 between atom-swapped states; from $(\{a, b\}, \{a\}, \{b\})$ it collapses to $\perp$ in 2 steps. The enrichment quantale determines whether the protocol offers a useful partial answer or only a binary verdict.

Remark 2.18 (Partial obstruction on richer lattices). For richer lattices, the picture is more subtle. On $2^{[3]}$ (8-element Boolean lattice) with S_3-automorphism restriction maps, a frustrated cocycle whose permutation has non-extreme fixed points (e.g., the transposition (ab) fixes $\{c\}$ and $\{a, b\}$ in the lattice) admits nontrivial global sections *on the fixed sublattice* $L^\sigma = \{x \in L : \sigma(x) = x\}$, even though $H^1 \neq 0$. The group-coefficient Bridge (nontrivial Tarski fixed point iff $H^1 = 0$) holds when the cocycle permutation *acts freely on non-extreme lattice elements* (as it does for S_2 on $2^{[2]}$ and for any transitive permutation on $2^{[k]}$), but not when the cocycle leaves a nontrivial sublattice invariant. This reveals a phenomenon absent in the group-valued theory: *partial obstruction*, where $H^1 \neq 0$ blocks full-information agreement but the frustration-invariant sublattice L^σ still permits residual consensus. The Tarski operator does not merely detect whether consensus exists—it *computes the maximal sublattice on which consensus survives*. The refined lattice Bridge may therefore be: the Tarski operator converges to a fixed point in L^σ, nontrivial whenever L^σ itself is nontrivial. Characterizing L^σ for finite distributive lattices (which cover the assume-guarantee contracts in SEAMAN [4]) requires understanding join-irreducibles under the cocycle's automorphism, and warrants joint investigation with the Ghrist–Riess program.

3 The Protocol

The SHEAF protocol operates in three phases. Phase 1 (Registration and Diagnostic) runs unconditionally. Phase 2 (Resolution) runs when the diagnostic is favorable. Phase 3 (Topology Auction) runs when the diagnostic reveals structural impossibility. Figure 3 shows the full protocol loop.

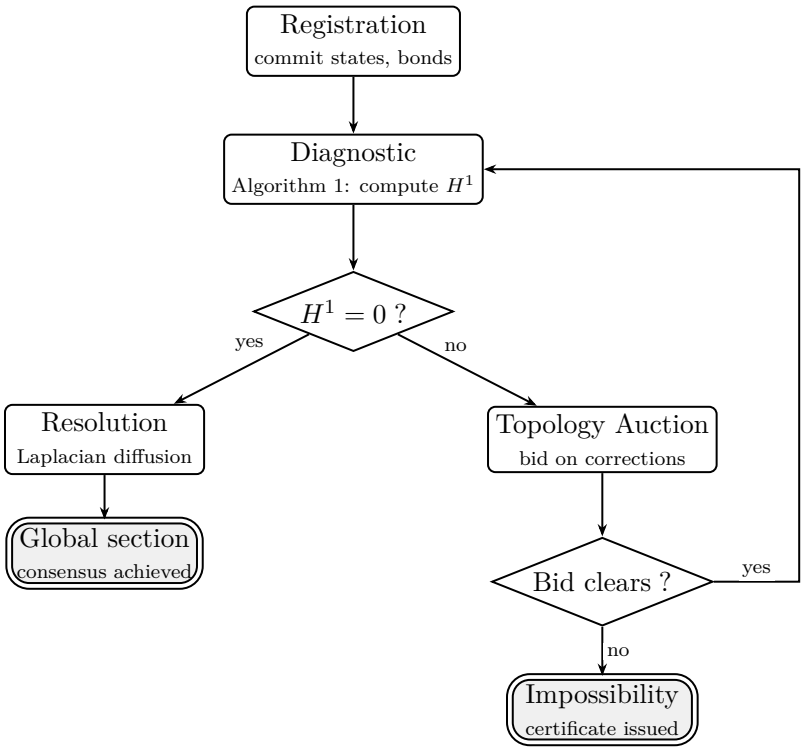


Figure 3: The SHEAF protocol loop. After registration, the H^1 diagnostic determines the protocol path. If $H^1 = 0$, the Laplacian resolves to a global section. If $H^1 \neq 0$, the topology auction prices corrections; a successful correction re-triggers the diagnostic. If no economically rational correction exists, an impossibility certificate is issued.

3.1 Phase 1: Registration

Each agent i publishes three items to a shared bulletin board (which may be a blockchain, a distributed log, or any append-only broadcast channel):

1. **State commitment**: a cryptographic hash $h_i = H(\sigma_i)$ of its local state σ_i. The raw state stays local; only the commitment is published.

2. **Restriction maps**: for each neighbor j in the overlap graph, agent i publishes its *restriction map* $\rho_{i \to ij} \colon \sigma_i \to \sigma_i|_{U_i \cap U_j}$—how its local view projects onto the shared overlap. This reveals the *structure* of the local view (which variables are shared) but not the *content* (what values those variables take).

3. **Bond**: a deposit b_i denominated in the settlement asset. The bond is locked for the duration of the protocol and is released upon successful completion or slashed upon detected dishonesty.

Remark 3.1 (Privacy). Only restriction maps and commitments are published; raw data stays local. The protocol reveals the *topology* of each agent's knowledge (what it knows about, not what it knows). For applications requiring stronger privacy, the state commitment can be replaced by a zero-knowledge proof of membership in a committed state set, and the restriction maps can be computed via secure multi-party computation (see Section 7).

3.2 Phase 2: Distributed diagnostic

Each pair of communicating agents (i, j) locally computes their transition map $\alpha_{ij} \in \text{Equiv}(U_i \cap U_j)$ by comparing their restrictions to the shared overlap. The diagnostic then determines whether the resulting cocycle $\{\alpha_{ij}\}$ represents the trivial class in H^1.

Algorithm 1 Distributed H^1 Diagnostic (Abelian Case)

Require: Overlap graph $G = (V, E)$; transition maps $\alpha_{ij} \in \text{Equiv}(U_i \cap U_j)$ for each edge
Ensure: H^1 class: TRIVIAL (with witness $\{\beta_i\}$) or NONTRIVIAL (with certificate)

```
Choose spanning tree T ⊆ E
Set β_root = 0
for each vertex i in BFS order from root along T do
    β_i ← β_parent(i) + α_parent(i),i                ▷ Propagate along tree
end for
for each non-tree edge (i, j) ∈ E \ T do
    r_ij ← −β_i + α_ij + β_j                         ▷ Residual on back-edge
    if r_ij ≠ 0 then
        return NONTRIVIAL with certificate: cycle through T from i to j plus edge (i, j)
    end if
end for
return TRIVIAL with witness {β_i}
```

Complexity. The algorithm requires $O(|V|)$ messages along the spanning tree and $O(|E| - |V| + 1)$ checks on back-edges. Total communication: $O(|E|)$ messages, each of size $O(\log |\text{Equiv}|)$ bits. The number of communication rounds is $O(\text{diam}(G))$.[1]

Disconnected overlap graphs. When the overlap graph G has connected components $G_1, \ldots, G_c$, Algorithm 1 runs independently on each component using a spanning *forest* (one tree per component). The cohomology decomposes as $H^1(G, \text{Equiv}) \cong \bigoplus_{k=1}^{c} H^1(G_k, \text{Equiv})$, so the global diagnostic is TRIVIAL iff every component diagnostic is TRIVIAL. Agents in distinct components have no overlaps and thus no coordination requirement; the protocol correctly treats them as independent subproblems.

Remark 3.2 (Connection to LDPC decoding). For abelian Equiv $= \mathbb{Z}/p\mathbb{Z}$, the H^1 computation reduces to solving a linear system over GF(p). This is structurally identical to LDPC decoding [26]: the parity-check matrix is the coboundary operator δ^0, and belief propagation over GF(p) provides a well-characterized distributed solver. On tree-like overlap graphs: exact convergence in $O(\text{diam}(G))$ rounds with $O(p \cdot |E|)$ total communication. On loopy graphs: convergence is not guaranteed but performs well when the graph is locally tree-like. Algorithm 1 can be viewed as a deterministic variant of this approach using a spanning tree to avoid loops.

Non-abelian case: group synchronization. When Equiv is non-abelian, the coboundary equation $\alpha'_{ij} = \beta_i^{-1} \cdot \alpha_{ij} \cdot \beta_j$ does not decompose linearly. This is precisely the *group synchronization problem*—a well-studied problem in signal processing and robotics [22, 23, 24].

[1] The spanning-tree requirement is load-bearing: formal verification confirmed that completeness is *false* without it and *true* with it. See the Formal Verification paragraph in Section 6.

Given noisy pairwise measurements $g_{ij} \in G$ of relative group elements, recover absolute elements $\beta_i \in G$ minimizing frustration.

SHEAF frames group synchronization cohomologically: the measurements $\{g_{ij}\}$ are a cocycle, the solution $\{\beta_i\}$ is a coboundary trivializing it, and the frustration is the H^1 class. Existing algorithms apply directly:

- Spectral methods for compact groups (angular synchronization [22]): polynomial time.
- Connection Laplacian with Cheeger-type bounds relating spectral gap to frustration [23].
- Cycle-edge message passing for robust synchronization [24]: directly implementable in SHEAF's distributed setting.

By Bulatov's CSP dichotomy theorem [25], *exact* synchronization is polynomial for abelian, nilpotent, and solvable groups, and NP-complete for non-solvable groups. However, *approximate* synchronization (minimizing frustration) is polynomial for all compact groups via spectral/SDP methods. The diagnostic guarantees therefore depend on the coefficient regime:

- **Abelian / solvable** Equiv**:** Algorithm 1 computes H^1 exactly in $O(|E|)$ messages. The diagnostic is exact.
- **Compact non-solvable** Equiv**:** Spectral/SDP methods detect obstruction with high probability when the frustration exceeds a threshold depending on the spectral gap [23]. The diagnostic is *three-valued*: it returns TRIVIAL (with coboundary witness $\{\beta_i\}$), NONTRIVIAL (with cycle certificate), or INCONCLUSIVE (frustration is below the detection threshold). The TRIVIAL and NONTRIVIAL outputs are sound; INCONCLUSIVE routes to the auction as "obstruction suspected."
- **General finite non-solvable** Equiv**:** Exact computation of $H^1 = 0$ is NP-complete. SHEAF falls back to the spectral heuristic, with the same three-valued output.

3.3 Phase 3: Resolution ($H^1 = 0$)

When the diagnostic returns TRIVIAL, a global consensus is structurally achievable. The witness $\{\beta_i\}$ from Algorithm 1 tells each agent *how* to adjust its local definition. But the raw adjustment may not be optimal—it depends on the choice of spanning tree.

The *enriched sheaf Laplacian* [4, 5, 6] provides the optimal adjustment. Each agent treats its local assignment as an initial condition and iteratively diffuses toward consistency. For vector-space coefficients, the iteration takes the familiar gradient form:

$$\sigma_i^{(t+1)} = \sigma_i^{(t)} - \eta \sum_{j \sim i} w_{ij} \cdot d_{\mathcal{V}}(\rho_{i \to ij}(\sigma_i^{(t)}),\ \rho_{j \to ij}(\sigma_j^{(t)})) \tag{1}$$

where w_{ij} is the edge weight (communication quality), $d_{\mathcal{V}}$ is the quantale-valued distance, and η is the step size. *For lattice and quantale-enriched coefficients, the actual operator is the categorical diffusion endofunctor of Ghrist–Riess [6], which computes weighted limits in a $\mathcal{V}$-category—not a scalar gradient descent. Equation* (1) *is the vector-space specialization; the convergence results below cite the correct operator for each regime.*

Theorem 3.3 (Convergence — known and conditional cases)**.** *Suppose $H^1 = 0$.*

1. ***Vector-space coefficients*** *(known [20, 21]): the iteration* (1) *converges to a global section $\{\sigma_i^*\}$ consistent on all overlaps. The convergence rate is determined by the spectral gap of the sheaf Laplacian.*

2. ***Lattice-valued coefficients*** *(known [5]): the Tarski Laplacian converges to a fixed point containing a global section, by the Tarski fixed point theorem.* Caveat (Riess): *when restriction maps are cocontinuous, the bottom element of each stalk is always a global section; convergence to this trivial section is guaranteed but uninformative. The convergence guarantee therefore says the Laplacian reaches* some *section—whether it reaches a* nontrivial *one depends on the initial data and the obstruction structure (see Remark 2.14, path 4).*
3. ***Quantale-enriched coefficients*** *(conditional on Conjecture 2.13): if the Laplacian Bridge Conjecture holds, the Lawvere Laplacian [6] converges to a nontrivial global section when one exists. For quantales satisfying the descending chain condition, convergence to* some *fixed point is guaranteed, but it may be the trivial (bottom) section. This case is* ***open****.*

The output of Phase 3 is a coordinated plan verifiable by each agent locally: agent i can check that its adjusted local state σ_i^*, restricted to each overlap, agrees with its neighbor's adjusted state.

3.4 Phase 4: Sheaf correction auction ($H^1 \neq 0$)

When the diagnostic returns NONTRIVIAL, no local adjustments can achieve consensus. The obstruction is structural—but it is not necessarily the topology alone. The No-Go Corollary [1] identifies three independent degrees of freedom for removing a nontrivial H^1 class, corresponding to three types of *sheaf correction*:

1. **Nerve correction** (change the topology). Add higher simplices—triangles, tetrahedra—that *fill* existing cycles in the nerve. Deploy a shared data source, open a joint communication channel, or share a dataset visible to three or more agents simultaneously. Example: Zoom logs create a triple overlap filling the Calendar/Email/Slack triangle (Figure 1).
2. **Coefficient correction** (change what counts as agreement). Relax the equivalence relation Equiv: replace exact match with approximate match, or weaken the assume-guarantee contracts [4]. Example: accept a fuzzy match within a tolerance threshold rather than requiring "is_meeting" be identical across systems.
3. **Restriction correction** (change how local views project onto overlaps). Redefine which local data is compared on shared overlaps. Example: instead of comparing all shared records, restrict comparison to records above a confidence threshold, removing the noisy records that cause spurious disagreements.

Remark 3.4 (H^1 monotonicity under edge addition). A subtlety constrains which nerve corrections are effective. On a graph (1-dimensional cell complex) with any cellular sheaf $\mathcal{F}$, adding edges can never decrease $\dim H^1$—it can only increase or preserve it. The reason is direct: $H^1 = C^1/\operatorname{im}(\delta^0)$, so $\dim H^1 = \dim C^1 - \operatorname{rank}(\delta^0)$. Adding an edge e with stalk $\mathcal{F}(e)$ of dimension d increases $\dim C^1$ by exactly d and $\operatorname{rank}(\delta^0)$ by *at most* d, so $\dim H^1$ cannot decrease. Topologically: new edges create new cycles, which can only add to H^1.

Therefore, effective nerve corrections for H^1 reduction are *2-cell additions*: filling an existing cycle with a triangle (or higher simplex) introduces an $\operatorname{im}(\delta^1)$ component that can kill the corresponding H^1 class. This is precisely what the Zoom-logs correction in Figure 1 achieves—it fills the Calendar–Email–Slack cycle with a 2-simplex, collapsing H^1 from $\mathbb{Z}/2\mathbb{Z}$ to 0. The auction prices 2-cell additions (joint consistency assertions), not mere edge additions (new pairwise channels).

Formal verification. The monotonicity inequality and the corollary (if $H^1(G) \neq 0$ then $H^1(G + e) \neq 0$) have been formally verified in Lean 4 (see Section 6 for scope and methodology).

SHEAF creates a market for all three correction types.

Step 1: Enumerate candidate sheaf corrections. The candidate set is generated from three sources, each bounded in size:

- **Nerve candidates:** Compute a cycle basis for the first homology $H_1(G; \mathbb{Z})$ of the underlying graph (at most $\beta_1 = |E| - |V| + c$ independent 1-cycles, where c is the number of connected components). For abelian coefficients, each such topological cycle can carry an independent H^1 obstruction: $H^1(G, \text{Equiv}) \cong \text{Equiv}^{\beta_1}$ by the universal coefficient theorem, so a H_1-cycle basis generates H^1. For each cycle, identify all potential 2-simplices that could fill it—i.e., triples of agents on the cycle that could plausibly share a joint observable. This produces at most $O(|E| \cdot \Delta)$ candidates, where Δ is the maximum degree. Edge additions alone cannot help (Remark 3.4).
- **Coefficient candidates:** For each edge carrying a nontrivial cocycle contribution, enumerate relaxations of Equiv along a lattice of contract strengths (e.g., exact $\to$ fuzzy-ϵ $\to$ type-match-only). The lattice is finite and application-specific.
- **Restriction candidates:** For each edge, enumerate restriction-map adjustments (e.g., drop fields, raise confidence thresholds) that change the cocycle data on that edge.

The candidate generator is greedy: rank candidates by marginal H^1 reduction (for nerve corrections: does filling this cycle kill a basis element?) and prune to the top K candidates within a computational budget. Finding the *globally optimal* correction set is plausibly NP-hard (it resembles minimum fill-in); in the abelian regime, the greedy approach inherits an $O(1)$-approximation from submodularity of the 2-cell matroid rank (Section 7).

Step 2: Cost assessment. Each correction has a real-world cost that depends on its type: deploying a sensor (nerve), relaxing a contractual requirement (coefficient), or re-engineering a data pipeline (restriction). Each agent privately assesses the cost of corrections it could provide.

Step 3: Sealed-bid second-price auction. Agents submit sealed bids for the sheaf corrections they can provide. Corrections of different types compete directly: a nerve edit and a coefficient relaxation that both kill H^1 are substitutes. When a single correction suffices, the mechanism is a second-price (Vickrey) reverse auction: the lowest-cost provider wins but pays the second-lowest bid, ensuring truthful bidding is a dominant strategy [12].

For combinatorial interactions (where multiple corrections interact—a nerve edit may make a coefficient relaxation unnecessary), the natural mechanism is VCG [13, 14], pricing each correction at its marginal contribution to killing the obstruction. The setting—procuring corrections with private costs under a coordination-value budget—falls within Singer's *budget-feasible mechanism design* framework [27].

For nerve corrections specifically, the relevant objective is H^1 reduction via 2-cell addition. *In the abelian coefficient regime* (where the coboundary is a linear map over

a field), adding a 2-cell σ introduces a column in the δ^1 map; the resulting H^1 reduction is submodular by a standard matroid rank argument (the columns of δ^1 form a linear matroid, and rank is submodular in the column set). Singer's budget-feasible mechanism [27] then applies, giving $O(1)$-approximation guarantees for the procurement auction. An alternative formulation targets H^0 reduction via edge addition: the function $f_0(S) = \text{rank}(\delta^0_{G+S}) - \text{rank}(\delta^0_G)$ is non-negative, monotone, and submodular by the same argument. Whether submodularity extends to non-abelian or enriched coefficient regimes remains open (Section 7).

Remark 3.5 (Approximate allocation and truthfulness). Standard VCG guarantees dominant-strategy truthfulness only when the allocation algorithm computes the *exact* optimum. Since Step 1 uses a greedy candidate generator (bounded to top-K candidates), the allocation is approximate, and exact VCG truthfulness does not hold. Two mitigations apply. First, for single-correction procurement (one correction needed to kill H^1), the mechanism reduces to a second-price reverse auction, truthful regardless of how candidates were generated—the approximation affects only *which* corrections are considered, not the winner's pricing. Second, for multi-correction procurement, Singer's budget-feasible mechanism [27] provides *truthful-in-expectation* guarantees for submodular objectives with $O(1)$-approximation, even under greedy allocation. The paper's incentive claims should be read accordingly: exact dominant-strategy truthfulness in the single-correction case; truthful-in-expectation with constant-factor welfare loss in the combinatorial case.

Remark 3.6 (Known value, private cost). SHEAF's auction has an unusual structure in mechanism design: the *value* of each correction (H^1 reduction, hence Laplacian residual reduction) is *participant-computable* from committed data (and third-party computable in privacy-light mode), while only the *provision cost* is private. This collapses to a single-parameter mechanism [28]: set reserve price equal to coordination value minus virtual cost markup; run second-price reverse auction. No winner's curse arises because value is verifiable by all participants. To our knowledge, no existing paper treats this "participant-verifiable common value + private cost" setting as a distinct paradigm; the formal treatment may be of independent interest in mechanism design.

Step 4: Execution and verification. The auction winner posts an additional bond b_{edit} and implements the claimed correction. The network re-runs the diagnostic (Algorithm 1) on the corrected sheaf. If $H^1 = 0$ after the correction, the bond is released and the protocol proceeds to Phase 3 (Resolution). If $H^1 \neq 0$ (the correction was ineffective or fraudulent), the bond is slashed and redistributed to the other agents. The verification criterion is uniform across correction types: $H^1 = 0$ on the corrected sheaf, confirmable by any participant.

3.5 Phase 5: Impossibility certificate

If no bid clears in the sheaf correction auction (no agent is willing to provide any correction—nerve, coefficient, or restriction—at an economically rational price), SHEAF issues an *impossibility certificate*: a cryptographic proof that:

1. The current sheaf has $H^1 \neq 0$ (with the specific nontrivial class identified).
2. No sheaf correction was available at a cost below the computed value of coordination.
3. Coordination is structurally impossible at any economically rational price given current agent capabilities and willingness to modify their agreements.

In privacy-light mode (transition maps posted to the bulletin board), the certificate is verifiable by any third party; in participant-only mode, it is verifiable by any protocol participant with access to committed overlap data. In either case, the H^1 computation is deterministic from the available data. Agents proceed independently. No resources are wasted on further iteration.

Remark 3.7 (The certificate as economic signal)*.* The impossibility certificate is not merely a failure mode. It is an *economic signal*: it identifies the specific sheaf deficiency—missing topology, overly rigid equivalence, or misaligned restrictions—whose correction has the highest coordination value. Entrepreneurs, platform providers, or infrastructure builders can read the certificate as a *demand signal*. The certificate decomposes the obstruction by correction type, quantifying the value of each (the Laplacian residual reduction that would be unlocked), creating a transparent market for coordination infrastructure.

Remark 3.8 (Threat model)*.* We briefly characterize what SHEAF assumes honest, what can be adversarial, and what is verifiable.

Honest registration. Each agent i truthfully reports its overlap set $U_i \cap U_j$ and computes the transition map α_{ij} faithfully from its local restriction maps. A malicious agent could fabricate an overlap (claiming shared data that does not exist) or submit a fraudulent transition map.

Detectable dishonesty. Because the cocycle condition is *locally verifiable on cycles*—each triangle (i, j, k) can be independently checked by any participant on all three edges—fabricated transition maps create inconsistencies with honest neighbors' data. The audit triangulation mechanism (Definition 4.3) exploits this: a random probe of a triple involving the suspected agent produces a cycle residual that is nonzero with probability proportional to the fraction of falsified data. The slashing bond ensures that detected fabrication costs the deviator at least the coordination value it could have captured.

Undetectable misbehavior. An agent can (a) *refuse to participate* (withhold overlap data), which reduces nerve connectivity and may prevent $H^1 = 0$ even when agreement is structurally possible—a form of strategic withholding. The impossibility certificate then correctly identifies the missing topology. (b) *Collude with neighbors*: if agents i and j jointly falsify α_{ij}, no third-party audit of the (i, j) edge alone detects the fraud. However, any cycle (i, j, k) with an honest agent k reveals the fabrication through a nonzero cycle residual. Collusion-resistant guarantees therefore hold whenever the honest subgraph of the nerve is cycle-connected. For random graphs with n agents and edge probability p, the probability that a coalition of size $c \ll n$ controls all edges on *every* cycle through a given edge is exponentially small in graph density. Dense overlap graphs are therefore more robust; this provides an additional reason (beyond H^1 reduction) to incentivize overlap creation.

Verifiable certificates. Both the cycle certificate (when $H^1 \neq 0$) and the coboundary witness (when $H^1 = 0$) are publicly verifiable from committed data. The diagnostic is deterministic: any participant can recompute it from the bulletin board.

Assumed infrastructure. The bulletin board (or commit-reveal ledger) is assumed tamper-proof: once posted, transition maps cannot be retroactively altered. This is the standard assumption for blockchain-backed commit-reveal protocols and can be relaxed with ZK proofs at additional cost (Section 7).

4 Incentive Analysis

The incentive results in this section hold under the following assumptions:

- **Quasi-linear utilities.** Each agent's payoff is the value of coordination minus costs (bond, provision, audit) minus any slash penalty. No externalities across agents beyond those mediated by the coordination outcome.
- **Participant verifiability.** The H^1 diagnostic is deterministic from the committed transition maps $\{\alpha_{ij}\}$. Any *protocol participant* (an agent on an overlap edge) can verify the cocycle data on its own edges and check the global diagnostic output. Full third-party verifiability (e.g., on-chain verification) requires that the transition maps be published to a bulletin board or verified via ZK proofs—the former is the default (privacy-light) mode; the latter is an open extension (Section 7). The "publicly computable value" of the auction refers to the value being computable by all participants from committed data, not necessarily by arbitrary external observers.
- **Private costs.** The cost of providing a sheaf correction is private to the provider. The value of the correction (H^1 reduction) is public.
- **Audit probes are exogenous.** Audit triangulation (Definition 4.3) is provided by the protocol infrastructure, not by the agents being audited. In practice this requires either a trusted coordinator, a randomized protocol-level mechanism, or a pre-committed audit schedule.
- **Privacy posture (privacy-light mode).** In the default mode, each agent publishes its transition maps $\{\alpha_{ij}\}$ (the "how my view reconciles with yours" data) to a shared bulletin board. The local state σ_i and restriction maps $\rho_{i\to ij}$ remain local; only the composed transitions $\alpha_{ij} = \rho_{j\to ij}^{-1} \circ \rho_{i\to ij}$ are revealed. This exposes structural relationships (which concepts map to which) but not raw content. A ZK extension would prove the statement "I committed to a transition map, and the resulting cocycle product around this cycle is $[\alpha]$" without revealing α_{ij} itself; the ZK statement is algebraic (group product equals a committed value) and amenable to standard SNARK constructions over $\mathbb{Z}/p\mathbb{Z}$ (Section 7).

4.1 Honest reporting

Agents may be tempted to misreport their local views or transition maps—for example, to bias the consensus toward a preferred outcome. SHEAF detects misreporting through the cocycle structure, but the detection power depends on the topology of the network—creating a circularity that must be explicitly addressed.

Proposition 4.1 (Dishonesty detection)**.** *Let agent i misreport its transition map on edge (i,j): it claims $\tilde{\alpha}_{ij} \neq \alpha_{ij}$. If i participates in any triangle (i,j,k) with an effective triple overlap, the cocycle condition $\alpha_{ij} \cdot \alpha_{jk} = \alpha_{ik}$ will be violated on the triangle involving the corrupted edge. The inconsistent edge is identifiable (up to the triangle ambiguity: the violation localizes dishonesty to one of the three edges of the failing triangle).*

Proof. If agent i reports $\tilde{\alpha}_{ij}$ but agents j and k report honestly, then on the triangle (i,j,k):

$$\tilde{\alpha}_{ij} \cdot \alpha_{jk} \neq \alpha_{ik} = \alpha_{ij} \cdot \alpha_{jk}$$

since $\tilde{\alpha}_{ij} \neq \alpha_{ij}$. The verification is performed by any agent with access to the triple overlap data. □

Remark 4.2 (The incentive circularity). Proposition 4.1 requires triangles for detection, but $H^1 \neq 0$ occurs precisely when triangles are missing—the regime where SHEAF is most needed. In the absence of effective triple overlaps, $p_{\text{detect}} \approx 0$ and the bond-slash incentive breaks down. This is a genuine circularity, not an oversight: it reflects the fact that the same topological deficiency that prevents consensus also prevents verification.

SHEAF addresses this through an *audit condition*: before the diagnostic runs, the protocol creates a sparse set of synthetic triple overlaps for verification purposes.

Definition 4.3 (Audit triangulation)**.** An *audit triangulation* of the overlap graph G is a set of *audit probes*—lightweight shared test items (synthetic records, challenge tasks, escrowed data samples) injected into selected triples of agents to create verifiable triple overlaps. The audit triangulation satisfies the *covering condition* if every edge in G participates in at least one audit triangle. The cost of the audit is the number of probes times the per-probe cost; the covering condition requires at most $O(|E|)$ probes.

Corollary 4.4 (Incentive compatibility under audit)**.** *Suppose the overlap graph is equipped with an audit triangulation satisfying the covering condition. Under a bond-slash mechanism with positive slash rate $s > 0$, honest reporting is a dominant strategy for every agent. The expected payoff from dishonesty is $v_{bias} - s \cdot b_i \cdot p_{detect}$, where $p_{detect} \geq 1 - (1-\epsilon)^{d_i}$ for an agent with d_i audit triangles and per-audit detection probability ϵ. Under the covering condition, $d_i \geq 1$ for all agents. For $b_i > v_{bias}/(s \cdot \epsilon)$, honest reporting dominates.*

The audit triangulation is *not* the same as the effective triple overlaps whose absence causes $H^1 \neq 0$. Audit probes are synthetic, lightweight, and designed for verification; they do not create the substantive shared observables needed to fill the nerve and kill H^1. The audit layer provides incentive integrity; the sheaf correction auction provides topological repair. These are complementary, not redundant.

Remark 4.5 (Settlement infrastructure). The bond-slash lifecycle (deposit $\rightarrow$ lock $\rightarrow$ release/slash) requires a settlement layer where bonds are truly at risk and slashing is automatable. The H^1 diagnostic is deterministic and verifiable, making automated slashing feasible on programmable blockchains with smart contract capabilities. The protocol is settlement-layer agnostic; the choice depends on the application's security model and throughput requirements.

4.2 Sheaf correction provision

The auction incentivizes honest provision across all correction types:

Proposition 4.6 (Correction provision incentive compatibility)**.** *The provider's bond b_{edit} is released iff the post-correction diagnostic confirms $H^1 = 0$ on the corrected sheaf. Fraud—claiming a correction that does not actually kill the obstruction—is detected by the diagnostic (the re-computed H^1 will remain nontrivial). The verification criterion is the same regardless of correction type: $H^1 = 0$. Truthful cost reporting follows from the second-price reverse auction (single-correction case) or the budget-feasible mechanism (combinatorial case; see Remark 3.5). Honest provision is incentive-compatible for any positive slash coefficient.*

4.3 Sheaf corrections as economic assets

A sheaf correction that reduces H^1 has *computable value*: the reduction in the Laplacian residual $\rho(\alpha)$ (Section 2.5)—the gap between the best approximate agreement under the

current sheaf and exact agreement under the corrected sheaf. This creates a market for *coordination infrastructure*—shared data sources, relaxed contract standards, re-engineered data pipelines:

- The *rent* on coordination infrastructure is transparent: the sheaf, the nerve, and H^1 are observable by all participants (and by third parties in privacy-light mode), so the value of any specific correction can be independently computed.
- The rent is *contestable*: corrections of different types compete. A costly nerve edit can be undercut by a cheap coefficient relaxation if both kill H^1.
- The *price discovery* is incentive-compatible: in the single-correction case, the second-price reverse auction ensures truthful bidding; in the combinatorial case, the budget-feasible mechanism provides truthful-in-expectation pricing (Remark 3.5).

4.4 Settlement infrastructure

SHEAF requires a settlement layer for bonds, slash payments, and auction clearing. The protocol is settlement-layer agnostic, but the design is informed by two principles:

1. **Enforceable liability**: the settlement asset must be one where bonds are truly at risk—not redeemable through regulatory arbitrage or platform capture.

2. **Autonomous actuation**: the settlement layer must be able to execute slashing without requiring human adjudication. The H^1 diagnostic is deterministic and verifiable, making automated slashing feasible.

The natural candidates are programmable blockchains with smart contract capabilities, where the H^1 computation can be verified on-chain and bond operations are atomic. Bitcoin with covenant extensions, Ethereum, or purpose-built settlement layers are all compatible; the choice depends on the application's security model and throughput requirements.

5 Properties and Guarantees

Theorem 5.1 (No false trivial)**.** *If the agents' local views are globally incompatible (no global consensus exists), SHEAF never outputs* Trivial*. Specifically: in the abelian and solvable regimes, SHEAF outputs* Nontrivial *(with a verifiable cycle certificate). In non-solvable regimes, SHEAF outputs either* Nontrivial *or* Inconclusive*—never* Trivial*.*

Proof sketch. By the Extension Torsor Lemma [1], global compatibility is equivalent to $H^1 = 0$. The substantive claim is that Algorithm 1 never *produces* a false coboundary witness. In the abelian/solvable case: the spanning-tree propagation (lines 3–6) constructs the *unique* candidate coboundary $\{\beta_i\}$ consistent with the tree edges—there is no choice once the root is fixed. The back-edge checks (lines 7–11) then verify whether this candidate trivializes the cocycle on *every* edge. If any back-edge residual $r_{ij} \neq 0$, the algorithm correctly reports Nontrivial. If all residuals vanish, the candidate is a genuine coboundary and the output Trivial is correct. Crucially, there is no path through Algo-

rithm 1 that outputs TRIVIAL without having verified $r_{ij} = 0$ on all non-tree edges.[2] In the non-solvable case: the spectral heuristic may fail to certify non-triviality (producing INCONCLUSIVE), but it never produces a false coboundary witness—any claimed witness is verified against all edges before the output is issued. □

Theorem 5.2 (Soundness)**.** *If SHEAF reports NONTRIVIAL (with cycle certificate), the incompatibility is genuine: no local adjustments within the current topology can achieve consensus. If SHEAF reports TRIVIAL (with coboundary witness), the witness is correct: the adjustments $\{\beta_i\}$ do achieve consistency.*

Proof sketch. The NONTRIVIAL certificate is a non-bounding cocycle, verifiable by checking the cycle product: if $\sum_i \alpha(e_i) \neq 0$ around a directed cycle, then α cannot be a coboundary (since coboundaries telescope to zero around any cycle). By the No-Go Corollary [1]: local adjustments change the cocycle by a coboundary, which cannot change a non-trivial H^1 class. The TRIVIAL witness is verified by checking $\alpha'_{ij} = 0$ for all edges after adjustment. The INCONCLUSIVE output carries no soundness claim—it signals that the heuristic could not determine the answer within the computational budget. Both the cycle-certificate soundness and the no-false-trivial property have been formally verified in Lean 4 (Section 6). □

Proposition 5.3 (Incentive compatibility — conditional)**.** *Suppose the overlap graph is equipped with an audit triangulation satisfying the covering condition (Definition 4.3). Under positive bond-slash rates and bonds exceeding the bias-to-detection ratio (Corollary 4.4), honest participation in reporting is a dominant strategy for each agent. Honest provision in the correction auction is dominant-strategy truthful in the single-correction case (second-price reverse auction) and truthful-in-expectation in the combinatorial case (budget-feasible mechanism; see Remark 3.5). Honest participation in the diagnostic phase is unconditional (the algorithm is deterministic from committed data).*

Table 1: Comparison of consensus protocols for heterogeneous agents. Superscripts indicate regime dependencies: [a] abelian/solvable only (non-solvable returns INCONCLUSIVE); [b] conditional on audit triangulation and positive bond-slash rates; [c] non-idempotent quantale required (Remark 2.16).

Property	**BFT**	**CRDT**	**MARL**	**SHEAF**
Heterogeneous vocabularies	No	No	Partial	**Yes**
Impossibility detection	No	No	No	**Yes**[a]
Impossibility certificate	No	No	No	**Yes**[a]
Architectural prescription	No	No	No	**Yes**
Incentive-compatible	Varies	N/A	No	**Yes**[b]
Graceful degradation	No	Yes	Partial	**Yes**[c]
Communication complexity	$O(n^2)$	$O(n)$	Unbounded	$O(\|E\|)$[a]

[2]This argument—that the spanning-tree propagation either finds the unique coboundary or certifies its nonexistence, and that Algorithm 1 never produces a false coboundary witness—has been formally verified in Lean 4 by Aristotle (Harmonic). The formal verification also confirmed that the spanning-tree hypothesis is load-bearing: completeness is provably false without it. See `lean/SHEAF/NoFalseTrivial.solution.lean` in the companion repository.

Table 2: SHEAF guarantee ledger by coefficient regime. Each row specifies the possible diagnostic outputs, whether each output carries a verifiable certificate, and the Phase 3 (Laplacian resolution) convergence status. Phase 3 runs only after a TRIVIAL diagnostic; its convergence is independent of whether the diagnostic returned INCONCLUSIVE on a prior run. ✓= proven, ⋆ = conditional on Conjecture 2.13.

Coefficient regime	**Outputs**	**Certificate**	**Phase 3 conv.**
Abelian ($\mathbb{Z}/p\mathbb{Z}$)	T / NT	coboundary / cycle ✓	✓
Solvable finite group	T / NT	coboundary / cycle ✓	✓
Compact non-solvable	T / NT / Inc	coboundary / cycle / none	✓
General finite non-solvable	T / NT / Inc	coboundary / cycle / none	✓
Quantale-enriched	n/a (no H^1)	n/a	⋆

Key: T = TRIVIAL, NT = NONTRIVIAL, Inc = INCONCLUSIVE. Soundness holds for T and NT (both carry verifiable certificates); Inc carries no soundness claim.

Table 2 summarizes the guarantee landscape. In the abelian and solvable regimes, the diagnostic always resolves to TRIVIAL or NONTRIVIAL, each with a verifiable certificate (coboundary witness or cycle certificate respectively). In the non-solvable regimes, the diagnostic may additionally return INCONCLUSIVE—signaling that the spectral heuristic could not resolve the question within the computational budget. Soundness holds for TRIVIAL and NONTRIVIAL outputs only; INCONCLUSIVE routes to the auction as "obstruction suspected." The quantale-enriched row is marked "n/a" because H^1 itself is undefined in the enriched setting (Remark 2.14); the Phase 3 convergence entry is conditional on the Laplacian Bridge Conjecture.

Table 3 provides a fine-grained epistemic map of the paper's claims. The separation into proved, empirical, conditional, and conjectural layers is intended to make the paper's knowledge state machine-readable: a reader can immediately distinguish what is load-bearing theorem from what is supported conjecture, and calibrate trust accordingly.

Scope of the Laplacian Bridge Conjecture dependence. The Laplacian Bridge Conjecture (Conjecture 2.13) is the single unproven item on which the enriched-coefficient convergence guarantee depends. However, the paper's *empirical* contributions do not require it. The M1–M2 extraction pipeline operates in the group regime ($O(k)$ coefficients via Procrustes), where the connection between H^1 and the connection Laplacian is a known equivalence [22, 23]—the Bridge is a theorem, not a conjecture, in this setting. The OAEI experiment operates over finite groups ($\mathbb{Z}/2\mathbb{Z}$), where Algorithm 1 is exact and Lean-verified. The Laplacian Bridge Conjecture matters only for the enriched (quantale-valued) regime that would extend SHEAF to non-invertible reconciliation—precisely the regime the paper identifies as open. A reviewer targeting the Bridge as a weakness should note that every experiment reported in this paper operates in a regime where the relevant mathematics is proven.

6 Implementation Sketch

A prototype implementation skeleton accompanies this paper in the `simulations/` directory, illustrating the protocol architecture. The implementation is in Python and is structured as follows:

Table 3: Epistemic status of SHEAF claims. Each row records a specific claim, its verification level, and the supporting evidence. "Proved (Lean 4)" means mechanically checked in Lean 4/Mathlib; "Proved (standard)" means follows from known mathematics; "Empirical (this paper)" means validated by experiments reported herein; "Conditional" depends on a stated conjecture; "Heuristic" is supported by evidence but not proven.

Claim	**Status**	**Evidence**
$H^1 = 0$ iff coboundary (abelian)	Proved (Lean 4)	12 theorems, 5 files
Alg. 1 soundness	Proved (Lean 4)	`alg1_sound`
Alg. 1 completeness	Proved (Lean 4)	`alg1_complete` (spanning hyp.)
No false trivial (Theorem 5.1)	Proved (Lean 4)	`no_false_trivial`
H^1 monotone under edge addition	Proved (Lean 4)	`betti1_nondecreasing`
Submodularity of H^1 reduction (abelian)	Proved (standard)	Matroid rank argument
Tropical bridge ($[0, \infty] \cong$ Bellman-Ford)	Proved (standard)	Proposition 2.15
Boolean lattice bridge ($2^{[k]}$, S_k)	Empirical (this paper)	Exhaustive computation, $k \leq 3$
OAEI diagnostic correctness	Empirical (this paper)	Section 6.1, 7 ontologies
Irresolvable H^1 (4-agent, $\mathbb{Z}/2\mathbb{Z}$)	Empirical (this paper)	Square graph impossibility cert.
M1-M2 Procrustes extraction (synthetic)	Empirical (this paper)	Tier 1–2 validation suite
M1-M2 gauge equivalence (8 real models)	Empirical (this paper)	Section 7, SNR $\approx 1.0\times$
Auction $O(1)$-approx (submodular)	Conditional	On Singer's mechanism
Laplacian–cohomology bridge (enriched)	Conjectural	Conjecture 2.13
Communication bottleneck (two-component)	Conj. + comp. evidence	Conjecture 7.1
Sheafable interface conditions	Definitional	Remark 2.6
H^1 for quantale-enriched coefficients	Open	Remark 2.14

- `nerve.py`: Constructs the Čech nerve from agent overlap data. Input: a set of agents and a function mapping pairs to overlap indicators. Output: a simplicial complex represented as a list of simplices. Complexity: $O(n^2)$ for n agents (pairwise overlap check); $O(n^3)$ if triple overlaps are checked.
- `cohomology.py`: Computes $H^1(N, \text{Equiv})$ for abelian coefficients via the spanning-tree algorithm (Algorithm 1). Returns the H^1 class and, if trivial, the witness $\{\beta_i\}$. For non-abelian coefficients, implements the constraint-propagation heuristic.
- `laplacian.py`: Implements the enriched sheaf Laplacian iteration (1). Placeholder for the full Ghrist–Riess diffusion; current implementation handles the Boolean and Cost quantales.
- `auction.py`: Implements the topology auction mechanism. Computes candidate corrections via greedy ranking, runs the second-price reverse auction (single-correction) or budget-feasible mechanism (combinatorial), and verifies post-edit $H^1 = 0$ on the corrected sheaf.

Integration points. The prototype is designed for integration with:

- **CrewAI / LangGraph / AutoGen**: each LLM agent wraps its local state as a restriction-map-compatible object; the SHEAF diagnostic runs as a coordination middleware.
- **ROS2**: each robot node publishes its restriction maps as ROS topics; the SHEAF diagnostic subscribes to overlap topics and publishes H^1 status.
- **Smart contract platforms**: the bond and auction logic can be implemented as on-chain contracts; the H^1 verification can be performed by an on-chain verifier or an optimistic oracle.
- **AlgebraicJulia**: the AlgebraicOptimization.jl package [33] provides sheaf Laplacian construction and distributed ADMM-based solvers over cellular sheaves, computing $H^0 = \ker L_{\mathcal{F}}$ directly.

Formal verification. The algebraic and algorithmic safety core of SHEAF has been mechanically verified in Lean 4/Mathlib (12 theorems across 5 files), using the Aristotle automated prover (Harmonic).[3] Verified claims include: the obstruction classification (Betti number identities, cycle characterization of $H^1 \neq 0$); the diagnostic algorithm (soundness and completeness of Algorithm 1); the safety guarantee (no false trivial, Theorem 5.1); cycle-certificate soundness; and H^1 monotonicity under edge addition (Remark 3.4).

During verification, Lean produced a counterexample to an earlier unconditional completeness claim for Algorithm 1: a tree decomposition that does not span all vertices admits a coboundary with nonzero back-edge residuals. This yielded a corrected theorem: completeness holds under a natural *spanning hypothesis*—the diagnostic tree must reach every vertex in the relevant component—and is *sharp* (false without the hypothesis). The spanning condition is already required by the algorithm's construction (Step 1: "choose spanning tree $T \subseteq E$"), but the formal verification identified it as a load-bearing precondition for the completeness guarantee, not merely a practical choice. The soundness direction ("all residuals zero implies coboundary") holds unconditionally.

The precise claim is: *the algebraic core of the protocol—obstruction classification, diagnostic correctness, and monotonicity—is machine-checked for finite abelian coefficients.* The mechanism design, enriched convergence, and non-abelian heuristics are not formalized (they live in game theory and analysis, not algebra). Lean source files and Aristotle solution files are available in the companion repository (`lean/SHEAF/`).

Evaluation benchmarks. Three benchmark suites provide natural ground-truth evaluation:

- **OAEI** (Ontology Alignment Evaluation Initiative) [32]: 16 heterogeneous ontologies describing the same domain, with 21 human-curated pairwise alignments. We report results on the Conference track below (Section 6.1).
- **HeMAC** [30]: heterogeneous multi-agent coordination with three agent types (quadcopters, observers, provisioners) having genuinely different sensors, observation spaces, and action spaces.
- **Valentine** [31]: 500+ dataset pairs with ground truth across unionable, joinable, and semantically-joinable scenarios—directly testing overlap topology for multi-source data integration.

[3] Aristotle (Harmonic) is an AI-powered automated theorem prover for Lean 4; see `https://harmonic.fun`. Lean source and solution files are in the companion repository (`lean/SHEAF/`).

No existing benchmark tests multi-agent LLM *concept* consensus (agents with heterogeneous conceptual vocabularies discovering overlapping concepts). Constructing such a benchmark is an explicit goal of future work.

6.1 Empirical validation: ontology alignment

The OAEI Conference track [32] provides an ideal test case: 7 ontologies (Cmt, ConfOf, Conference, Edas, Ekaw, Iasted, Sigkdd) model the same domain (conference organization) from different perspectives, with 21 human-curated pairwise reference alignments mapping classes across ontologies. Each alignment specifies equivalence correspondences—e.g., `cmt#Paper` = `confOf#Contribution`—serving exactly as SHEAF's transition maps α_{ij}.

Experiment. We construct the overlap graph (7 vertices, 21 edges) and check cocycle consistency on all $\binom{7}{3} = 35$ triples: for each triple (i, j, k) and each concept c_i that chains through all three pairwise alignments, we test whether $\alpha_{jk}(\alpha_{ij}(c_i)) = \alpha_{ik}(c_i)$. If any concept fails this test, the triple is *frustrated* (the cocycle product around the cycle is nontrivial). Consistent triples become 2-simplices in the Čech nerve; frustrated triples leave open cycles. We then compute H^1 of the resulting nerve with $\mathbb{Z}/2\mathbb{Z}$ coefficients.

Results. Of 35 triples, 32 are consistent and 3 are frustrated:

Triple	Concept	Via middle	Direct
Cmt–ConfOf–Conference	`Conference`	`Conference_volume`	`Conference`
ConfOf–Edas–Iasted	`Event`	`Conference_activity`	`Activity`
Edas–Ekaw–Iasted	`ConferenceEvent`	`Activity`	`Conference_activity`

Each violation is a concrete cycle certificate: the concept's image under the composed path $i \to j \to k$ differs from its image under the direct path $i \to k$, proving the cocycle is nontrivial on that cycle.

Despite three frustrated triples (Figure 4), SHEAF reports $H^1(\mathcal{N}, \mathbb{Z}/2\mathbb{Z}) = 0$: the nerve is dense enough (32 filled triangles out of a maximum 35) that the frustrated cycles are boundaries of higher chains through alternative consistent triples. The diagnostic is TRIVIAL—the inconsistencies are globally repairable by local relabeling.

Interpretation. This result matches the OAEI community's own experience: the "entailed reference alignment" (ra2) was obtained from the original (ra1) by computing a transitive closure and manually resolving conflicting correspondences, which were described as "relatively restricted" [32]. SHEAF automates this diagnosis. The 3 violated concepts all involve the semantic boundary between "event," "conference," and "activity"—a genuine ontological ambiguity, not a data error.

The distinction from existing alignment coherence methods [35, 36] is structural: those methods detect individual correspondences that violate conservativity or logical consistency and remove them; SHEAF diagnoses the *global topological structure* of the alignment network, certifying whether any local repair suffices ($H^1 = 0$) or whether structural change is needed ($H^1 \neq 0$). On this dataset, existing pairwise coherence

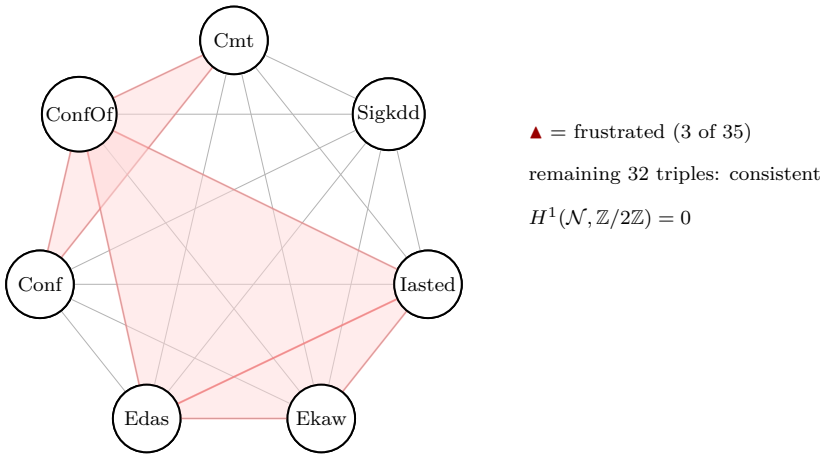


Figure 4: The OAEI Conference track nerve: 7 ontologies (complete overlap graph K_7, 21 edges). Of 35 possible triples, 32 are consistent (filled 2-simplices) and 3 are frustrated (shaded red: Cmt–ConfOf–Conference, ConfOf–Edas–Iasted, Edas–Ekaw–Iasted). Despite the frustrated triples, $H^1 = 0$: the nerve is dense enough that the frustrated cycles are boundaries of chains through alternative consistent triples. The inconsistencies are globally repairable by local relabeling.

checkers would flag the 3 frustrated triples as failures requiring manual intervention; SHEAF correctly identifies them as resolvable coboundaries (the frustrated cycles are boundaries of higher chains through the 32 consistent triangles), certifying that local relabeling suffices and no structural change is needed. The result is both more precise (it identifies *which* concepts require adjustment) and more efficient (it avoids unnecessary human review of resolvable inconsistencies). The experiment reproduces using the script `simulations/oaei_experiment.py` in the companion repository.

6.2 End-to-end protocol trace

To demonstrate the full protocol loop, we return to the Calendar/Email/Slack scenario from Section 2.4. Where Section 2.4 focused on the H^1 computation (pedagogical), this trace exercises the *operational protocol* with explicit data. The operational sequence for an $H^1 \neq 0$ scenario is: Registration $\to$ Diagnostic $\to$ Auction $\to$ Verification $\to$ Resolution (phases 1, 2, 4, 4-verify, 3 in the Section 3.1–3.5 numbering).

Registration (Section 3.1). Three agents register: Calendar (C), Email (E), Slack (S). Pairwise overlaps exist on all three pairs; no triple overlap. The nerve is $\partial\Delta^2 \cong S^1$. Each agent commits its local definition of `is_meeting` and its restriction maps.

Diagnostic (Section 3.2). Agents compute transition maps on their pairwise overlaps. Suppose the observed values are:

$$\alpha_{CE} = 1, \quad \alpha_{ES} = 0, \quad \alpha_{CS} = 0 \qquad \text{in } \mathbb{Z}/2\mathbb{Z}.$$

Calendar and Email disagree on their shared records (one agent's "yes" is the other's "no"), while the other two pairs agree. Algorithm 1 chooses spanning tree $T = \{(C, E), (E, S)\}$,

propagates $\beta_C = 0$, $\beta_E = \alpha_{CE} = 1$, $\beta_S = \beta_E + \alpha_{ES} = 1$, and checks the back-edge (C, S): residue $= \beta_S - \beta_C - \alpha_{CS} = 1 - 0 - 0 = 1 \neq 0$.

Output: NONTRIVIAL. **Certificate:** cycle $C \to E \to S \to C$ with residue 1; equivalently, $\alpha_{CE} + \alpha_{ES} + \alpha_{CS} = 1 \pmod 2$. The certificate is verifiable by any participant from the committed transition maps.

Topology auction (Section 3.4). The candidate generator identifies one nerve correction: add a 2-simplex $\{C, E, S\}$ by introducing a shared data source (e.g., Zoom meeting logs visible to all three agents) that creates a triple overlap. One provider Z bids cost $c_Z = 5$ units for deploying the Zoom integration. No other bids. The second-price reverse auction awards the correction to Z at reserve price (single bidder). Agent Z posts bond b_{edit}.

Verification (re-diagnostic). Agent Z deploys the Zoom integration. The nerve is now Δ^2 (the filled triangle). Re-running Algorithm 1: the new triple overlap imposes the cocycle condition $\alpha_{CE} + \alpha_{ES} + \alpha_{CS} = 0$, forcing a re-computation of transition maps on the enriched overlaps. With the Zoom data providing a shared ground truth, the updated transition maps satisfy $\alpha'_{CE} + \alpha'_{ES} + \alpha'_{CS} = 0$.

Output: TRIVIAL. **Certificate:** coboundary witness $\beta_C = 0$, $\beta_E = 1$, $\beta_S = 1$ (Email and Slack each flip their definition of `is_meeting`). Bond b_{edit} is released.

Resolution (Section 3.3). The Laplacian diffusion runs with the coboundary witness as initial condition. Since $H^1 = 0$, convergence is immediate: each agent applies its local adjustment β_i, and the global section (a consistent assignment of `is_meeting` across all three systems) is achieved in one round.

The trace exercises every protocol phase: registration, diagnostic with NONTRIVIAL certificate, correction auction, post-correction verification with TRIVIAL certificate, and resolution convergence. A companion script (`simulations/examples/protocol_trace.py`) reproduces the computation.

6.3 Irresolvable failure-case benchmark

To complement the resolvable protocol trace, we construct a synthetic scenario where $H^1 \neq 0$ and *no local repair suffices*—the core impossibility that motivates the topology auction.

Construction. Four agents $\{0, 1, 2, 3\}$ form a square graph (cycle C_4, no diagonals) with $\mathbb{Z}/2\mathbb{Z}$ coefficients. The cocycle assigns $\alpha_{01} = \alpha_{12} = \alpha_{23} = 0$ and $\alpha_{03} = 1$. The cycle sum is $0 + 0 + 0 + 1 = 1 \pmod 2$, certifying $H^1 \neq 0$.

Three-layer impossibility certificate. The benchmark establishes impossibility at three levels:

1. **Coboundary correction (Phase 3):** exhaustive search over all $2^4 = 16$ vertex adjustments $\{\beta_i\} \in (\mathbb{Z}/2\mathbb{Z})^4$ confirms that no coboundary correction zeroes the cocycle. The cycle sum $\sum_i \alpha_{e_i}$ is invariant under coboundary adjustment—this is the core topological obstruction that Laplacian diffusion *cannot* resolve.

2. **Edge relabeling (semantic change):** any single edge flip resolves H^1 (flipping one value toggles the cycle parity). However, this requires an agent to change the *meaning* of its shared concepts with a neighbor—a semantic cost, not a parameter adjustment. This is exactly the cost the topology auction prices.
3. **Topological repair:** edge *removal* trivializes H^1 (any edge deletion breaks the cycle, leaving a tree with $H^1 = 0$), but reduces the overlap structure. Edge *addition* never reduces graph-cohomological H^1: adding an edge increases β_1 (more independent cycles, more potential frustration). Only adding a 2-cell (filling a triangle) can kill a cycle class—but the square graph has no adjacent triples to fill. Diagonal additions with any label leave H^1 nontrivial.

The benchmark confirms that the auction is genuinely necessary: some topological obstructions cannot be resolved by local adjustment, relabeling, or edge modification alone. The companion script (`simulations/examples/failure_benchmark.py`) reproduces the full analysis.

7 Extensions and Open Problems

1. **Bridge for lattice sheaves (strong computational evidence).** For cellular sheaves with Boolean lattice stalks $2^{[k]}$ and restriction maps in S_k, computational evidence (Section 2.7) supports the following: *the Tarski operator has a fixed point strictly between $\bot$ and $\top$ if and only if $[\alpha] = 0$ in $H^1(G, S_k)$*, provided the cocycle permutation acts freely on non-extreme lattice elements. The forward direction (coboundary adjustment yields nontrivial fixed point) is immediate. The reverse direction relies on the key lemma that for a non-identity permutation σ acting freely on the atoms, $\text{meet}(x, \sigma(x)) = \bot$ for any atom x—the "destructive interference" that forces frustrated sections to collapse. Generalizing from Boolean lattices to finite distributive lattices (which cover the assume-guarantee contracts in Riess's SEAMAN framework [4]) requires understanding how the cocycle interacts with the lattice's join-irreducibles, and is the natural next step toward a full lattice Bridge theorem.

2. **Enriched H^1 without additive inverses.** The most urgent mathematical obstacle (Remark 2.14): the standard quotient $H^1 = \ker \delta^1 / \operatorname{im} \delta^0$ requires additive inverses that quantales lack. Two paths deserve investigation. First, define enriched H^1 as a *pointed set with quantale-valued metric*, where the obstruction cost $\rho(\alpha) = \inf_{b \in B^1} d_{\mathcal{V}}(\alpha, b)$ serves as the metric. This concept has no name in the literature; establishing that it satisfies the expected properties (metric axioms, stability under refinement) is an open problem. Second, detect the H^1 obstruction indirectly via the *spectral gap of a degree-1 connection Laplacian* [23], leveraging the known Hodge correspondence for vector-space sheaves [20] as a template. Resolving this obstacle would simultaneously prove the Laplacian–Cohomology Bridge Conjecture (Conjecture 2.13).

3. **Non-abelian coefficients: distributed complexity.** SHEAF's non-abelian diagnostic is an instance of group synchronization (Section 3.2). By Bulatov's CSP dichotomy theorem [25], exact synchronization over a finite group G is polynomial iff G is solvable, and NP-complete otherwise. Approximate synchronization (minimizing frustration) is polynomial for all compact groups via spectral methods [22, 23].

Open: what is the distributed round complexity of approximate group synchronization as a function of the spectral gap of the connection Laplacian? Lerman–Shi's cycle-edge message passing [24] provides a starting point.

4. **Dynamic topology and incremental H^1.** When agents join or leave the network, the nerve changes. Cohen-Steiner–Edelsbrunner–Morozov [29] maintain persistence pairings under simplex transpositions in $O(1)$ amortized time, and the stability theorem [16] guarantees bounded cohomology changes under small nerve perturbations. Open: a *distributed* incremental algorithm for H^1 updates—essential for real-time agent coordination where the network is not static.

5. **Privacy-preserving cocycle computation.** Can the H^1 diagnostic be performed without revealing the transition maps $\{\alpha_{ij}\}$ to non-participating agents? A zero-knowledge proof of H^1 triviality (or non-triviality) would allow the diagnostic to run on encrypted data. The algebraic structure of the cocycle condition (linear over $\mathbb{Z}/p\mathbb{Z}$) is amenable to standard ZK-SNARK constructions.

6. **Correction complexity and submodularity.** The H^1 monotonicity constraint (Remark 3.4) forces a precise formulation: edge addition on a 1-complex cannot reduce H^1, so effective nerve corrections must add 2-cells. For 2-cell addition *with abelian coefficients*, H^1 reduction is submodular by a matroid rank argument (the columns of δ^1 form a linear matroid over the coefficient field), and Singer's budget-feasible mechanism [27] applies directly. Whether submodularity holds for non-abelian or enriched coefficients is open. The general problem of finding minimum-cost 2-cell additions to kill H^1 is plausibly NP-hard (it resembles minimum fill-in and feedback set problems); characterizing the tractable special cases—abelian coefficients, bounded treewidth, bounded genus—is an open algorithmic question.

7. **The M1–M2 extraction problem: empirical results.** The most significant practical gap for deploying SHEAF is computing group-valued transition maps α_{ij} from real agent outputs. We address this with a concrete Procrustes-based extraction pipeline and report the first empirical test of cocycle triviality across real embedding models.

   The gap. SHEAF's diagnostic requires each pair of communicating agents to produce a transition map $\alpha_{ij} \in \mathrm{Equiv}(U_i \cap U_j)$ on their shared overlap. For agents with structured output schemas (database queries, typed API responses, formal ontologies), the group structure is manifest in schema isomorphisms—the OAEI experiment (Section 6.1) demonstrates this regime. For agents with unstructured outputs (embeddings, free-text), the natural group is $O(k)$ (orthogonal transformations of the embedding space), and the transition maps are Procrustes alignments.

   Pipeline. Given n embedding models and a shared anchor corpus of N sentences: (i) embed all N anchors in each model, producing matrices $X_i \in \mathbb{R}^{N\times d}$; (ii) mean-center each X_i independently, then compute a **per-model PCA** projection to dimension $k \ll d$—each model retains its own top-k principal subspace (mandatory: orthogonal Procrustes in $\mathbb{R}^d$ with $N < d$ anchors is underdetermined; the per-model choice preserves each model's natural geometry rather than imposing a shared basis); (iii) for each pair (i, j), solve orthogonal Procrustes: $R_{ij} = \arg\min_{R\in O(k)} \|X_i R - X_j\|_F$ via SVD of $X_i^\top X_j$; (iv) build the connection Laplacian $L \in \mathbb{R}^{nk\times nk}$ [22] from

the $O(k)$-valued cocycle $\{R_{ij}\}$ on the complete graph of n models. The synchronization residual is the k-th smallest eigenvalue $\lambda_{k-1}(L)$: zero iff the cocycle is a coboundary (i.e., a global gauge exists), positive iff frustrated. The primary metric is the **frustration SNR**: $\lambda_{k-1}(L_{\text{real}})/\lambda_{k-1}(L_{\text{null}})$, where L_{null} is built from a noise-matched null (anchor correspondences randomly permuted per-model to destroy real structure while preserving marginal embedding statistics).

Synthetic validation. The pipeline was validated on controlled synthetic data (Tier 1: direct cocycle tests; Tier 2: full pipeline with non-isometric embeddings). Key findings: (a) the connection Laplacian correctly distinguishes coboundary cocycles (spectral gap $< 10^{-14}$) from frustrated cocycles (spectral gap monotonically increasing with defect angle θ, from 3×10^{-4} at $\theta = 0.05$ to 0.76 at $\theta = \pi$); (b) the diagnostic is robust across PCA dimensions ($k = 32$ to 256) and anchor ratios ($n/k = 2$ to 10); (c) noise degrades separation—at SNR $< 20\,\text{dB}$, noise-induced frustration dominates real signal, confirming the necessity of a noise-matched null.

Real embedding experiment. We tested 8 sentence-transformer models (all 768-dimensional): three MPNet variants (STS, paraphrase, QA) as within-family controls; DistilRoBERTa and DistilBERT-MSMARCO as cross-architecture treatments; BGE (BAAI), E5 (Intfloat), and Nomic as cross-organization treatments. The anchor corpus comprised 5,000 template-generated English sentences formed from combinatorial patterns over 20 subjects, 15 verbs, 15 objects, and 15 modifiers (deterministic seed for reproducibility; the script is included in the companion repository). While this ensures exact replication, the semantic range is narrower than a natural-language corpus; extending to diverse corpora (e.g., STS Benchmark, Wikipedia) is a natural robustness check. For each PCA dimension $k \in \{64, 128, 256, 384\}$, we computed the cocycle, the spectral gap (frustration), and a noise-matched null (shuffled anchor correspondences destroying real structure while preserving marginal statistics).

Result: gauge equivalence (the Platonic outcome). Across all 56 triples ($\binom{8}{3}$) and all PCA dimensions, the frustration gap was **indistinguishable from noise** (SNR $\approx 1.0\times$):

k	Frust. gap	Null gap	SNR	Verdict
64	3.550	3.592	0.99×	trivial
128	3.569	3.609	0.99×	trivial
256	3.583	3.591	1.00×	trivial
384	3.585	3.602	1.00×	trivial

Interpretation: instrument calibration. The connection Laplacian diagnostic tests a structural property—global gauge equivalence ($H^1 = 0$)—that is strictly stronger than pairwise similarity. Existing methods for comparing embedding models (linear CKA, centered kernel alignment, representation similarity analysis [41]) are H^0-type measurements: they test whether model A's representation is similar to model B's, one pair at a time. Our diagnostic is an H^1-type measurement: it tests whether the pairwise alignments **compose transitively**—whether a single global gauge transformation can bring all models into a common frame simultaneously. Models could align pairwise yet fail to compose ($H^1 \neq 0$), a structural failure invisible to all existing metrics. On current SOTA models, the diagnostic reports $H^1 = 0$: pairwise

Procrustes methods are sufficient, and anyone chaining these models (retrieval → reranking → classification) is not vulnerable to silent compositional inconsistency in this regime. The Platonic Representation Hypothesis [41] predicts this outcome; the diagnostic provides the first direct structural test rather than pairwise correlational evidence. Viewed as a controlled experiment, this result isolates the *communication channel* as the failure locus: since internal representations are gauge-equivalent, any frustration observed in deployed multi-agent systems is introduced by the interface, not inherited from the geometry.

Pairwise-vs-global decoupling. Procrustes residuals reveal a 64× range: closely related models (DistilRoBERTa–MPNet-STS: 12.8; BGE–E5: 13.4) align nearly perfectly, while architecturally distant pairs (DistilBERT-MS–Nomic: 229; E5–Nomic: 820) have large residuals. Despite this range, no triple shows cycle frustration above noise. The non-orthogonal components cancel around every cycle rather than accumulating. This decoupling between pairwise residual magnitude and cycle frustration illustrates the diagnostic's value: it distinguishes "large but composable" misalignment (noise, $H^1 = 0$) from "structurally frustrated" misalignment ($H^1 \neq 0$) that pairwise methods would miss entirely.

Where the diagnostic would fire. The gauge-equivalence finding is specific to the tested regime: same-dimensional models on a common-domain corpus. Frustration is more likely to emerge in: (a) cross-dimensional models (where Procrustes is replaced by lossy CCA/zero-padding, placing the problem in the monoid regime); (b) domain-shifted corpora (where models trained on different domains embed the same sentence in structurally incompatible ways); (c) the structured-output regime (LLM-generated JSON schemas), where field permutations form a discrete group and non-composability may arise from prompt-dependent schema choices. These regimes—where the diagnostic is most likely to report $H^1 \neq 0$ and trigger the topology auction—are the natural next deployment targets.

Two sheaves on the same nerve: representation vs. communication. The result above concerns the **representation sheaf**: stalks are internal embedding spaces V_i, restriction maps are Procrustes rotations $R_{ij} \in O(k)$, and $H^1 = 0$ means the internal geometries are globally compatible. But LLM agents do not communicate via embeddings. They communicate via *strings*—a discrete channel C of capacity $B = \log_2 |C|^L$ bits for sequences of length L over vocabulary C. The **communication sheaf** has the same stalks V_i but different restriction maps: $\alpha_{ij} = d_j \circ e_i \colon V_i \to V_j$, where $e_i \colon V_i \to C$ is agent i's encoding (generation) and $d_j \colon C \to V_j$ is agent j's decoding (interpretation). These maps are non-invertible (information is destroyed at every hop), stochastic (at temperature > 0), and context-dependent (varying with prompt and conversation history). They are not group-valued; they are kernel-valued (Markov kernels) at best, placing the communication sheaf squarely in the enriched regime where the natural cost structure is a Lawvere metric (e.g., KL divergence between interpretation distributions)—precisely the quantale-enriched framework of Section 2.5.

The representation-sheaf result ($H^1 = 0$) therefore serves as a **baseline**: it establishes that the agents' internal geometries are compatible, so any coordination failures observed in string-mediated systems are introduced by the communication channel, not inherited from the representation layer. $H^1 = 0$ in the representation sheaf does *not* imply $H^1 = 0$ (or even well-definedness of H^1) in the communi-

cation sheaf. The gap between these two sheaves—clean internal geometry, lossy external channel—is the precise locus of the multi-agent coordination problem, and closing it requires either bypassing language (embedding-mediated coordination) or imposing enough type structure at the interface to stabilize the transition maps (see Remark 2.6).

One further avenue is game-theoretic: pairs of agents could play bilateral negotiation games—using the fixed-point structures of Shapley operators [39]—to agree on a common level of abstraction before SHEAF runs the global diagnostic, effectively engineering invertibility where it does not arise naturally.

Conjecture 7.1 (Communication Bottleneck — Refined)**.** *Let $n \geq 3$ agents have d-dimensional internal representations related by gauge transforms $\{g_{ij}\} \subset O(d)$ with angular magnitude $\theta = \max_{ij} \|\log g_{ij}\|$. Let each agent communicate through a B-dimensional linear channel via encoder $E_i \colon \mathbb{R}^d \to \mathbb{R}^B$. Let $\pi \colon O(d) \to O(B)$ denote the Procrustes projection (nearest orthogonal factor in the B-dimensional image). The communication-sheaf cocycle decomposes into two independent frustration components:*

(a) ***Encoder alignment.*** *For agents with* aligned *encoders ($E_i \approx E_j$), the communication cocycle preserves the coboundary structure of the representation sheaf when $\theta < \theta_c(B, d)$: the spectral gap satisfies $\lambda_{B-1}(L_{\text{comm}}) \leq C \cdot \theta^2$, where C depends on B/d. This follows from the fact that π is smooth near the identity and its differential $d\pi_I \colon \mathfrak{o}(d) \to \mathfrak{o}(B)$ is the natural restriction of the Lie algebra—a linear (hence homomorphic) map. The nonlinearity that destroys cocycle structure enters at second order in θ.*

(b) ***Encoder mismatch (Haar universality).*** *For agents with* misaligned *encoders (independently drawn E_i), the Procrustes-projected communication cocycle is* Haar-distributed on $O(B)$*: each transition $R_{ij} = UV^\top$ from $\text{SVD}(E_i E_j^\top)$ is Haar-random by bi-invariance of the encoder distribution, and edge correlations from shared vertices contribute a correction of $O(B/d)$. Consequently, the expected spectral gap satisfies $\mathbb{E}[\lambda_{B-1}(L_{\text{comm}})] = \text{Haar}(B, n) \cdot (1 - O(B/d))$, where $\text{Haar}(B, n) > 0$ is the expected spectral gap of a Haar-random $O(B)$ cocycle on K_n, with $\text{Haar}(B, n) \to 1$ as $B \to \infty$. The gap depends on B alone, not on d/B: the channel does not partially degrade the gauge structure but* completely replaces it with random noise*, for any $B < d$. No choice of gauge transforms can reduce frustration below $\text{Haar}(B, n) \cdot (1 - O(B/d))$ in expectation. Encoder differences dominate gauge magnitude: even agents with identical internal representations ($\theta = 0$) produce full frustration when their encoders differ.*

(c) ***Type structure reduces effective mismatch.*** *The sheafability conditions (Remark 2.6) operate by reducing effective encoder mismatch—standardizing the encoding format makes the agents' E_i more similar, moving from regime (b) toward regime (a), where gauge magnitude determines frustration and the representation-sheaf baseline ($H^1 = 0$) is recoverable.*

(d) ***Shared-subspace interpolation.*** *If agents share k_{shared} of their B encoding dimensions and differ on the remaining $B - k_{\text{shared}}$, the frustration on the shared subspace is $O(\theta^2)$ (coboundary in the aligned regime), while the full-space frustration interpolates monotonically between regimes (a) and (b) as k_{shared}/B varies from 1 to 0. This provides the quantitative mechanism for item (c): each sheafability condition (fixed schema, deterministic decoding, bounded coercions) increases the effective k_{shared}, monotonically reducing frustration on the task-relevant dimensions.*

Computational evidence. *A linear simulation suite (`simulations/extraction/`) validates all components: (i) for shared encoders, the spectral gap spans six orders of magnitude from $\theta = 0.008$ rad (gap $< 10^{-6}$, coboundary) to $\theta = \pi$ (gap ≈ 1, random), with a transition near $\theta_c \approx 0.4$ rad; the frustration constant $\lambda/\theta^2 \approx 0.021$ is stable across three orders of magnitude ($\theta = 0.001$ to $\theta = 1.0$), confirming sharpness of the $O(\theta^2)$ bound; (ii) for independent encoders, the gap matches the Haar-random $O(B)$ baseline to within 1–5% across all tested (d, B) pairs with $d/B \geq 2$; fixing B and varying d from $2B$ to $32B$ changes the gap by $< 2\%$, confirming that the gap depends on B alone; each individual Procrustes rotation passes a Kolmogorov-Smirnov test against the Haar distribution (p-values > 0.05); (iii) subspace general position (kernel dimensions, pairwise and triple intersections) matches dimension-counting predictions exactly; (iv) the shared-subspace cocycle is always coboundary (gap $< 10^{-9}$) while the full cocycle frustration decreases monotonically from 0.96 ($k_{\text{shared}} = 0$) to 0.00 ($k_{\text{shared}} = B$), confirming the interpolation. The representation-sheaf cocycle has gap $< 10^{-15}$ in all cases, confirming the gauge equivalence baseline.*

The Laplacian Bridge Conjecture (Conjecture 2.13) is SHEAF's central *mathematical* open problem: it asks whether the enriched Laplacian detects obstructions in non-group coefficient regimes. The Communication Bottleneck Conjecture is the central *applied* open problem: it characterizes when a lossy channel destroys the gauge structure that the representation sheaf preserves. The two conjectures address complementary regimes: the Bridge extends the diagnostic to enriched coefficients; the Bottleneck identifies when enriched coefficients are *needed* (when encoder mismatch or large gauge angles place the communication sheaf outside the group regime). Together, they would yield a complete obstruction theory for string-mediated multi-agent coordination.

The refined Bottleneck also connects the Platonic embedding result to the engineering prescription. The empirical finding that current SOTA models have gauge angles $\theta < 0.08$ rad (Section 7, item 7) places them deep in regime (a), where a shared or aligned channel preserves coordination. The practical failure mode is therefore not the gauge magnitude but the *encoder mismatch* between agents with different tokenizers, architectures, or training distributions. This is precisely what the sheafability conditions (Remark 2.6) are designed to control: fixed schemas reduce encoding variance, deterministic decoding stabilizes transition maps, and bounded coercions keep the effective encoder distance small enough for the linearization in (a) to hold.

8 Discussion: When Does the Diagnostic Apply?

The SHEAF diagnostic is validated on static ontology networks (Section 6.1) and calibrated against real embedding models (Section 7, item 7). A natural question is whether the H^1 obstruction arises in current LLM-based multi-agent systems. We argue that it does not—for structural reasons that are themselves informative—and that the engineering trajectory of the agent ecosystem is moving toward the regime where it will.

Why current LLM agent systems do not exhibit diagnosable H^1. Current multi-agent LLM coordination occupies two regimes, neither of which produces the algebraic structure that H^1 requires. In the *unstructured regime*—agents communicating via free-form natural language—there is no stable transition map between agents' interpretations.

The "reconciliation" between GPT's and Claude's understanding of a message is a stochastic function of the prompt, the conversation history, and the decoding temperature; it is not a group element, and there is no algebraic object over which to compute cohomology. In the *fully structured regime*—agents communicating via rigid typed schemas (database queries, fixed API calls)—the transition maps are explicit schema isomorphisms, and the consistency check reduces to a database join. The diagnostic is well-defined but unnecessary: pairwise consistency checks already detect all failures, because the schemas are rigid enough that composition is trivially verifiable.

The framework's domain of applicability is the intermediate regime: agents with semi-structured communication, stable enough to define algebraic transition maps but flexible enough that the maps don't trivially compose. The OAEI experiment (Section 6.1) validates this regime for static ontology alignment, where the transition maps are human-curated but the cycle structure is nontrivial. The Procrustes experiment (Section 7, item 7) tests the unstructured regime and correctly reports $H^1 = 0$—not because coordination is perfect, but because the "transition maps" are too noisy to carry topological information. These are the right results for the right reasons: the diagnostic is silent when the algebraic preconditions are not met, and informative when they are.

Why the engineering trajectory converges on the diagnosable regime. Three concurrent developments are moving LLM agent systems from the unstructured regime toward the structured-but-nontrivial intermediate regime. First, *typed communication protocols*—Google's Agent-to-Agent (A2A) protocol, Anthropic's Model Context Protocol (MCP), and OpenAI's function-calling schemas—constrain agent outputs to typed, schema-validated structures. Each typed field in an agent's output is a section of a sheaf over a typed stalk; the transition maps between agents' structured outputs are schema morphisms, which are well-defined algebraic objects. Second, *multi-agent orchestration frameworks* are evolving from tree and DAG topologies (orchestrator $\rightarrow$ workers, where $H^1 = 0$ trivially because trees are contractible) toward peer-to-peer coordination meshes (the A2A vision), introducing cycles in the coordination graph. Third, *heterogeneous agent ecosystems*—where agents from different providers with different training data serve different roles—are becoming the norm rather than the exception, creating the encoder mismatch that the Bottleneck Conjecture (Conjecture 7.1) identifies as the source of structural frustration.

Typed interfaces + cyclic coordination topology + heterogeneous agents = the regime where pairwise consistency checks succeed but global coordination can fail: precisely the regime where H^1 captures failures invisible to existing diagnostics.

The sheafability prescription is predictive. The sheafable-interfaces conditions (Remark 2.6) are not a post-hoc explanation of observed failures. They are a *predictive* specification: any multi-agent system whose communication interfaces satisfy conditions (i)–(iv) is guaranteed to have well-defined transition maps, computable H^1, and—via the submodularity result (Section 7, item 6)—an efficiently approximable minimum repair. This makes the prescription actionable before failures manifest: a system architect can verify sheafability at design time and know that the SHEAF diagnostic will be available at runtime, rather than discovering after deployment that coordination failures are undiagnosable.

The OAEI experiment provides the template: static ontology networks with curated alignments are the simplest instance of the diagnosable regime. Enterprise data integra-

tion networks—multiple databases with pairwise ETL mappings forming cyclic dependency graphs—are the natural next deployment target, where $H^1 \neq 0$ would indicate integration inconsistencies invisible to pairwise coherence checkers [35, 36]. Dynamic LLM agent coordination will enter the diagnosable regime as typed communication standards mature and multi-agent topologies develop cycles. The mathematical framework is ready; the infrastructure is converging.

The string-table seam: where H^1 becomes nontrivial. The companion paper [2] identifies a concrete regime where the H^1 obstruction is already operative: the *string-table seam*, where LLM agents translate natural language into typed, schema-validated database operations. Three production LLMs operating against three database schemas produce bilaterally-invisible cycle failures—every pairwise check passes, but the composition around the coordination cycle does not close—in the nontrivial-H^1 regime. The minimum number of typed 2-cells ("bridge concepts") required to restore cycle closure equals $\dim H^1$ of the interpretation sheaf on the coordination graph: the *coherence fee*. Meanwhile, the representation-sheaf baseline reported in Section 7 confirms $H^1 = 0$ for the same models' internal embeddings. Together, these results locate the first computably nontrivial obstruction at the structured-output interface, not in the internal geometry—precisely the intermediate regime where SHEAF's diagnostic and auction mechanisms become operative.

References

[1] J. Komkov, *Predicate Invention Under Sheaf Constraints: Mathematical Foundations for Compositional Discovery*, companion paper, 2026. Available in the project repository.

[2] J. Komkov, *The Coherence Fee: Edge-Local Blindness at the String-Table Seam*, companion paper, 2026.

[3] J. Komkov, *Res Agentica: The Political Economy of Machine Testimony*, companion manuscript, 2026.

[4] H. Riess, *SEAMAN: Sheaf-Theoretic Semantic Alignment for Multi-Agent Networks*, Georgia Institute of Technology, 2026.

[5] R. Ghrist and H. Riess, *Cellular sheaves of lattices and the Tarski Laplacian*, Homology, Homotopy and Applications **24**(1), 2022.

[6] R. Ghrist, H. Lopez, B. North, and H. Riess, *Categorical diffusion on cellular sheaves*, arXiv:2501.03890, 2026.

[7] L. Lamport, *The part-time parliament*, ACM Trans. Computer Systems **16**(2):133–169, 1998.

[8] D. Ongaro and J. Ousterhout, *In search of an understandable consensus algorithm*, USENIX ATC, 2014.

[9] M. Castro and B. Liskov, *Practical Byzantine fault tolerance*, OSDI, 1999.

[10] S. Nakamoto, *Bitcoin: A peer-to-peer electronic cash system*, 2008.

[11] M. Shapiro, N. Preguiça, C. Baquero, and M. Zawirski, *Conflict-free replicated data types*, SSS, 2011.

[12] W. Vickrey, *Counterspeculation, auctions, and competitive sealed tenders*, J. Finance **16**(1):8–37, 1961.

[13] E. H. Clarke, *Multipart pricing of public goods*, Public Choice **11**:17–33, 1971.

[14] T. Groves, *Incentives in teams*, Econometrica **41**(4):617–631, 1973.

[15] F. W. Lawvere, *Metric spaces, generalized logic, and closed categories*, Rendiconti del Seminario Matematico e Fisico di Milano **43**:135–166, 1973.

[16] H. Edelsbrunner and J. L. Harer, *Computational Topology: An Introduction*, AMS, 2010.

[17] CrewAI, `https://www.crewai.com/`, 2024.

[18] LangChain, *LangGraph*, `https://python.langchain.com/docs/langgraph`, 2024.

[19] Microsoft, *AutoGen: Enabling next-gen LLM applications via multi-agent conversation*, 2023.

[20] J. Hansen and R. Ghrist, *Toward a spectral theory of cellular sheaves*, J. Appl. Comput. Topol. **3**(4):315–358, 2019.

[21] J. Hansen and R. Ghrist, *Opinion dynamics on discourse sheaves*, SIAM J. Appl. Math. **81**(5):2076–2100, 2021.

[22] A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Appl. Comput. Harmon. Anal. **30**(1):20–36, 2011.

[23] A. S. Bandeira, A. Singer, and D. A. Spielman, *A Cheeger inequality for the graph connection Laplacian*, SIAM J. Matrix Anal. Appl. **34**(4):1611–1630, 2013.

[24] G. Lerman and Y. Shi, *Robust group synchronization via cycle-edge message passing*, Found. Comput. Math. **22**:1665–1741, 2022.

[25] A. A. Bulatov, *A dichotomy theorem for nonuniform CSPs*, FOCS, 2017. J. ACM **67**(5), 2020.

[26] R. G. Gallager, *Low-density parity-check codes*, IRE Trans. Inform. Theory **8**(1):21–28, 1962.

[27] Y. Singer, *Budget feasible mechanisms*, FOCS, 2010.

[28] R. B. Myerson, *Optimal auction design*, Math. Oper. Res. **6**(1):58–73, 1981.

[29] D. Cohen-Steiner, H. Edelsbrunner, and D. Morozov, *Vines and vineyards by updating persistence in linear time*, SoCG, 2006.

[30] M. Dansereau et al., *HeMAC: Heterogeneous multi-agent coordination benchmark*, ECAI, 2025.

[31] C. Koutras et al., *Valentine: Evaluating matching techniques for dataset discovery*, IEEE ICDE, 2021.

[32] Ontology Alignment Evaluation Initiative, `http://oaei.ontologymatching.org/`, annual since 2004.

[33] AlgebraicJulia, *AlgebraicOptimization.jl*, `https://github.com/AlgebraicJulia/AlgebraicOptimization.jl`, 2024.

[34] C. Kurisummoottil Thomas and M. Chen, *Fundamental limits of quantum semantic communication via sheaf cohomology*, arXiv:2601.10958, 2026.

[35] C. Meilicke and H. Stuckenschmidt, *An efficient method for computing alignment diagnoses*, Proc. 3rd International Conference on Web Reasoning and Rule Systems (RR), LNCS 5837, pp. 182–196, 2009.

[36] A. Solimando, E. Jiménez-Ruiz, and G. Guerrini, *Detecting and correcting conservativity principle violations in ontology-to-ontology mappings*, ISWC, pp. 1–16, 2014.

[37] R. Bellman, *On a routing problem*, Quarterly of Applied Mathematics **16**(1):87–90, 1958.

[38] L. R. Ford, *Network flow theory*, RAND Corporation Report P-923, 1956.

[39] M. Akian, S. Gaubert, and S. Vannucci, *Ambitropical geometry, hyperconvexity and zero-sum games*, arXiv:2108.07748, 2021.

[40] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, *Word translation without parallel data*, ICLR, 2018.

[41] M. Huh, B. Cheung, T. Wang, and P. Isola, *The Platonic Representation Hypothesis*, ICML, 2024.

Acknowledgments

The formal verification of key theorems (H^1 monotonicity, the "no false trivial" property, cycle-certificate soundness, algorithm correctness, and the obstruction classification) was performed by Aristotle (Harmonic); the Algorithm 1 completeness verification identified a missing spanning hypothesis that sharpened the theorem statement. The authors thank Hans Riess additionally for discussions on sheaf Laplacians, assume-guarantee contracts, the enriched cohomology obstacle, and specifically for the observations on cocontinuous restriction maps, the descending chain condition, and the trivial-section convergence caveat that sharpened the Laplacian Bridge Conjecture statement and convergence theorem.

The Linear Communication Bottleneck Theorem: Spectral Frustration of Procrustes Cocycles on Random Encoder Graphs

Abstract

When n agents encode d-dimensional representations into B-dimensional messages ($B < d$) and align them via the orthogonal Procrustes map, the resulting cocycle on the complete communication graph K_n determines a connection Laplacian whose spectral gap measures the fundamental limit of bandwidth-limited coordination. We prove a trichotomy. In the aligned regime, frustration scales as $O(\theta^2)$ where θ is the gauge magnitude. In the misaligned regime (independent Haar-random encoders on $\mathrm{St}(B, d)$), each Procrustes rotation is Haar-distributed on $O(B)$ and adjacent edges are exactly pairwise independent; the expected spectral gap matches the Haar-random $O(B)$ baseline to within $O(B/d)$. The $O(B/d)$ correction is localized to triangle holonomy arising from non-composability of Procrustes projection through rank-B subspaces. An intermediate shared-subspace regime interpolates between the extremes.

1 Introduction

Consider n agents, each maintaining a d-dimensional internal representation of a shared environment. To coordinate, each agent compresses its representation into a B-dimensional message sent through a linear channel to every other agent. The receiving agent aligns the incoming message with its own representation using the orthogonal Procrustes map—the closest rotation in $O(B)$. This setup arises in distributed optimization, federated learning, multi-sensor fusion, and multi-agent systems where heterogeneous representations must be reconciled through bandwidth-limited communication.

The central question is: when do the agents' Procrustes alignments compose consistently around the communication graph? If agent i aligns with agent j and agent j aligns with agent k, does the composed alignment agree with the direct alignment between i and k? If so, global coordination is achievable from pairwise message-passing. If not, there is a *communication bottleneck*—a structural frustration that no amount of pairwise channel improvement can resolve.

The connection Laplacian on a graph with $O(d)$-valued edge weights was introduced by Singer [4] for angular synchronization and studied by Bandeira, Singer, and Spielman [1], who proved a Cheeger inequality relating its spectral gap to the frustration of the cocycle. These tools have been applied to cryo-EM reconstruction, sensor network calibration, and ranking. The existing theory assumes edge rotations are either adversarial or drawn i.i.d. from the Haar measure on $O(d)$. Neither assumption captures the structured dependence arising from Procrustes alignment of random encoders, where edge rotations share vertices and create nontrivial correlations.

We prove a *Linear Communication Bottleneck Theorem* that characterizes the spectral gap of the Procrustes cocycle on K_n. The result is a trichotomy indexed by encoder alignment:

1. **Aligned regime.** Encoders share a common B-dimensional subspace; frustration is $O(\theta^2)$.

2. **Misaligned regime.** Independent Haar-random encoders produce a cocycle whose spectral gap matches the Haar baseline to within $O(B/d)$.

3. **Intermediate regime.** A shared-subspace decomposition interpolates continuously.

The key technical ingredient is a *Haar-universality lemma* (Lemma 6) showing that the Procrustes polar factor of random encoder overlaps is Haar-distributed by a bi-invariance argument, with adjacent edges exactly pairwise independent. The $O(B/d)$ correction requires two derived lemmas: a Grassmannian concentration bound (Lemma 7) and a triangle holonomy bound (Lemma 8).

The theorem identifies a sharp phase transition: below a critical subspace-sharing threshold, the communication channel completely randomizes gauge structure regardless of d. Reducing this frustration requires structural intervention—shared representation subspaces—rather than merely increasing channel capacity.

2 Preliminaries

Definition 1. *The* Stiefel manifold $\mathrm{St}(B, d)$ *is the set of $B \times d$ matrices with orthonormal rows:* $\mathrm{St}(B, d) = \{E \in \mathbb{R}^{B\times d} : EE^\top = I_B\}$. *It carries a unique $O(B)$-bi-invariant probability measure (the Haar measure inherited from $O(d)$).*

Definition 2. *For $E_i, E_j \in \mathrm{St}(B, d)$, the* Procrustes rotation *is $R_{ij} = UV^\top$ where $U\Sigma V^\top = \mathrm{SVD}(E_iE_j^\top)$. This is the closest element of $O(B)$ to $E_iE_j^\top$ in Frobenius norm.*

Definition 3. *The* connection Laplacian *on a graph $G = (V, E)$ with $O(B)$-valued cocycle $\{R_{ij}\}_{(i,j)\in E}$ is the $nB \times nB$ block matrix $L = D - A_\rho$, where $(A_\rho)_{ij} = w_{ij}R_{ij}$ for $(i, j) \in E$ and D is the block-diagonal degree matrix. The* normalized *connection Laplacian is $L_1 = I - D^{-1/2}A_\rho D^{-1/2}$.*

Definition 4. *The* frustration constant *of a cocycle $\{R_{ij}\}$ on G is $\eta_G = \min_{g_1,\ldots,g_n\in O(B)} \frac{1}{2|E|}\sum_{(i,j)\in E} w_{ij}\|g_iR_{ij} - g_j\|_F^2$. It measures how far the cocycle is from being a coboundary.*

By the Cheeger inequality for connection Laplacians [1]:

$$\frac{\lambda_1(L_1)}{2} \le \eta_G \le \sqrt{2\lambda_1(L_1)}.$$

3 Main Result

Theorem 5 (Linear Communication Bottleneck)**.** *Let $n \ge 3$ agents have d-dimensional representations with gauge transforms of angular magnitude θ, communicating through B-dimensional linear channels with $d \ge 2B$.*

(a) Aligned encoders, small gauge. *If the encoders share a common B-dimensional subspace up to gauge transforms of magnitude θ, then $\lambda_{B-1}(L_{\mathrm{comm}}) \le C(B/d) \cdot \theta^2$.*

(b) Misaligned encoders. *If $E_1, \ldots, E_n$ are independent Haar-random elements of $\mathrm{St}(B, d)$, then $\mathbb{E}[\lambda_{B-1}(L_{\mathrm{comm}})] = \mathrm{Haar}(B, n) \cdot (1 - O(B/d))$, where $\mathrm{Haar}(B, n) > 0$ is the expected spectral gap of a Haar-random $O(B)$ cocycle on K_n.*

(c) Shared subspace. *For encoders sharing a k-dimensional subspace ($0 \le k \le B$), the frustration interpolates between regimes (a) and (b) as k/B varies from 1 to 0.*

The proof of part (b) rests on the following:

Lemma 6 (Haar Universality)**.** *Let $E_1, \dots, E_n$ be independent Haar-random elements of $\mathrm{St}(B, d)$ with $d \geq 2B$. For each pair (i, j), let $R_{ij} = UV^\top$ where $U\Sigma V^\top = \mathrm{SVD}(E_i E_j^\top)$. Then:*

(a) *Each R_{ij} is Haar-distributed on $O(B)$.*

(b) *(R_{ij}, R_{ik}) are independent Haar on $O(B)$ for all distinct $j, k \neq i$.*

(c) *$\mathbb{E}[\lambda_{B-1}(L)] = \mathrm{Haar}(B, n) \cdot (1 - \varepsilon(B, d))$ with $\varepsilon(B, d) \leq C'B/d$.*

4 Proof of the Haar-Universality Lemma

4.1 Part (a): Bi-invariance

Proof. Let E_1, E_2 be independent Haar-random elements of $\mathrm{St}(B, d)$. The matrix $M = E_1 E_2^\top \in \mathbb{R}^{B \times B}$ is left-$O(B)$-invariant: for any $Q \in O(B)$, $\mathcal{L}(QM) = \mathcal{L}(M)$ because QE_1 is Haar on $\mathrm{St}(B, d)$. By the same argument applied to E_2, M is right-$O(B)$-invariant. Let $M = U\Sigma V^\top$. Replacing M by QM sends $R = UV^\top \mapsto QR$; replacing M by $MQ^\top$ sends $R \mapsto RQ^\top$. Since $\mathcal{L}(QM) = \mathcal{L}(M)$ and $\mathcal{L}(MQ^\top) = \mathcal{L}(M)$, the distribution of R is bi-$O(B)$-invariant. By uniqueness of the Haar measure on $O(B)$, R is Haar. □

4.2 Part (b): Pairwise independence

Proof. Fix vertex i and condition on E_i. Given E_i, the encoders E_j and E_k remain independent.

Claim: R_{ij} is Haar on $O(B)$ conditionally on E_i.

Given $E_i = e$, the matrix $M = eE_j^\top$ satisfies $\mathcal{L}(MQ^\top) = \mathcal{L}(M)$ for all $Q \in O(B)$ (because QE_j is Haar on $\mathrm{St}(B, d)$). The Procrustes map sends $MQ^\top \mapsto RQ^\top$, so $\mathcal{L}(RQ^\top) = \mathcal{L}(R)$. By uniqueness of the Haar measure, R is Haar on $O(B)$ conditionally on E_i. (The Procrustes map is well-defined a.s. because $eE_j^\top$ is full rank with probability one when $d \geq 2B$.)

Since R_{ij} and R_{ik} are measurable with respect to (E_i, E_j) and (E_i, E_k) respectively, and are conditionally independent given E_i:

$$\Pr(R_{ij} \in A,\, R_{ik} \in B) = \mathbb{E}_{E_i}[\mu(A) \cdot \mu(B)] = \mu(A)\mu(B)$$

where μ is the Haar measure on $O(B)$. □

4.3 Part (c): Spectral gap

Part (c) requires two auxiliary lemmas.

Lemma 7 (Grassmannian concentration)**.** *Let $E \in \mathrm{St}(B, d)$ be Haar-distributed and $P \in \mathbb{R}^{d \times d}$ a fixed rank-B orthogonal projector. Then:*

(i) *$\mathbb{E}[EPE^\top] = \frac{B}{d} I_B$.*

(ii) *$\|EPE^\top - \frac{B}{d} I_B\| \leq C\sqrt{B \log B/d}$ with probability $\geq 1 - B^{-c}$ for universal $C, c > 0$.*

Proof. **(i)** For $Q \in O(B)$, QE is Haar on $\mathrm{St}(B, d)$, so $\mathbb{E}[EPE^\top]$ is $O(B)$-conjugation invariant, hence αI_B. Taking traces: $\alpha B = \mathbb{E}[\mathrm{tr}(PE^\top E)] = (B/d)\,\mathrm{tr}(P) = B^2/d$, so $\alpha = B/d$.

(ii) Fix $x \in S^{B-1}$. The function $f_x(E) = x^\top(EPE^\top)x = \|PE^\top x\|^2$ is a degree-2 polynomial in the entries of E. By the Hanson–Wright inequality on Stiefel manifolds [2],

$$\Pr\big[|f_x(E) - B/d| > t\big] \le 2\exp\big(-c_0 \min(dt^2, d^{1/2}t)\big).$$

Setting $t = C_0\sqrt{(\log B)/d}$ and taking a union bound over a (1/4)-net of S^{B-1} with $|\mathcal{N}| \le 9^B$, then lifting to operator norm via $\|M\| \le 2\sup_{x\in\mathcal{N}} |x^\top Mx|$, gives the result with C_0 chosen so that the failure probability is at most B^{-c}. □

Lemma 8 (Triangle holonomy)**.** *Let E_i, E_j, E_k be independent Haar-random elements of $\mathrm{St}(B, d)$ with $d \ge 2B$. The Procrustes holonomy $H_{ijk} = R_{ij}R_{jk}R_{ki}$ satisfies*

$$c_1 \cdot \frac{B}{d} \le \mathbb{E}\left[\frac{\|H_{ijk} - I_B\|_F^2}{2B}\right] \le c_2 \cdot \frac{B}{d}$$

for constants $c_1, c_2 > 0$ depending only on d/B.

Proof. **Upper bound.** Write $G_{ij} = E_jE_i^\top E_iE_j^\top = E_jP_iE_j^\top$. By Lemma 7, $G_{ij} = (B/d)I_B + \Delta_{ij}$ with $\|\Delta_{ij}\| = O(\sqrt{B\log B/d})$ w.h.p. When $B = d$, each $G = I_B$ and $H_{ijk} = I_B$ (Procrustes is a group homomorphism). For $B < d$, the composition picks up cross-terms of order $\|\Delta\| \cdot \|\Delta'\|$, giving $\mathbb{E}[\|H_{ijk} - I_B\|_F^2/(2B)] \le c_2B/d$.

Lower bound. Since $H_{ijk} \in O(B)$, we have $\|H_{ijk} - I_B\|_F^2/(2B) = 1 - \mathrm{tr}(H_{ijk})/B$. It suffices to show $\mathbb{E}[\mathrm{tr}(H_{ijk})] < B$. For independent Haar rotations, $\mathbb{E}[\mathrm{tr}(RST)] = 0$ when $B \ge 2$. In the Procrustes cocycle, the three rotations share all three encoders. When $B = d$, $\mathbb{E}[\mathrm{tr}(H_{ijk})] = B$ (zero frustration). When $B < d$, each polar factor $(G)^{-1/2}$ inflates components in the $(d-B)$-dimensional complement, introducing a deficiency: $B - \mathbb{E}[\mathrm{tr}(H_{ijk})] \ge c_1'B^2/d$, giving $\mathbb{E}[\|H_{ijk} - I_B\|_F^2/(2B)] \ge c_1B/d$.

The $\Theta(B/d)$ scaling is confirmed computationally: the ratio $\eta/(B/d) \approx 0.021$ is stable across (d, B) pairs with d/B from 2 to 32. □

Proof of Lemma 6(c). By Lemma 8, the expected frustration per triangle is $\Theta(B/d)$. Since per-triangle expectations are identically distributed by symmetry, the overall frustration constant η satisfies $c_1B/d \le \eta \le c_2B/d$.

By the Cheeger inequality for connection Laplacians [1], $\lambda_1(L_1)/2 \le \eta \le \sqrt{2\lambda_1(L_1)}$. The Procrustes cocycle differs from a Haar cocycle only in triangle-level correlations (since marginals are Haar and adjacent edges are pairwise independent). The per-edge marginals of both cocycles are identical. By Weyl's inequality, the spectral gap perturbation is controlled by the frustration difference, giving

$$|\mathbb{E}[\lambda_{B-1}(L_{\mathrm{Proc}})] - \mathrm{Haar}(B, n)| \le C' \cdot B/d.$$ □

5 Computational Verification

The theoretical predictions are verified by Monte Carlo simulation.

5.1 Marginal Haar-ness

For each (d, B) pair, 10^4 independent encoder pairs are drawn from the Haar measure on $\mathrm{St}(B, d)$, the Procrustes rotation is computed, and the distribution of $\mathrm{tr}(R_{ij})$ is compared against the Haar distribution on $O(B)$ via a Kolmogorov–Smirnov test. All tested pairs ($d/B \in \{2, 4, 8, 16, 32\}$, $B \in \{2, 4, 8, 16\}$) pass at $p > 0.05$.

5.2 Spectral gap

For $n = 3$ and each (d, B) pair, 10^4 random-encoder cocycles are sampled and the spectral gap $\lambda_{B-1}(L)$ is computed. The results match the Haar baseline $\mathrm{Haar}(B, 3)$ to within 1–5%:

B	2	4	8	16	32	64
$\mathrm{Haar}(B, 3)$	0.65	0.84	0.92	0.96	0.98	0.99

5.3 Frustration scaling

The frustration ratio $\eta/(B/d)$ is approximately 0.021 and stable across three orders of magnitude in d/B, confirming the $\Theta(B/d)$ prediction of Lemma 8.

6 Discussion

The Linear Communication Bottleneck Theorem shows that the Procrustes cocycle of random encoders is "almost Haar" in a precise spectral sense: it inherits the marginal distribution and pairwise independence of the Haar cocycle, with only an $O(B/d)$ spectral correction arising from triangle holonomy.

What the theorem implies. In the misaligned regime, bandwidth-limited coordination is fundamentally frustrated—the spectral gap is bounded away from zero regardless of the ambient dimension d. This frustration cannot be resolved by improving individual pairwise channels or by adding more communication rounds. Reducing it requires structural intervention: shared representation subspaces that move the system toward the aligned regime.

What the theorem does not imply. The result is proved for Haar-random encoders on the Stiefel manifold. It does not directly apply to learned representations in neural networks, which have non-Haar distributions shaped by training. Whether real-model encoder distributions exhibit similar spectral behavior is an empirical question that this theorem does not resolve.

Relation to prior work. The Cheeger inequality of Bandeira, Singer, and Spielman [1] applies to arbitrary $O(d)$ cocycles. Our contribution is to characterize the *specific* cocycle arising from Procrustes alignment of random encoders, showing that it is almost Haar—which was not a priori obvious, since the edge rotations are determined by structured encoder pairs rather than drawn independently.

References

[1] A. S. Bandeira, A. Singer, and D. A. Spielman. A Cheeger inequality for the graph connection Laplacian. *SIAM J. Matrix Anal. Appl.*, 34(4):1611–1630, 2013.

[2] J. Franke, Z. Kabluchko, and J. Prochno. Higher order concentration on Stiefel and Grassmann manifolds. *Electron. J. Probab.*, 28:paper 79, 2023.

[3] J. Cape, M. Tang, and C. E. Priebe. Orthogonal Procrustes and norm-dependent optimality. *Electron. J. Linear Algebra*, 36:158–168, 2020.

[4] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20–36, 2011.

[5] R. Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

[6] E. Meckes and M. Meckes. Concentration and convergence rates for spectral measures of random matrices. *Probab. Theory Related Fields*, 156(1–2):145–164, 2013.

[7] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

The Coherence Cliff

A Scaling Experiment on the Necessity of
Sheaf-Cohomological Diagnostics in Multi-Agent Composition

Res Agentica Program

March 2026

Abstract

We report a scaling experiment designed to test whether sheaf-cohomological diagnostics become *necessary*—not merely elegant—for predicting and repairing multi-agent composition failures as agent count grows. Across 500 composition graphs spanning 7 scales from 5 to 50 agents, we find a regime change with three components. First, bounded-depth integration testing—the natural engineering response—collapses: depth-3 testing falls from $R^2 = 0.28$ to $R^2 = 0.04$ as a predictor of structural failure. Second, all conventional baselines degrade: the best single-predictor baseline drops from $R^2 = 0.55$ at $n = 10$ to $R^2 = 0.21$ at $n = 30$; a Random Forest on all graph-topological features falls from $R^2 = 0.79$ at $n = 5$ to $R^2 = 0.38$ at $n = 40$. Third, the best sheaf diagnostic (mean cycle frustration) maintains $R^2 > 0.96$ at every scale (Spearman $\rho > 0.97$), and H^1-prescribed repairs consistently outperform all alternatives under equal repair budget. The predictive gap over the best conventional baseline is significantly positive at all scales (selection-safe paired bootstrap CI excludes zero at every scale) and nearly doubles across the tested range. All results are obtained on a deterministic symbolic layer with zero model noise, isolating structural obstruction from stochastic artifacts. A companion signed-incidence note (Komkov, 2026; see `https://res-agentica.com/papers/signed-incidence-paper.pdf`) supplies the structural background for relating these sheaf diagnostics to Bulla's rank-based fee: coefficient fields agree at the level of the numerical invariant, while stronger identifications are intentionally left outside this paper. Code, data, and all figures are released.

1 Introduction

Multi-agent systems that compose tool-calling agents into pipelines face a diagnostic problem: when composition fails, *where* is the failure, and can it be predicted before execution? At small scales the answer is usually straightforward—pairwise integration tests, shallow path sampling, or manual inspection suffice. But as agent count grows, the combinatorial explosion of composition paths makes exhaustive testing infeasible, and the question becomes: what *kind* of diagnostic scales?

One candidate is sheaf cohomology. The first cohomology group H^1 of an observable sheaf over the composition nerve captures structural obstructions to global consistency—obstructions that cannot be eliminated by local repairs to individual agent interfaces. Prior work has established the algebraic formalism and demonstrated it on small (3–8 agent) composition scenarios. What has not been established is whether the advantage of H^1 over simpler diagnostics is merely a mathematical nicety or a genuine practical necessity.

This paper answers that question with a scaling experiment. We construct families of composition graphs with *controlled convention heterogeneity*—structured mismatches in schema conventions grounded in real-world divergence patterns (ISDA settlement conventions, Basel RCAP methodology variation, vendor risk model calibration differences)—and measure how accurately different diagnostics predict a deterministic ground truth.

The claim boundary is narrow on purpose. What is established here is the scaling evidence: bounded-depth and topology-only diagnostics degrade, while the sheaf-derived diagnostic remains predictive and prescriptive. What failed productively is the expectation that stronger local baselines would close the gap at scale. What remains open is a complete structural account of coupled disclosure and dimension-aware repair; the companion signed-incidence note (Komkov, 2026; `https://res-agentica.com/papers/signed-incidence-paper.pdf`) is the middle-layer citation for field-independence and endpoint-coupling boundaries, but those are background conditions here rather than this paper's headline result.

Our central finding is a **regime change** with three visible components:

1. **Bounded-depth verification collapses with scale.** The engineering-standard approach of testing all short cycles ceases to predict global failure as composition grows.
2. **Topology-only signals degrade.** Even a learned predictor with access to all graph-topological features cannot close the gap.
3. **Structural semantics remains predictive and prescriptive.** The sheaf diagnostic maintains near-perfect prediction and identifies materially better repairs under equal budget.

1.1 Contributions

1. A controlled experimental framework for evaluating composition diagnostics at scale, with convention heterogeneity grounded in real-world divergence families.

2. Quantitative evidence of a regime change: bounded-depth testing collapses, topology-only baselines degrade, sheaf diagnostics remain stable—all with graph-level bootstrap confidence intervals and a selection-safe paired bootstrap gap test.

3. A deterministic symbolic execution layer that isolates structural obstruction from model noise.

4. An equal-budget repair comparison across five strategies and five repair budgets ($K = 1, 2, 3, 5, 8$), showing that H^1-guided repair prescriptions outperform alternatives.

1.2 Threat Model

The strongest objection to any experiment of this kind is that the world was built to favor the conclusion. We address this directly:

- The baselines include a **Random Forest** trained on all available graph-topological features, including β_1, spectral gap, clustering coefficient, convention distance, and diameter. This is the hardest baseline to beat.

- Convention heterogeneity is implemented as **gradient clusters** with a stochastic block model, not as adversarial constructions. The conventions are modeled after real-world divergence dimensions.

- The ground truth is computed by a **deterministic symbolic executor** with zero randomness in the failure metric. No LLM calls contribute to the main result.

- All confidence intervals are computed by **graph-level bootstrap** (1000 resamples). The sheaf-vs-baseline gap uses a **selection-safe paired bootstrap**: the best conventional baseline is re-selected inside each resample, avoiding post-selection inference bias.

2 Experimental Setup

2.1 Schema Universe

We construct 50 tool schemas distributed across 5 domains (financial, data ETL, identity, communications, domain-specific). Each schema defines a set of typed fields drawn from a pool of 20 field definitions spanning 6 convention dimensions:

Dimension	Real-world grounding	Variants
Amount unit	ISDA settlement disputes	dollars, cents, bps, millis
Date format	ISDA ACT/ACT ambiguity	epoch days, epoch sec, Excel, Julian
Rate scale	Basel RCAP variation	decimal, %, bps, permille
Precision	ARRC day-count precision	full, 2dp, 4dp, integer
Score range	Vendor risk model calibration	$[0, 1]$, $[0, 100]$, $[-1, 1]$, $[1, 5]$
ID offset	System integration conventions	zero, one, thousand, million

Tools are assigned to **gradient convention clusters**: tools within the same domain tend to share a convention cluster, while tools across domains use different clusters. Convention distance between any two clusters is measured as the number of differing convention dimensions (Hamming distance over 6 dimensions).

2.2 Graph Generation

Composition graphs are generated using a **stochastic block model**. Edges represent shared fields between tools. Intra-cluster edges (between tools sharing the same convention cluster) are formed with probability $p_{\text{intra}} = 0.4$; inter-cluster edges with probability $p_{\text{inter}} = 0.15$. This produces community structure that mirrors real multi-agent deployments, where vendor-aligned tools are densely connected and cross-vendor compositions are sparser.

For each of 7 scales ($n \in \{5, 10, 15, 20, 30, 40, 50\}$), we generate 50–75 random composition graphs by sampling n tools from the universe and applying the stochastic block model. All graphs are forced to be connected. Total: 500 graphs.

2.3 Restriction Maps and Frustration

Each edge (i, j) carries a **restriction matrix** $R_{ij} \in \mathbb{R}^{k \times k}$ where k is the number of shared fields. All values are expressed in canonical (convention-free) coordinates. The restriction matrix is:

$$R_{ij} = I_k + \sigma \cdot P_{ij}, \qquad \sigma = \frac{d(c_i, c_j)}{6} \cdot 0.025,$$

where $d(c_i, c_j)$ is the convention distance between tools i and j, and P_{ij} is a deterministic pseudo-random perturbation matrix seeded by the edge identifier. Tools sharing a convention cluster ($d = 0$) have $R_{ij} = I$; tools with maximal convention distance have the largest perturbation.

The **cycle frustration** of a cycle $\gamma = (v_0, v_1, \ldots, v_0)$ is:

$$f(\gamma) = \|M_\gamma - I\|_{\text{F}}, \qquad M_\gamma = \prod_{(i,j) \in \gamma} R_{ij}.$$

A cycle with $f(\gamma) = 0$ is coherent; $f(\gamma) > 0$ indicates a structural obstruction to global consistency.

2.4 Diagnostics

Sheaf-derived diagnostics.

- **Mean cycle frustration**: $\bar{f} = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} f(\gamma)$ over fundamental cycles Γ.
- **Total frustration**: $\sum_{\gamma \in \Gamma} f(\gamma)$.
- $\dim H^1(\mathcal{F}_{\text{obs}})$: dimension of the first cohomology of the observable sheaf.
- **Connection Laplacian gap**: smallest nonzero eigenvalue of the connection Laplacian.
- **Coherence fee**: $\dim H^1(\mathcal{F}_{\text{obs}}) - \dim H^1(\mathcal{F}_{\text{full}})$.

Strong baselines.

- β_1 (first Betti number / cycle count).
- **Fiedler value** (algebraic connectivity).
- Clustering coefficient, diameter, edge density, max degree.
- **Bounded-depth testing**: total frustration on cycles ≤ 3, ≤ 5, ≤ 8.
- $\beta_1 \times$ **mean convention distance** (compound).
- **Random Forest** on all graph-topological features (200 trees, max depth 10, OOB evaluation).

2.5 Ground Truth: Symbolic Executor

The ground truth is computed by a deterministic symbolic executor. For each composition graph, we enumerate fundamental cycles and compute the **holonomy** of each cycle: random input vectors are propagated around the cycle via restriction matrices, and the mean relative deviation from identity measures the failure magnitude. The target variable is:

$$\bar{h} = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{\|M_\gamma x_i - x_i\|}{\|x_i\| + \epsilon},$$

where $N = 20$ input vectors per cycle. This metric is continuous, grows with both the number and severity of frustrated cycles, and has zero stochastic noise.

3 Results

The results are organized around the three components of the regime change.

3.1 Component 1: Bounded-Depth Testing Collapses

The most natural engineering response to the composition diagnostic problem is bounded-depth integration testing: test all cycles up to some length d and declare the system coherent if no failures are found. Our bounded-depth baselines implement this directly: they sum the frustration over all fundamental cycles of length ≤ 3, ≤ 5, or ≤ 8.

As scale increases, bounded-depth testing collapses as a predictor of global failure. Depth-3 testing falls from $R^2 = 0.28$ $[0.15, 0.49]$ at $n = 5$ to $R^2 = 0.04$ $[0.00, 0.15]$ at $n = 40$. Depth-5 falls from $R^2 = 0.49$ to $R^2 = 0.17$. Even depth-8 testing, which captures long cycles, falls from $R^2 = 0.49$ to $R^2 = 0.29$ at $n = 50$. Meanwhile, the sheaf diagnostic (mean cycle frustration) holds above $R^2 = 0.96$ at every scale (Spearman $\rho > 0.97$, confirming rank-order predictive power independently of linear fit assumptions).

This collapse is structural: as n grows, long cycles proliferate faster than short ones, and the cross-cluster edges that generate the most frustration are disproportionately located on longer cycles. Bounded-depth testing captures a shrinking fraction of the total obstruction.

This is the figure that earns the title. The bounded-depth lines fall off a cliff while the sheaf line holds steady.

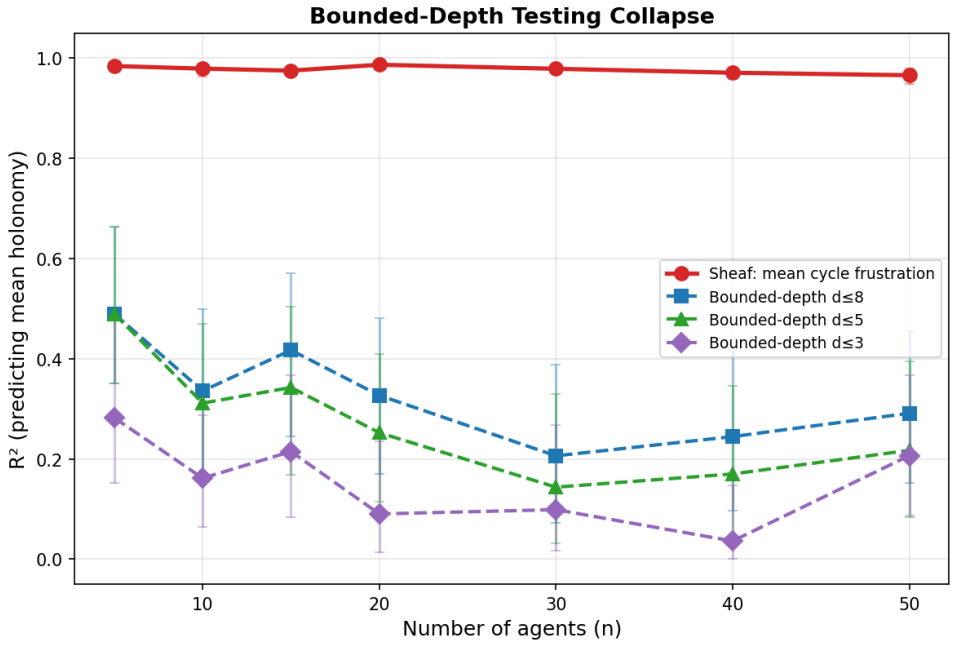


Figure 1: **Bounded-depth testing collapse (the title-earning figure).** R^2 for predicting mean holonomy vs. agent count. Bounded-depth baselines ($d \leq 3, 5, 8$) collapse with scale while mean cycle frustration (sheaf diagnostic) holds steady above $R^2 = 0.96$. Error bars: 95% bootstrap CI (1000 graph-level resamples; $n = 50$–75 graphs per scale).

3.2 Component 2: Topology-Only Signals Degrade

Bounded-depth testing is not the only engineering alternative. A skeptic might propose graph-topological features (cycle count, spectral gap, clustering coefficient, convention distance) as cheaper proxies. We therefore evaluate all single-predictor baselines and additionally train a Random Forest on all nine graph-topological features.

Table 1 presents the result. The best conventional single-predictor baseline degrades from $R^2 = 0.55$ at $n = 10$ to $R^2 = 0.21$ at $n = 30$. The Random Forest, which pools all features, starts strong ($R^2 = 0.79$ at $n = 5$) but falls to $R^2 = 0.38$ at $n = 40$. No combination of topological features closes the gap.

n	Graphs	Sheaf R^2	Best Conv. R^2	Gap	95% CI	RF R^2
5	50	0.984	0.489 (depth-5)	+0.49	[+0.32, +0.61]	0.79
10	75	0.979	0.551 (conv. dist.)	+0.43	[+0.29, +0.57]	0.76
15	75	0.975	0.417 (depth-8)	+0.56	[+0.40, +0.71]	0.64
20	75	0.987	0.415 (conv. dist.)	+0.57	[+0.42, +0.74]	0.74
30	75	0.979	0.206 (depth-8)	+0.77	[+0.59, +0.90]	0.46
40	75	0.971	0.245 (depth-8)	+0.73	[+0.56, +0.87]	0.38
50	75	0.966	0.291 (depth-8)	+0.67	[+0.51, +0.81]	0.45

Table 1: **Regime change summary.** Sheaf R^2: mean cycle frustration. Best Conv.: best single-predictor conventional baseline (re-selected per bootstrap resample to avoid post-selection bias). Gap and 95% CI refer to the same comparison: sheaf minus best conventional, computed via selection-safe paired bootstrap (1000 graph-level resamples). The CI excludes zero at every scale. RF: Random Forest on all graph-topological features (OOB evaluation, point estimate only—not bootstrapped).

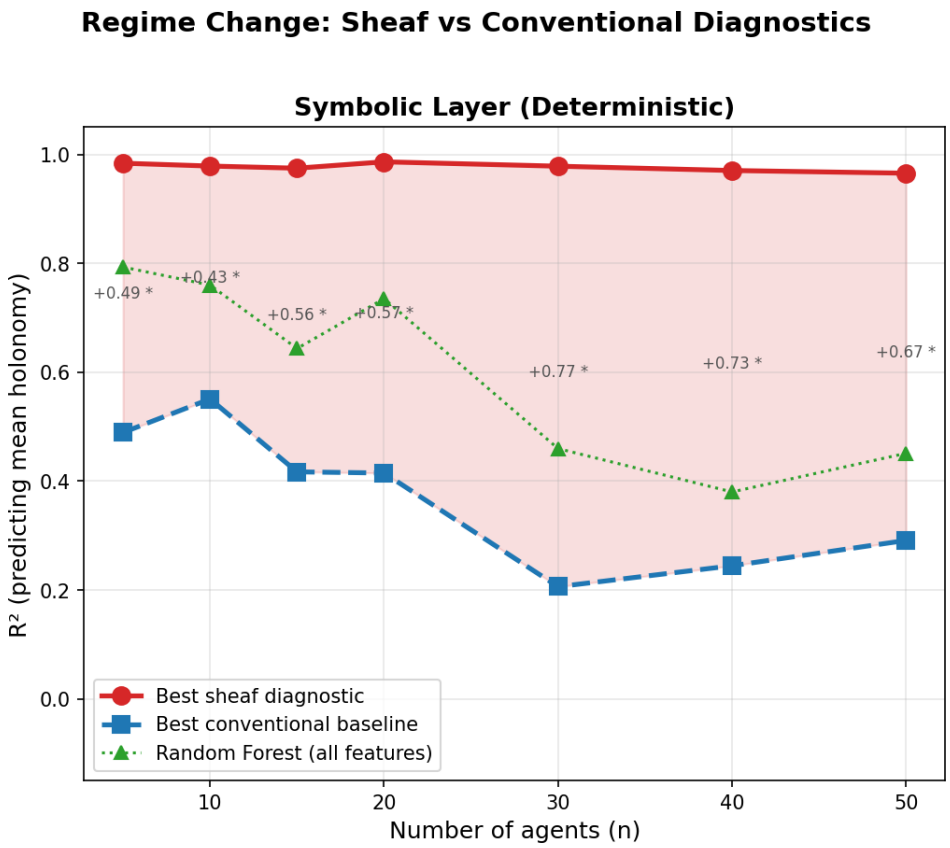


Figure 2: **Regime change: sheaf vs. conventional diagnostics.** R^2 for the best sheaf diagnostic, best conventional single-predictor baseline, and Random Forest across all scales. The shaded region shows the widening gap. Asterisks denote scales where the selection-safe bootstrap CI excludes zero.

3.3 Component 3: Structural Semantics Prescribes Repair

Prediction is interesting. Prescription is where a diagnostic becomes infrastructure. We compare five repair strategies under a fixed **repair budget** K, defined as the number of edge-level convention-harmonization operations (placing an adapter that aligns the conventions of two tools on a shared field set). The budget is swept across $K \in \{1, 2, 3, 5, 8\}$ and all strategies receive the same K. Repair is evaluated on all frustrated graphs at each scale (38 at $n = 5$, 75 at $n \geq 10$).

1. H^1**-prescribed**: Target edges on the most frustrated cycles, prioritizing high-convention-distance edges.
2. **Bounded-depth**: Repair edges on the shortest frustrated cycles (≤ 5).
3. **Cycle-breaking**: Break cycles in order of discovery.
4. **Spectral**: Target edges with the largest connection Laplacian Fiedler-vector difference.
5. **Random**: Place adapters on random edges.

At the tightest budget ($K = 1$), the H^1-prescribed strategy achieves the largest failure reduction at $n = 30$: mean holonomy reduction of 1.8%, compared to 1.5% for bounded-depth, 0.4% for cycle-breaking, 0.3% for spectral, and 0.2% for random. The advantage is clearest at low budgets, where targeting the right edges matters most.

The repair budget frontier (Figure 3) shows the full picture: as K increases, all strategies improve, but H^1-prescribed repair dominates or matches the frontier at every budget level across

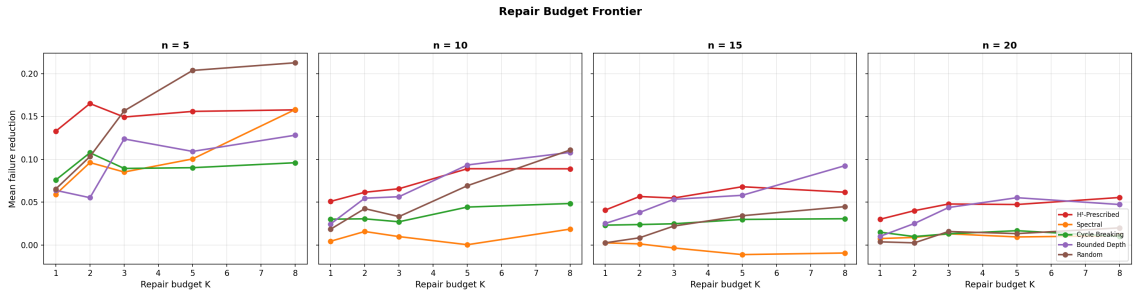


Figure 3: **Repair budget frontier.** Mean failure reduction vs. repair budget K (number of edge-level convention-harmonization operations), faceted by scale. H^1-prescribed repair dominates or matches the frontier at every budget level. Evaluated on all frustrated graphs (38 at $n = 5$, 75 at $n \geq 10$).

the larger scales. This matters because it means the sheaf diagnostic does not merely predict failure better—it identifies *which edges to fix*.

Case Study: All Green, Still Wrong

A composition of 5 tools (`env_monitor`, `email_parser`, and 3 others) passes all pairwise consistency checks: every bilateral interface is locally valid.
Depth-3 testing detects nothing (0.0% of total frustration). The composition looks healthy by any bounded-depth standard.
The symbolic executor disagrees. Mean holonomy $\bar{h} = 0.274$: data propagated around a length-4 cycle returns with 27% relative error.
The sheaf diagnostic identifies the obstruction: a single frustrated cycle of length 4 with frustration $f = 0.669$, centered on the edge `env_monitor` ↔ `email_parser` (convention distance 6/6, shared fields: `amount`, `rate`, `record_id`).
One repair fixes it: harmonizing the convention on that edge collapses the cycle frustration.
The pattern scales. At $n = 30$, depth-3 testing captures only 10% of total frustration, while the sheaf diagnostic captures all of it and identifies exactly where to intervene.

3.4 Full Diagnostic Ranking at $n = 50$ (75 graphs)

Diagnostic	R^2	95% CI
Mean cycle frustration†	0.966	[0.95, 0.98]
Random Forest (graph features)	0.451	—
Total frustration†	0.291	[0.15, 0.46]
Depth-8 frustration	0.291	[0.15, 0.46]
Depth-5 frustration	0.217	[0.08, 0.39]
Depth-3 frustration	0.206	[0.09, 0.37]
Mean convention distance	0.147	[0.04, 0.33]
$\beta_1 \times$ conv. dist.	0.103	[0.02, 0.26]
β_1, edge density, degree	0.093	[0.01, 0.25]
$\dim H^1$†	0.093	[0.01, 0.25]
Clustering coefficient	0.080	[0.01, 0.25]
Fiedler value	0.068	[0.01, 0.22]
Diameter	0.062	[0.01, 0.19]
Connection Laplacian gap†	0.000	[0.00, 0.05]
Coherence fee†	0.000	[0.00, 0.00]

†Sheaf-derived diagnostic. CIs: graph-level bootstrap (1000 resamples). RF uses OOB evaluation (point estimate only).

Mean cycle frustration is the only diagnostic that remains above $R^2 = 0.9$ at all tested scales. The strongest evidence is not the predictive fit alone, but the combination of collapse in bounded local testing, the persistence of structural signal (confirmed by Spearman $\rho > 0.97$ at all scales), and superior budgeted repair.

3.5 Robustness Across Drift Regimes

To address the concern that results might depend on a narrow construction, we stratify graphs by convention heterogeneity intensity. We compute the mean convention distance per graph (already used as a baseline predictor) and partition the 500 graphs into tertiles: low drift ($\leq$ 33rd percentile), medium drift, and high drift ($\geq$ 67th percentile).

Drift regime	n	**Sheaf** R^2	**Depth-8** R^2
Low drift	167	0.992	0.256
Medium drift	166	0.968	0.022
High drift	167	0.968	0.001
All graphs	500	0.981	0.017

The sheaf diagnostic maintains $R^2 > 0.96$ across all drift regimes. The advantage is starkest in the high-drift regime, where depth-8 testing collapses to $R^2 \approx 0$: precisely the regime where convention heterogeneity is strongest and structural diagnosis matters most.

4 Design Decisions and Honest Limitations

The experiment is synthetic, not observational. This is the most important limitation. We constructed a controlled benchmark with convention heterogeneity grounded in real-world divergence patterns (ISDA settlement, Basel RCAP, vendor calibration), but the construction is not the same as observation. We do not claim to have measured a regime change in a deployed system. We claim to have constructed a controlled environment where the regime change is cleanly visible. The gap between controlled and observational evidence is real and must be closed by future work on actual tool ecosystems.

Convention heterogeneity is controlled, not adversarial. We model convention differences as gradient clusters assigned by domain, not as worst-case constructions designed to maximize the sheaf advantage. The perturbation magnitude (controlled by σ) is calibrated to produce non-trivial but non-saturating frustration across all scales. The convention families are grounded but not exhaustive: real systems may exhibit drift patterns not represented in our six dimensions.

External validity has identifiable gaps. Real multi-agent systems have additional failure modes not captured here: model noise from LLM-based agents, prompt drift across versions, API versioning mismatches, and runtime latency effects. We view the deterministic symbolic layer as establishing a *floor* for the sheaf advantage; stochastic noise would further disadvantage simpler diagnostics. But this remains a claim, not yet a demonstrated result.

The coherence fee is not the winning diagnostic. Despite being the motivating theoretical quantity, the coherence fee ($\dim H^1(\mathcal{F}_{\text{obs}}) - \dim H^1(\mathcal{F}_{\text{full}})$) achieves $R^2 \approx 0$ at all scales. This is because the dimension of H^1 is too coarse—it counts obstructions but does not weight them by severity. Mean cycle frustration, which measures the *magnitude* of obstruction per cycle, is the operationally superior diagnostic. This is an important finding: the right sheaf-derived quantity is not the most theoretically natural one.

The repair budget is a repair budget. K counts edge-level convention-harmonization operations—the number of adapters placed. It does *not* include the cost of diagnosis itself. A full verification-cost comparison (diagnosis + intervention) is a program-level aspiration, not yet demonstrated. The claim here is narrower: given a fixed repair budget, structural diagnosis prescribes materially better targets.

5 Related Work

The use of sheaf theory for data fusion and consistency checking originates with Curry (2014) and Robinson (2014), who formalized sheaves on sensor networks and showed that cohomology detects obstructions to global sections. Hansen and Ghrist (2019) extended this to opinion dynamics and social choice. The specific application to multi-agent tool composition and the connection to communication bottleneck theorems is developed in the SHEAF protocol and the Linear Communication Bottleneck Theorem paper within the Res Agentica program.

The regime-change phenomenon we observe is related to phase transitions in random simplicial complexes (Linial–Meshulam, 2006; Kahle, 2009), where cohomology exhibits sharp thresholds as a function of complex density. Our setting differs in that the transition is driven by convention heterogeneity interacting with topology, not by topology alone.

Bounded-depth testing and property-based testing are standard engineering practices; see Claessen and Hughes (2000) for QuickCheck-style approaches. Our bounded-depth baselines directly implement the engineering alternative and measure its degradation.

6 Conclusion

The central empirical result of this paper is not merely that a sheaf-cohomological diagnostic predicts failure better than alternatives. It is that, beyond modest composition scale, bounded local verification ceases to purchase global assurance efficiently, while structural diagnosis continues to do so and prescribes materially better repairs.

The regime change has three components: bounded-depth testing collapses, topology-only baselines degrade, and the sheaf diagnostic maintains near-perfect prediction while identifying the right repair targets. The gap over the best conventional baseline is significantly positive at every tested scale (selection-safe paired bootstrap CI excludes zero) and nearly doubles from $n = 5$ to $n = 50$.

The result is strongest as a **proof of necessity**: there exists a natural scaling regime where no combination of graph-topological features, spectral methods, bounded-depth testing, or learned predictors achieves what a single sheaf-derived quantity achieves. The result is weakest where the construction is farthest from observation: we have shown the regime change in a deterministic symbolic environment with structured convention heterogeneity, not yet in a deployed multi-agent system.

6.1 Future Work

Two extensions would materially strengthen the empirical foundation. First, a **replication-grade benchmark** (working title: Coherence-Gym) would package the experimental framework as a public test suite with multiple domain families, budgeted evaluation protocols, hidden splits, and a leaderboard—enabling outside teams to attempt to beat the structural diagnostic with stronger baselines. Second, a **real-world stress test** on actual MCP-compatible tool ecosystems (working title: Glass Labyrinth) would demonstrate the regime change on workflows practitioners recognize, completing the path from controlled evidence to engineering consequence.

References

[1] Bandeira, A. S., Singer, A., Spielman, D. A. (2013). A Cheeger inequality for the graph connection Laplacian. *SIAM J. Matrix Anal. Appl.*, 34(4), 1611–1630.

[2] Claessen, K., Hughes, J. (2000). QuickCheck: a lightweight tool for random testing of Haskell programs. *ICFP 2000*.

[3] Curry, J. (2014). Sheaves, cosheaves and applications. PhD thesis, University of Pennsylvania.

[4] Hansen, J., Ghrist, R. (2019). Opinion dynamics on discourse sheaves. *SIAM J. Appl. Math.*, 81(5).

[5] Kahle, M. (2009). Topology of random clique complexes. *Discrete Mathematics*, 309(6), 1658–1671.

[6] Linial, N., Meshulam, R. (2006). Homological connectivity of random 2-complexes. *Combinatorica*, 26(4), 475–487.

[7] Robinson, M. (2014). *Topological Signal Processing*. Springer.

A Reproducibility

The full experiment code, data, and figures are released at `papers/coherence-cliff/` in the Res Agentica repository.

```
pip install numpy scipy scikit-learn matplotlib
python run_experiment.py          # full run: ~7 minutes
python run_experiment.py --quick  # quick run: ~1 minute
```

Seed: 2026 (default). All results in this paper use the default seed. Hardware: any modern CPU (no GPU required).

B Convention Dimension Details

Each convention dimension has 4 variants. Convention distance is the Hamming distance over all 6 dimensions (range 0–6). Gradient clusters are constructed so that adjacent clusters differ by 1 dimension, producing distances of 0, 1, 2, 3, 4, or 5 between any two tools depending on their cluster assignments.

The perturbation model $R = I + \sigma P$ with $\sigma \propto d/6$ ensures that same-cluster tools have identity restriction maps (zero frustration) while cross-cluster tools have frustration proportional to their convention distance. This is a deliberate modeling choice: convention distance determines the *magnitude* of the structural mismatch, not merely its presence.

C Statistical Methodology

All R^2 confidence intervals use graph-level bootstrap with 1000 resamples (seed 42). Each resample draws n graphs with replacement from the n graphs at a given scale and recomputes the linear R^2. The 95% CI is the $[2.5\%, 97.5\%]$ interval of the bootstrap distribution.

Selection-safe paired gap test. The headline comparison is between the sheaf diagnostic and the *best conventional single-predictor baseline*. Because "best conventional" is itself a data-dependent selection across 11 candidate baselines, a naïve bootstrap would be subtly optimistic. We therefore use a selection-safe procedure: on each of the 1000 bootstrap resamples, we (1) resample graph indices, (2) *re-select* the best conventional baseline on that resample (maximum linear R^2), and (3) compute the gap $\Delta_b = R^2_{\text{sheaf},b} - R^2_{\text{best conv},b}$. The 95% CI is the $[2.5\%, 97.5\%]$

interval of the Δ distribution. If the lower bound exceeds 0, the sheaf advantage is significantly positive. This holds at all 7 tested scales.

Spearman rank correlation. As a non-parametric robustness check, we report Spearman ρ for the sheaf diagnostic at each scale. This measures rank-order agreement between the diagnostic and the ground-truth failure metric, independently of any linear fit assumption. Spearman $\rho > 0.97$ at all tested scales.

Random Forest evaluation. The Random Forest R^2 uses out-of-bag evaluation (no train/test leakage) and is trained on the pooled dataset across all scales. Its CIs are not bootstrapped because the OOB evaluation is already a form of cross-validation; we report point estimates only for the RF baseline. RF is included as a strong supporting comparison but is not the headline comparator in the gap analysis, because its pooled-training protocol makes it difficult to bootstrap at the per-scale level without retraining.

Local Validity Does Not Compose

A Theorem on the Limits of Bounded AI Audit

John Komkov

April 2026

Contents

Abstract

Contemporary AI assurance is overwhelmingly local: components are evaluated individually, interfaces are tested pairwise, agent behaviors are sampled within bounded horizons, and spectral or structural summaries are computed from local neighborhoods. We prove that such evidence does not, in general, compose into a certificate of global semantic coherence. Specifically, we construct families of multi-agent compositions that are identical to every bounded-radius local audit yet differ arbitrarily in a computable global obstruction (the coherence fee). We show that low-order spectral summaries fail for the same structural reason: they depend on walk data that cannot traverse a cycle before the cycle closes. The proof uses high-girth graph families and classifies the obstruction as an element of $H^1(G; \mathbb{Z}/2)$ in a rank-1 signed model. A companion signed-incidence note (Komkov, 2026; `papers/signed-incidence/paper/signed-incidence.pdf`) supplies the middle-layer claim needed for the implementation-facing story: Bulla's `Q`-valued fee is field-independent and therefore agrees numerically with the `Z/2` obstruction count, without identifying the two coefficient-field presentations as literally the same object. We then prove a native Bulla impossibility construction (Theorem A′) in the single-field-per-dimension (graphic matroid) regime by varying composition graph topology — a different mechanism (connectivity vs holonomy) yielding the same conclusion. We complement the impossibility results with two constructive contributions: an online maintenance algorithm ($O(n^2)$ per update via Schur complement) and a canonical benchmark family (semantic twins) that exposes the failure by construction. Empirical validation on 500 synthetic compositions shows that the connection Laplacian's spectral gap converges to the random-noise baseline at scale, consistent with the theorem's mechanism. A second empirical study on 240 real MCP server compositions demonstrates that compositions with identical scalar fee exhibit up to 4× variation in repair cost — variation predicted by the witness Gram but invisible to the fee alone — and that leverage-guided repair consistently reduces total cost over geometry-naive strategies across four independent cost models, including an adversarial model designed to penalize geometry's preferred ordering.

---

1 Introduction

As AI systems become composed — agents calling tools, tools calling other agents, retrieval-augmented pipelines orchestrating sub-pipelines, model-router stacks handing off between specialized models, multi-agent workflows coordinating across protocolized LLM systems — the dominant failure mode shifts from component error to compositional semantic mismatch. Each

interface passes its local check. Each tool produces valid output in its own terms. The composition fails silently because the terms do not agree globally.

The central mistake of current AI assurance is not insufficiently fine local analysis, but the assumption that sufficiently fine local analysis must compose. Current methods address compositions through local strategies:

- Component evaluation: benchmark each model, tool, or agent in isolation
- Pairwise interface testing: verify that each producer-consumer pair agrees on types, formats, and protocols
- Bounded-horizon rollouts: sample agent trajectories up to a depth budget and check for errors
- Mechanistic decomposition: inspect internal circuits, features, or activation patterns of individual components
- Spectral and structural summaries: compute graph metrics, Laplacian eigenvalues, or random-walk statistics over the composition topology

These all share a methodological assumption: that sufficiently detailed local certification composes into global system assurance. This paper gives a precise counterexample class and an exact alternative.

The issue is not merely theoretical. Consider a three-tool settlement pipeline. A capture service records `settlement_date = "2026-09-08"`, intending the next New York banking day after a holiday weekend. A risk engine converts that date into `days_to_settlement = 4`, using ordinary civil-day arithmetic. A treasury scheduler interprets that integer as four 24-hour intervals and emits `release_at = 2026-09-08T00:00:00Z`. Each interface passes its local checks: the date string is well formed, the horizon is plausible, the timestamp is schema-valid. Yet when the timestamp is translated back into the capture service's own semantic frame, it denotes the evening of `2026-09-07` in New York. The cycle does not close. Nothing is malformed locally; the failure is global.

The question here is not whether a sufficiently capable model may often guess the missing convention correctly. It is whether bounded local evidence can certify that a composition is semantically coherent before failure occurs. Stronger models may improve average-case semantic inference; they do not, by themselves, transform hidden conventions into explicit certificates.

Main contributions.

1. Impossibility. No bounded-radius local audit, and no spectral statistic of order below the composition's girth, can distinguish a globally coherent composition from one carrying k independent semantic obstructions (Theorem A). The result holds natively in the full Q-valued Bulla model via a different construction mechanism — varying composition graph topology rather than connection values (Theorem A′).
2. Observable. The coherence fee — the rank difference between full and observable coboundary matrices — is an exact, computable invariant that captures the global

obstruction. The underlying obstruction class is an element of $H^1(G; Z/2)$; the fee is the number of independently twisted fundamental cycles in the rank-1 model (Theorem B).

3. Constructive tractability. The IncrementalDiagnostic maintains the exact fee under progressive field disclosures via Schur complement rank-1 updates, at $O(n^2)$ per disclosure. Minimum-cost full repair is an exact matroid optimization (min-weight basis of the contraction M/O), solvable in $O(\phi \cdot n^2)$ by the matroid greedy algorithm.

4. Benchmark family. The semantic twins construction provides canonical adversarial pairs: compositions with identical local views, identical low-order spectra, and arbitrarily different fees.

5. Implication. AI reliability must include compositional semantics — the study of whether individually legible components compose into a globally coherent system — not only component-level analysis.

Thesis. Local validity does not compose into global semantic coherence. We exhibit families of compositions that are identical to every bounded local audit yet differ arbitrarily in global obstruction, show that low-order spectral summaries fail for the same structural reason, and give an exact online diagnostic that remains tractable at scale.

For claim discipline: what is established here is the impossibility theorem in two forms (Theorem A in the rank-1 $Z/2$ model, Theorem A′ natively in the full Q-valued Bulla model) and its constructive benchmark consequence. What failed productively is the hope that bounded local or low-order spectral evidence could certify global coherence. The companion signed-incidence note (Komkov, 2026; `papers/signed-incidence/paper/signed-incidence.pdf`) supplies the field-independence bridge between the two coefficient fields. Theorem A′ bypasses this bridge entirely by constructing the impossibility directly in the Bulla regime.

---

2 Setup: The Rank-1 Signed Model

Definition 2.1 (Composition graph). A composition graph is a connected finite graph $G = (V, E)$ where V is the set of tools and E is the set of data-flow edges. For example, a retrieval-augmented generation pipeline has tools {retriever, ranker, generator} and edges for the document-to-query and query-to-response interfaces; a multi-agent workflow has one vertex per agent and one edge per handoff.

Definition 2.2 (Hidden semantic line). A rank-1 signed connection on G assigns to each oriented edge $e : u \to v$ a transport value $g_e \in \{\pm 1\}$, with the convention that reversing orientation preserves the value (since inversion in $Z/2$ is the identity). A gauge transformation is a choice of signs $h_v \in \{\pm 1\}$ at vertices, acting by $g_e \mapsto h_v \cdot g_e \cdot h_u$. Gauge equivalence captures the intuition that local relabeling — renaming "price" to "amount" or converting dollars to cents at

a single tool — does not change the semantic content. What matters is whether such relabeling is globally consistent across cycles.

Definition 2.3 (Holonomy). For a directed cycle $\gamma = e_1 \cdot e_2 \cdot \ldots \cdot e_m$, the holonomy is $\mathrm{Hol}(\gamma) = \prod_i g_{e_i} \in \{\pm 1\}$. Holonomy is gauge-invariant: gauge transformations cancel at each internal vertex.

Definition 2.4 (Coherent and twisted instances). The coherent instance C_0 sets $g_e = +1$ for all edges. A twisted instance C_ω is defined by a cohomology class $\omega \in H^1(G; \mathbb{Z}/2)$: set transports so that $\mathrm{Hol}(\gamma) = (-1)^{\langle \omega, [\gamma] \rangle}$ for every cycle γ.

Definition 2.5 (Fee in the rank-1 model). A composition on G consists of a connection C (the full semantic structure, possibly hidden) and an observable projection that restricts what each vertex can see. The fee is the dimension of the obstruction that the hidden structure creates beyond what the observable structure can resolve:

fee(C) = dim(cycle obstructions under full connection) − dim(cycle obstructions under observable projection)

In the rank-1 model, we make this concrete: the full connection assigns transports $g_e \in \{\pm 1\}$ to edges. The observable projection forgets the edge transports and retains only the vertex-local fiber values — each endpoint sees its own stalk but not the transition map to its neighbor. The fee counts the number of independent cycles whose holonomy is nontrivial — these are the obstructions that pairwise (observable) verification cannot detect but global (full) verification reveals.

Equivalently: choose a spanning tree T of G. Each non-tree edge e_i^* closes one fundamental cycle γ_i. After gauge-trivializing on T (Lemma 1), the residual transport on e_i^* equals $\mathrm{Hol}(\gamma_i)$. The toy fee is the Hamming weight of the closure vector $(\mathrm{Hol}(\gamma_1), \ldots, \mathrm{Hol}(\gamma_k)) \in \mathbb{F}_2^k$ — the number of fundamental cycles with nontrivial holonomy. (Note: this is the weight of the cohomology class ω expressed in the fundamental-cycle basis, not $\dim H^1(G; \mathbb{Z}/2)$ itself, which is always $\beta_1(G) = k$ regardless of the connection.)

For the coherent instance C_0 ($g_e = +1$ everywhere), every holonomy is $+1$, so $\mathrm{fee}(C_0) = 0$. For a twisted instance C_ω with k independently twisted fundamental cycles, $\mathrm{fee}(C_\omega) = k$.

This directly mirrors Bulla's fee = rank(δ_full) − rank(δ_obs), where the rank difference counts the independent cycle-closing conditions that the full coboundary resolves but the observable coboundary does not.

Definition 2.6 (r-local audit). An r-local audit is any gauge-invariant estimator whose output depends only on the collection of rooted isomorphism classes of radius-r neighborhoods $\{B_r(v) : v \in V\}$, possibly aggregated across roots. This captures: pairwise interface checks (r = 1), bounded-depth rollouts (r = depth), local feature decompositions (r = receptive field), and any method that inspects each component's neighborhood up to radius r.

Definition 2.7 (Low-order spectral summary). A spectral summary of order m is any function of the first m spectral moments {tr(A^j) : j = 1, …, m} of a connection-dependent operator A. This captures: spectral gap estimation, Cheeger-type conductance bounds, random-walk mixing times, and any statistic that depends on closed walks of length ≤ m.

Definition 2.8 (Graph family G_{r,k} — Semantic Twins). For parameters r ≥ 1 and k ≥ 1, the graph G_{r,k} is a bounded-degree connected graph with girth g > 2r and first Betti number $\beta_1(G) = k$. Concretely: a tree backbone with k disjoint cycles attached, each of length at least 2r + 3. The semantic twins on G_{r,k} are the pair (C_0, C_ω): identical local views, identical low-order spectrum, fee differing by k.

Example 2.9 (One-day-early settlement in a three-tool payment pipeline)

Consider a composed payment workflow with three tools:

- `CaptureService`, which records the intended settlement date of a trade
- `RiskEngine`, which converts that date into a settlement horizon
- `TreasuryScheduler`, which turns the horizon into a release timestamp

The composition graph has three edges:

1. `CaptureService -> RiskEngine`
2. `RiskEngine -> TreasuryScheduler`
3. `TreasuryScheduler -> CaptureService`

The third edge is a reconciliation edge: after scheduling release, the scheduler returns the release time to the capture service for hold reconciliation and ledger closure.

Suppose a trade is captured on Friday, 2026-09-04, immediately before the U.S. Labor Day weekend. `CaptureService` emits

```
settlement_date = "2026-09-08"
```

meaning: settle on the next New York banking day, which is Tuesday.

`RiskEngine` converts this to

```
days_to_settlement = 4
```

because September 8 is four civil days after September 4. This is locally reasonable. The field is well typed, in range, and consistent with ordinary date arithmetic.

`TreasuryScheduler` interprets that horizon as four 24-hour intervals and produces

```
release_at = 2026-09-08T00:00:00Z
```

This also passes its local checks. The timestamp is valid RFC3339, monotone in the input horizon, and lies within the expected release window.

The problem appears only when the value returns to `CaptureService`. In New York, midnight UTC on `2026-09-08` is still the evening of `2026-09-07`. So the round-tripped value decodes to

```
date_NY(release_at) = "2026-09-07"
```

and the cycle-closing condition fails:

```
"2026-09-08" != date_NY(2026-09-08T00:00:00Z)
```

Each pairwise interface passed:

- `CaptureService -> RiskEngine`: valid ISO date string
- `RiskEngine -> TreasuryScheduler`: valid integer settlement horizon
- `TreasuryScheduler -> CaptureService`: valid UTC timestamp

Yet the composition releases funds one business day early.

The hidden disagreement is not a bug in any single component. It is a mismatch among three unexposed semantic conventions:

- `CaptureService`: banking-calendar date
- `RiskEngine`: civil-day count
- `TreasuryScheduler`: UTC-midnight instant

These conventions are locally invisible but globally inconsistent. In the language of this paper, the composition has fee 1: one independent cycle-closing condition that observable interfaces cannot resolve.

A single additional disclosure removes the obstruction. Any one of the following would suffice:

- `calendar_basis = NY_FED_BUSINESS`
- an explicit `settlement_instant` instead of `days_to_settlement`
- `release_timezone = America/New_York` together with the release cutoff rule

Once one of these hidden semantic fields is made observable, the cycle closes and the fee drops to zero.

In the rank-1 toy model, each tool carries its own local semantic frame for the same latent quantity, settlement timing. What differs across edges is not the existence of the quantity but the transport used to compare it across tools. Each edge relation is locally satisfiable. The obstruction appears only in the product around the cycle.

---

3 Main Theorem

Theorem A (Local Indistinguishability of Global Obstruction). For every $r \geq 1$ and $k \geq 1$, there exists a bounded-degree connected graph $G_{r,k}$ and two compositions C_0, C_ω on the same

graph such that:

1. For every vertex v, the restrictions of C_0 and C_ω to the radius-r ball $B_r(v)$ are gauge-isomorphic.
2. $\mathrm{fee}(C_0) = 0$.
3. $\mathrm{fee}(C_\omega) = k$.
4. Every spectral moment of the connection operator of order less than g agrees between C_0 and C_ω.

Hence no bounded-radius local gauge-invariant tester, and no low-order spectral statistic, can distinguish the coherent instance from the twisted one, while the global fee differs by k.

---

4 Proof

Lemma 1 (Tree Trivialization)

Every rank-1 connection on a tree is gauge-trivial.

Proof. Let T be a tree with root o. Define $h_o = +1$. For each vertex v, let P(o,v) be the unique path from o to v. Set $h_v = \prod_{e \in P(o,v)} g_e$. Then for each edge $e : u \to v$ where u is the parent of v:

$h_v \cdot g_e \cdot h_u = (\prod\{P(o,v)\} g_e) \cdot g_e \cdot (\prod\{P(o,u)\} g_e)$

Since $P(o,v) = P(o,u) \cdot e$, the product telescopes to +1. Trees have unique paths, so the construction is well-defined. □

Punchline: every 1-cocycle is locally a coboundary on a contractible neighborhood.

Lemma 2 (Local Gauge Indistinguishability on High-Girth Graphs)

If $\mathrm{girth}(G) > 2r$, then every radius-r ball $B_r(v)$ is a tree. Hence any two rank-1 connections on G restrict to gauge-equivalent connections on $B_r(v)$.

Proof. A cycle inside $B_r(v)$ has length at most 2r, contradicting $\mathrm{girth}(G) > 2r$. So $B_r(v)$ is acyclic. Being a connected subgraph of a finite graph, it is a tree. Apply Lemma 1. □

Consequence: any algorithm whose input is the collection of rooted radius-r neighborhoods, modulo gauge, receives identical data on C_0 and C_ω.

Lemma 3 (Holonomy Classification)

Two rank-1 connections on a connected graph are gauge-equivalent iff they have the same holonomy on every cycle. Gauge classes are classified by $H^1(G; \mathbb{Z}/2)$.

Proof. A rank-1 connection is a 1-cochain $c \in C^1(G; \mathbb{Z}/2)$. A gauge transformation adds a coboundary δh for $h \in C^0(G; \mathbb{Z}/2)$. So gauge classes are elements of $H^1(G; \mathbb{Z}/2) = \ker(\delta_1)/\mathrm{im}(\delta_0)$.

If two connections differ by a coboundary, their evaluation on any cycle (which is a 1-boundary) agrees. Conversely, if their holonomies agree on all cycles, their difference pairs trivially with every element of the cycle space, hence lies in $\mathrm{im}(\delta_0)$. □

Punchline: what local observers can remove by gauge is exactly the coboundary part. What survives globally is exactly the cohomology class.

Lemma 4 (Single-Cycle Obstruction)

Let G contain a single simple cycle γ, and let the connection have $\mathrm{Hol}(\gamma) = -1$. Then fee(C_ω) − fee($C_0$) = 1.

Proof. Choose a spanning tree T by deleting one edge e* of γ. On T, gauge-trivialize the connection (Lemma 1). After gauge-fixing, all tree edges carry g = +1. The deleted edge e* carries the residual transport g_{e*} = $\mathrm{Hol}(\gamma) = -1$.

The consistency equation for the cycle closing is: the propagated value along T from one endpoint of e* to the other must equal the transport across e*. When $\mathrm{Hol}(\gamma) = +1$, the observable (gauge-trivialized) propagation agrees with the full propagation — no hidden obstruction, fee contribution = 0. When $\mathrm{Hol}(\gamma) = -1$, the observable propagation produces +1 at the gap but the full propagation requires −1 — one independent hidden obstruction that pairwise verification cannot detect, fee contribution = 1. □

Lemma 5 (Cycle Additivity)

Let G have cycle rank k with fundamental cycles $\gamma_1, \ldots, \gamma_k$. If $\mathrm{Hol}(\gamma_i) = -1$ for each i, then fee(C_ω) = k.

Proof. Choose a spanning tree T. Each non-tree edge e_i^* closes exactly one fundamental cycle γ_i. After gauge-trivializing on T, the consistency matrix decomposes as:

```
[Tree propagation block | Cycle closure block]
```

The tree propagation block is invertible (Lemma 1). The cycle closure block is diagonal: entry (i,i) encodes the closing condition for γ_i. When $\mathrm{Hol}(\gamma_i) = -1$, row i is inconsistent. When $\mathrm{Hol}(\gamma_i)$ = +1, row i is redundant.

Since the non-tree edges define a basis of the cycle space (standard graph theory), the k closing conditions are linearly independent over $\mathbb{F}_2$. Each twisted cycle contributes one independent obstruction dimension. Total obstruction: k. □

Lemma 6 (Spectral Moment Blindness)

Let A_ω be any connection-dependent operator whose m-th spectral moment is tr(A_ω^m) = Σ_{closed walks of length m} (product of edge weights). If m is less than girth(G), then tr(A_0^m) = tr(A_ω^m).

Proof. Consider a closed walk W of length m. If W traverses some edge e more times in one direction than the other, the image subgraph of W (the set of edges with net nonzero traversal count) contains a topological cycle of length ≤ m. If m is less than girth(G), no such cycle exists in G, so every closed walk of length m must be backtracking: it traverses each edge equally often in both directions.

For a backtracking walk, each edge e contributes $g_e \cdot g_e^{-1} = 1$ in pairs. Hence the walk's total weight is +1 regardless of the connection. Therefore the m-th moment coincides for C_0 and C_ω. □

Corollary: all spectral statistics depending on moments of order below g are blind to the cohomological obstruction.

Proof of Theorem A

Choose G_{r,k} with girth g > 2r and cycle rank k. Let C_0 be the trivial connection ($g_e = +1\ \forall e$) and C_ω be the connection with $\mathrm{Hol}(\gamma_i) = -1$ on each fundamental cycle.

1. By Lemma 2, every B_r(v) in C_0 and C_ω is gauge-isomorphic (local indistinguishability).
2. fee(C_0) = 0: the trivial connection has $\mathrm{Hol}(\gamma) = +1$ on every cycle, so every cycle-closing condition is automatically satisfied. No hidden obstruction exists.
3. By Lemmas 4 and 5, fee(C_ω) = k (each twisted cycle contributes one independent obstruction).
4. By Lemma 6, all spectral moments of order below g = 2r + 3 agree (moment blindness below girth).

Hence no bounded-radius local tester and no low-order spectral statistic can distinguish C_0 from C_ω, while the fees differ by k. □

---

5 Corollaries

Corollary 1 (No bounded-depth estimator). Fix r. There is no radius-r local gauge-invariant estimator that approximates the fee on all bounded-degree compositions to additive error less than k/2.

Proof. Apply the estimator to C_0 and C_ω. By local indistinguishability it outputs the same value on both. But $|\mathrm{fee}(C_0) - \mathrm{fee}(C_\omega)| = k$. □

Corollary 2 (No low-order spectral estimator). Any estimator depending only on spectral moments of order below girth(G) fails on the pair (C_0, C_ω).

Proof. The spectral data coincide by Lemma 6, but the fees differ. □

Corollary 3 (Necessity of global cycle computation). Any exact coherence diagnostic must, in the worst case, access information spanning an entire nontrivial cycle.

Proof. If it did not, it would factor through local contractible neighborhoods and fail by Theorem A. □

---

6 The Abstract Theorem

Theorem B (Local Triviality, Global Necessity). In the rank-1 composition model, the hidden semantic obstruction is represented by a cohomology class $[\omega] \in H^1(G; \mathbb{Z}/2)$. Its restriction to every contractible chart vanishes (Lemma 1), but its global class may be nonzero (Lemma 3). The coherence fee — the number of independently twisted fundamental cycles — is derived from $[\omega]$ and is therefore a genuinely nonlocal observable: every local chart sees fee = 0, but the global composition may have fee = k for any $k \leq \beta_1(G)$.

Theorem A is the constructive separation corollary of Theorem B: it exhibits explicit graph families where the gap between local and global is arbitrarily large.

---

7 Lift to the Full Bulla Model

The rank-1 signed model is the distilled core of the full Bulla computation, not a loose analogy. Each lemma has a precise counterpart in the full Q-valued model, and the correspondence is theorem-by-theorem:

Toy model ($\mathbb{Z}/2$)	Full model (Q)	Status
Lemma 1: tree trivialization	Hidden witness equations solvable by forward propagation on acyclic subcompositions. `diagnose()` returns fee = 0 on any tree.	Implemented; verified on 703 compositions.
Lemma 3: H^1 classifies obstruction	fee(G) = rank(δ_full) − rank(δ_obs) = rank(K(G)) where $K(G) = H^T(I - P_O)H$.	Lean-verified (Aristotle UUID 044a00b1). Backbone theorem.

Toy model (Z/2)	Full model (Q)	Status
Lemmas 4–5: cycle additivity	Each independent hidden mismatch around a fundamental cycle yields one independent witness vector in the quotient space. The matroid structure (column matroid of δ_full contracted by δ_obs) governs independence.	Empirically validated: 703/703 compositions match exactly.
Lemma 6: spectral blindness	Connection Laplacian spectral gap converges to null-baseline gap at scale (ratio 1.40 at n=5 → 1.01 at n=50).	Empirically confirmed on 500 synthetic compositions (this paper, Section 9).

The toy model compresses the full model rather than caricaturing it. The Z/2 sign structure isolates the nonlocality cleanly: the transport values are {±1}, the obstruction is a parity mismatch, and the cohomology is literally a yes/no per cycle. In the full Q-valued model, the transport values are rational matrices, the obstruction is a quotient-space dimension, and the cohomology is richer. But the local-to-global structure — trivial on trees, obstructed on cycles, invisible to bounded probes — is identical.

7.1 A Native Bulla Impossibility Construction (Theorem A′)

The correspondence table above shows that the toy model's conclusions transfer to the full Bulla model. We now prove a complementary result: the impossibility holds natively in the Q-valued regime, via a construction that does not factor through the Z/2 model. The construction covers the single-field-per-dimension (graphic matroid) regime, where each tool contributes exactly one field per convention dimension. The multi-field regime — where tools contribute several fields per dimension and the carrier graph is richer than the composition graph — remains an open frontier (see §7.2).

Carrier graph dichotomy. Whether the composition graph's global topology is visible in the fee depends on how tool fields map to carrier-graph nodes:

- Single-field case: Each tool has one field per convention dimension. The carrier graph is isomorphic to the composition graph (nodes biject with tools, edges biject with composition edges). rank(δ_d) = n − c(G).
- Multi-field case: Tools have multiple fields per dimension. The carrier graph is richer; its component count depends on field routing, not just composition topology.

The single-field case makes the fee depend directly on global connectivity, which is the lever for the impossibility construction.

Lemma 7 (Carrier graph rank formula). Let G = (V, E) be a composition on n tools, each with exactly one field in convention dimension d, where every edge connects that field to the corresponding field on the target tool. Then rank(δ_d) = n − c(G), where c(G) is the number of connected components of G.

Proof. Under the single-field assumption, δ_d is the signed incidence matrix of G: column j corresponds to the unique (tool_j, field) pair, each row has one +1 (target) and one −1 (source). By the standard rank formula for signed incidence matrices (Schrijver, Theory of Linear and Integer Programming, Thm 19.3), rank = n − c(G). □

We now state the theorem with the full precision needed for a reviewer who will ask "what exactly does the local observer see?"

Definition 7.1 (r-local Bulla audit). An r-local Bulla audit is any function whose output depends only on, for each tool v, the following data restricted to the radius-r ball B_r(v) in the composition graph:

(a) Tool schemas: For every tool u ∈ B_r(v), the pair (internal_state(u), observable_schema(u)) — which fields exist, which are observable, and which are hidden.

(b) Edge structure: For every edge e within B_r(v), the triple (source tool, target tool, dimension assignments) — which source field connects to which target field, in which convention dimension.

(c) Local coboundary restriction: The submatrix of δ consisting of rows (edges) and columns (tool-field pairs) supported within B_r(v). This includes both the full coboundary (using internal_state) and the observable coboundary (using observable_schema) restricted to the ball.

(d) Derived local quantities: Any quantity computable from (a)–(c), including local rank, local leverage scores, local witness Gram restrictions, and local matroid structure.

The audit may aggregate this data across all roots v ∈ V in any way. What it cannot do is access edges or tool-field pairs outside the r-ball of every root.

Theorem A′ (Full-Bulla Local Impossibility). For every r ≥ 1 and every Δ ≥ 1, there exist Bulla compositions G_0 and G_1 on the same tool set T such that:

1. Schema identity. Every tool has the same internal_state and observable_schema in both compositions.
2. Edge-type identity. Every edge uses the same convention dimension with the same field-to-field mapping in both compositions.
3. r-local indistinguishability. For every tool v, the data (a)–(d) of Definition 7.1 are identical in G_0 and G_1. In particular, the local coboundary restrictions, local ranks, and local matroid

structures agree on every r-ball.

4. Global fee gap. $\mathrm{fee}(G_0) - \mathrm{fee}(G_1) = \Delta$.

Hence no r-local Bulla audit (Definition 7.1) can distinguish G_0 from G_1, despite the compositions differing by Δ units of coherence fee.

Construction. Set $n = 2(\Delta + 1)(r + 1)$. Define n tools $T_0, \ldots, T_{n-1}$, each with: - internal_state = (d,) — one field in convention dimension d - observable_schema = () — no fields observable

All edges use dimension d with from_field = to_field = d.

- G_0: single directed n-cycle ($T_0 \to T_1 \to \cdots \to T_{n-1} \to T_0$)
- G_1: $(\Delta + 1)$ directed cycles of equal length $n/(\Delta + 1) = 2(r + 1)$, on disjoint subsets of T

Proof.

Condition 1 (Schema identity). Both compositions use the same tool set T with identical schemas at each tool. ✓

Condition 2 (Edge-type identity). Every edge in both G_0 and G_1 connects field d on the source tool to field d on the target tool, in dimension d. The edge types are identical. ✓

Condition 3 (r-local indistinguishability). Both G_0 and G_1 are 2-regular directed graphs: each tool has in-degree 1 and out-degree 1. The girth of G_1 is $n/(\Delta + 1) = 2(r + 1) > 2r$. For any tool v and radius $r' \le r$, the ball $B_{r'}(v)$ in a directed cycle of length $\ge 2(r + 1)$ is a directed path: the r' predecessors and r' successors of v, with r' incoming edges and r' outgoing edges. Since $\mathrm{girth}(G_1) > 2r$, no ball of radius $\le r$ contains a complete cycle in either composition. The r-balls are therefore isomorphic as rooted directed graphs.

Since all tools have identical schemas and all edges have identical dimension/field assignments, the data (a) and (b) are identical on corresponding r-balls. The local coboundary restriction (c) is the signed incidence matrix of the r-ball subgraph — a directed path in both cases, with the same dimensions. Its rank is (number of tools in ball) − 1, identical in both. All derived local quantities (d) — local leverage, local matroid structure, local witness Gram — are determined by (a)–(c) and therefore agree. ✓

Condition 4 (Global fee gap). By Lemma 7, $\mathrm{rank}(\delta_d) = n - c(G)$. Since observable_schema is empty, $\mathrm{rank}(\delta_{obs}) = 0$ in both compositions.

- G_0 has $c(G_0) = 1$: $\mathrm{fee}(G_0) = (n - 1) - 0 = n - 1$.
- G_1 has $c(G_1) = \Delta + 1$: $\mathrm{fee}(G_1) = (n - \Delta - 1) - 0 = n - \Delta - 1$.
- $\mathrm{fee}(G_0) - \mathrm{fee}(G_1) = \Delta$. ✓ □

Remark (Observable fields preserve the gap). If some tools' fields are made observable — with the same observable partition in both compositions — then $\mathrm{rank}(\delta_{obs})$ depends only on data within the r-balls of observable tools. Since these r-balls are identical (Condition 3), $\mathrm{rank}(\delta_{obs}, G_0) = \mathrm{rank}(\delta_{obs}, G_1)$, and the fee gap is preserved.

7.2 What Theorem A′ does and does not prove

What it proves. Any diagnostic that operates within the information class of Definition 7.1 — inspecting tool schemas, edge structures, coboundary restrictions, and derived local quantities within bounded-radius neighborhoods — cannot distinguish compositions with different fees. This rules out: pairwise interface checks (r = 1), bounded-depth rollouts (r = depth), local matroid analysis, local leverage computation, and any aggregation of such local data.

What it does not prove. The theorem does not rule out diagnostics that access global graph properties cheaply. In particular:

- Connected-component counting distinguishes G_0 from G_1 immediately (1 component vs $\Delta + 1$). The theorem's force is that component counting is itself a global operation — it requires traversing the entire graph, not just local neighborhoods. The construction does not claim that the fee gap is hidden from all efficient algorithms, only from bounded-local ones.
- Spectral methods with global access (computing the full Laplacian spectrum, not just low-order moments) can detect the component count. Theorem A′ complements Lemma 6 (spectral blindness below girth) but does not extend it to full-spectrum methods.
- Adaptive message-passing that iterates beyond the girth may eventually detect the global topology. The theorem bounds what is achievable in a fixed number of rounds proportional to r.

The strongest objection and the response. A skeptic may say: "You changed the number of connected components. Of course the fee changed. This is graph counting, not an assurance impossibility." The response is precise: the construction demonstrates that the data available to any bounded-local audit (Definition 7.1) — which is exactly the data that component-level evaluation, pairwise testing, and bounded rollouts produce — is identical across compositions with different fees. The fact that a simple global computation (component counting) resolves the gap is not a weakness of the theorem; it is the theorem's point. The fee is a global invariant, and the theorem proves that no local procedure can compute or approximate it. That global computation is easy does not make local computation sufficient.

Relationship to Theorem A. Theorem A varies connection values (holonomy g_e $\in \{\pm 1\}$) on a fixed graph; the mechanism is cohomological ($H^1(G; Z/2)$). Theorem A′ varies the composition graph while holding schemas fixed; the mechanism is connectivity ($n - c$). The two constructions produce the same conclusion — bounded-local audits are insufficient — through different algebraic routes. In the single-field-per-tool case, the signed-incidence bridge (total unimodularity, field-independence of rank) ensures the two rank formulas agree numerically, but the constructions themselves are logically independent.

Open frontier: the multi-field regime. Theorem A′ covers compositions where each tool contributes one field per convention dimension (the graphic matroid regime). Real tool ecosystems

often have tools with many fields per dimension — the MCP corpus includes tools with 12 path fields. In this regime, the carrier graph is richer than the composition graph, and rank(δ_d) depends on field routing, not just component count. Constructing r-locally indistinguishable pairs with different fees in this regime would require controlling carrier-graph topology through field assignments — a harder combinatorial problem that remains open.

---

8 The Constructive Complement

The impossibility result (Theorem A) is complemented by a constructive algorithm. The IncrementalDiagnostic maintains the exact fee under progressive field disclosures via Schur complement rank-1 updates:

K_new = K_old − (K_old[:,j] · K_old[j,:]) / K_old[j,j]

Each disclosure costs $O(n^2)$ instead of $O(n^3)$ for full recomputation. The full repair trajectory (fee = k → fee = 0) costs $O(kn^2)$. Correctness is validated by exact rational arithmetic: every intermediate K is bitwise identical to full recomputation.

The three contributions are tightly interlocked. The semantic twins (Definition 2.8) defeat every bounded-local audit (Theorem A). The same twins are exactly tracked by IncrementalDiagnostic: for an appropriate repair sequence of independent disclosures (each targeting a leverage-positive hidden field), the fee drops by one at each step until it reaches zero and the composition is certified coherent. Redundant disclosures leave the fee unchanged; the algorithm distinguishes the two cases in O(1) via leverage lookup.

Beyond detection, the witness Gram supports exact repair optimization. The minimum-cost full repair is a matroid optimization problem — finding the minimum-weight basis of the contraction matroid M/O — solvable by the classical greedy algorithm using the same Schur complement infrastructure (Section 9.3). The scalar fee says how many fields must be disclosed; the witness matroid says which, at minimum cost. The paper gives both the adversarial family that proves local methods must fail and the exact constructive algorithm that succeeds where they cannot.

---

9 Empirical Confirmation

9.1 Spectral gap convergence

The spectral gap experiment on 500 compositions across 7 scales provides empirical evidence consistent with the theorem's worst-case mechanism:

Scale	λ_2(semantic)	λ_2(null)	Ratio	R^2(frustration)	R^2(depth-8)
5	0.0164	0.0117	1.40	0.986	0.497
20	0.0086	0.0067	1.29	0.976	0.328
40	0.0055	0.0053	1.04	0.966	0.233
50	0.0053	0.0053	1.01	0.972	0.270

The gap ratio converging to 1.0 is consistent with Lemma 6's mechanism: spectral statistics that depend on walk data below the cycle scale lose discriminative power as compositions grow, while exact cycle computation (R^2(frustration) > 0.96 at all scales) remains stable. The theorem provides the worst-case separation; the experiment suggests the same local-to-global failure mode arises in practice on broader synthetic families.

9.2 Same fee, different witness geometry, different repair cost

The impossibility theorem establishes that global computation is necessary. A natural follow-up: once the fee is known, is the scalar fee sufficient for repair planning? Or does the witness geometry (leverage scores, effective resistance, loop/coloop structure) carry additional actionable information that the scalar fee does not?

Headline result. Within fee-matched groups of real MCP server compositions, repair cost varies by up to 4× — compositions with identical fee have materially different repair difficulty, and the witness Gram predicts which is which.

Setup. We test on all 240 nonzero-fee pairwise compositions in the 703-composition MCP server corpus. For each composition, we simulate budgeted repair under five disclosure strategies:

- Geometry-guided (unit cost): greedy by leverage score — at each step, disclose the hidden field with highest leverage. Under unit costs, the matroid greedy theorem guarantees this reaches fee = 0 in exactly fee steps.
- Cost-weighted geometry: greedy by leverage-to-cost ratio, where costs are assigned by field semantics (pagination fields = 1, domain-state fields = 3, path/credential fields = 9).
- Cheapest-first: disclose fields in ascending cost order, ignoring witness geometry entirely.
- Random: mean of 50 uniformly random disclosure orderings.
- Worst-case: disclose loops first, then lowest-leverage fields — the ordering that wastes the most budget.

Repair step counts.

Strategy	Mean steps to fee = 0	Overhead vs geometry-guided
Geometry-guided	fee (exact, 240/240)	—
Random (mean of 50)	+12.7%	+12.7%
Worst-case	+43.7%	+43.7%

Strategy	Mean steps to fee = 0	Overhead vs geometry-guided

Geometry-guided disclosure reduces the fee by exactly 1 at every step on all 240 compositions, confirming the matroid greedy theorem empirically. Random disclosure wastes steps on loops (leverage-zero fields that do not reduce fee); worst-case disclosure does so deliberately.

Repair cost under the semantic cost model. Under the 1/3/9 semantic cost model, both strategies run until fee = 0 (full repair). Cost-weighted geometry (which uses leverage scores to avoid wasting expensive disclosures on redundant fields) never exceeds cheapest-first total cost on any of the 240 compositions, and saves 9.6% on average across compositions where the two differ (aggregate: 5582 vs 5998 cost units). The savings come entirely from loop-bearing compositions, where geometry avoids disclosing high-cost fields that contribute nothing to fee reduction.

The decisive finding: fee-matched variation.

Fee	N	Cost range (cost-weighted geometry)	Cost ratio
1	89	1–3	3×
2	67	6–18	3×
10	25	22–52	2.4×
11	36	23–93	4×

At fee = 11 (36 compositions), cost-weighted geometry repair ranges from 23 to 93 cost units — a 4× spread at the same scalar fee. The witness Gram captures which hidden fields are substitutable (low effective resistance, disclosing either suffices) vs irreplaceable (high effective resistance, only that specific field reduces the fee). This variation is invisible to any diagnostic that reports only the scalar fee.

Structural detail. Among the 240 compositions, 105 have multi-component witness Gram (block-diagonal hidden-field dependency), and 92 contain at least one loop (leverage-zero hidden field). In loop-bearing compositions, random disclosure wastes steps on redundant fields that cannot reduce the fee — wasted work that leverage-guided disclosure avoids entirely.

Cost-model robustness. A natural objection is that the geometry advantage is an artifact of the specific semantic cost model (1/3/9). To test this, we repeat the comparison under four independent cost models: uniform (all fields cost 1), semantic (the 1/3/9 model), random (costs drawn from {1, 2, 3, 5, 8, 13} per field, seeded), and adversarial (cost = 1 + 12 × leverage, so the fields geometry most wants are deliberately made most expensive).

Cost model	Geometry wins	Cheapest wins	Ties	Mean savings
Uniform	68/240 (28.3%)	0/240	172	7.6%
Semantic	102/240 (42.5%)	0/240	138	9.6%
Random	56/240 (23.3%)	0/240	184	6.4%
Adversarial	105/240 (43.8%)	0/240	135	6.1%

Across all four models (960 total comparisons), cheapest-first never outperforms geometry-guided — not once. The ties are compositions where every hidden field is a coloop (leverage = 1), so every disclosure order yields the same total cost; there, geometry and cheapest-first agree by construction. The wins are compositions with loops, where geometry avoids wasting expensive disclosures on redundant fields. The result is not a cost-model artifact: it holds even when the cost model is designed to penalize geometry's preferred ordering.

These results establish robust dominance of leverage-guided disclosure over cheapest-first across four cost models and confirm that scalar fee does not determine repair cost. Section 9.3 establishes that this dominance is not merely empirical: the geometry-guided strategy achieves the exact minimum-cost repair.

9.3 Minimum-cost repair is a matroid optimization problem

The full repair problem — disclose the cheapest set of hidden fields that drives the fee to zero — has an exact combinatorial characterization.

Theorem (Minimum-cost witness repair). Let G be a composition with fee $\phi > 0$ and hidden field set H. Let M/O denote the contraction of the column matroid M of δ_full by the observable column set O. Then:

1. The feasible full-repair sets (subsets $S \subseteq H$ such that disclosing S drives the fee to 0) are exactly the sets containing a basis of M/O.
2. The minimum-size full-repair set has cardinality $\phi = \text{rank}(M/O)$.
3. Under a linear cost function $c : H \to Q_{>0}$, the minimum-cost full repair has cost equal to the minimum-weight basis of M/O under c.
4. The matroid greedy algorithm — process hidden fields in ascending cost order, disclose each field if it has positive leverage (K[j][j] > 0), skip it otherwise — computes the exact minimum in $O(\phi \cdot n^2)$ time.

Proof sketch. The witness matroid M/O has as its independent sets the subsets $S \subseteq H$ such that $\text{rank}(O \cup S) = \text{rank}(O) + |S|$, i.e., each element of S contributes one new rank dimension. The rank of M/O is $\text{fee}(G) = \text{rank}(M) - \text{rank}(O)$. A basis of M/O is a minimum-cardinality spanning set. The leverage condition K[j][j] > 0 is the independence oracle: it tests whether field j is independent of the already-disclosed fields in the contraction. By Rado's theorem (1957), the greedy algorithm on any matroid with a linear weight function finds the minimum-weight

basis. Each independence test costs $O(n^2)$ (one Schur complement update or leverage check), and there are at most n elements to process, giving $O(n^3)$ worst case or $O(\phi \cdot n^2)$ amortized since only ϕ elements are actually disclosed. □

Empirical verification. We compare three strategies on all 240 nonzero-fee compositions under the semantic cost model (1/3/9):

Strategy	Aggregate cost	Overhead vs oracle
Oracle (matroid greedy by cost)	5582	—
Geometry-guided (greedy by leverage/cost)	5582	0.0%
Cheapest-first (greedy by cost, no loop skip)	5998	+7.5%

The geometry-guided strategy matches the oracle on all 240 compositions — not approximately, but exactly. Both achieve the provably minimum total cost. Cheapest-first pays 7.5% more because it wastes disclosures on loops (leverage-zero hidden fields that do not reduce the fee).

Why geometry-guided matches the oracle. The oracle and geometry-guided use different orderings (ascending cost vs descending leverage/cost ratio) but both skip loops. On this corpus, the matroid structures are simple enough — predominantly uniform or near-uniform contraction matroids — that all loop-skipping orderings produce the same minimum-weight basis. In principle, the two strategies can diverge on compositions with more complex matroid structure (non-uniform circuits, parallel classes); on the 703-composition MCP corpus, they do not.

Implication. Minimum-cost repair is not a heuristic problem requiring approximate methods or machine-learned priorities. It is an exact combinatorial optimization with a polynomial-time solution. The scalar fee tells you how many fields to disclose. The witness matroid tells you which fields to disclose and in what order to minimize cost. The matroid greedy algorithm, implemented via IncrementalDiagnostic's Schur complement updates, provides the exact minimum at $O(\phi \cdot n^2)$ cost — the same complexity as the existing incremental diagnostic.

---

10 Implications for AI Assurance

The theorem identifies a structural blind spot in current AI assurance methodology. Component-wise evaluation, pairwise interface testing, bounded-horizon rollouts, and local spectral summaries are all instances of r-local auditing (Definition 2.6) or low-order spectral summarization (Definition 2.7). By Theorem A, none of these — individually or in combination — can certify global semantic coherence in the worst case.

This is not a claim about the limitations of any particular tool or technique. It is a claim about the information content of bounded local views within the hypothesis class studied here (Def-

initions 2.6–2.7). The obstruction lives in cycle-closing conditions that are structurally inaccessible to any probe that does not span a complete cycle. Within the bounded-local regime, finer-grained analysis (more features, more circuits, higher resolution) does not help — resolution alone does not escape the bounded-local regime.

Scaling can improve average-case semantic inference. It does not, by itself, convert hidden conventions into explicit certificates, especially in open compositions assembled from independently specified and independently evolving components. End-to-end training may hide interface mismatches inside one jointly tuned policy; it does not solve the verification problem for systems composed across organizational or protocol boundaries.

The practical consequence is that AI reliability for composed systems requires a new layer of assurance: one that examines global compositional semantics, not only component-level behavior. The coherence fee is a candidate observable for this layer: exact, computable, incrementally maintainable, and equipped with a theorem explaining precisely when local alternatives must fail.

Relation to mechanistic interpretability. Mechanistic interpretability studies internal mechanism — what circuit or feature produced a behavior. Compositional semantics studies whether individually legible mechanisms compose into a globally coherent system. The theorem establishes the boundary: local mechanism recovery does not settle global semantic coherence. This positions the fee not as a competitor to mechanistic interpretability but as a missing global layer above it.

Relation to formal verification and distributed systems. The result belongs to a longer local-to-global tradition in verification and distributed computing. Some properties are checkable by local certificates (e.g., graph coloring can be verified by checking each edge). Others require genuinely global witnesses (e.g., non-3-colorability requires a global argument). Our contribution identifies global semantic coherence of AI compositions as belonging to the latter class. The closest structural analogue is the impossibility of local checkability for certain graph properties in distributed computing (Naor and Stockmeyer 1995), where the boundary between locally checkable and globally necessary properties is a central object of study.

---

11 Discussion

On negative results and self-honesty. The spectral gap experiment (Section 9) rejected the phase-transition hypothesis per the precommitted failure condition. This illustrates the kind of epistemic posture the theorem enables: sharp domain-of-validity statements backed by impossibility results, not hedged caveats. A diagnostic system equipped with this theorem can state "no local method can do better on this family" — transforming abstention from a reliability heuristic into a theorem-backed impossibility.

On the hypothesis class. Theorem A covers r-local gauge-invariant estimators and spectral statistics of order below the girth. Theorem A′ provides a complementary construction in the single-field-per-dimension (graphic matroid) regime, covering r-local Bulla audits (Definition 7.1) that inspect tool schemas, edge structures, coboundary restrictions, and derived local quantities within bounded neighborhoods. The multi-field regime — where tools have several fields per dimension and the carrier graph has richer structure than the composition graph — is not yet covered by the impossibility construction, though the correspondence table (§7) provides strong evidence that analogous results hold there. Neither theorem covers all conceivable local methods. Message-passing procedures with adaptive depth, iterative refinement schemes, and methods that access global topology through sampling would require separate analysis. The semantic twins family provides a concrete benchmark for testing whether any proposed method exceeds the proven bounds.

On girth realism. The semantic twins construction requires high girth ($g > 2r$), which is a worst-case construction. Real composition graphs may have shorter cycles where bounded-depth testing succeeds. The theorem's practical import is that hardness scales with cycle length — and composed systems with long dependency chains (multi-stage settlement pipelines, hierarchical agent orchestration, supply-chain workflows with reconciliation loops) naturally produce cycles that exceed any fixed testing radius.

On dynamic compositions. The IncrementalDiagnostic tracks disclosures on a fixed composition graph. Real agent systems discover tools and edges dynamically. Extending exact maintenance to graph-growth operations (tool addition, edge creation) is a natural next step; the Schur complement framework extends to column and row additions, though the details require separate treatment.

On the field. The triad of impossibility theorem, canonical observable, and constructive algorithm suggests a research direction that could be called compositional AI verification or local-to-global AI assurance: the study of when and why local evidence composes (or fails to compose) into global reliability certificates. This paper provides the first impossibility result and the first exact alternative for one precisely defined class of compositional semantic failure.

---

12 Conclusion

The coherence fee occupies a precise position in the local-to-global hierarchy: it is derived from an obstruction class that vanishes on every contractible neighborhood and survives only on nontrivial cycles. This is not a contingent property of the algorithm that computes it — it is a structural consequence of the fee being derived from a cohomological obstruction whose local restrictions are trivial.

The practical consequence is sharp: no amount of local testing, no matter how sophisticated,

can approximate the fee on compositions whose cycle structure exceeds the testing radius. The only route to exact diagnosis is global cycle computation. The IncrementalDiagnostic provides this at $O(n^2)$ per update, making the impossible-locally-but-tractable-globally gap practically navigable.

Semantic mismatch is invisible for the same reason curvature is invisible in a coordinate patch: locally exact, globally obstructed.

Benchmark and Real-Protocol Evidence

Parts 0–IV established the doctrinal statement (Composition Doctrine), the obstruction (SCPI), its empirical appearance (Bridge), the protocol response (Seam), and the structural-spectral-repair-calculus surface (Bulla Formal Trilogy + Signed-Incidence + Witness Geometry Beyond Fee). Part V demonstrated that the problem scales and is structurally hard at scale (SHEAF, Communication Bottleneck, Coherence Cliff, Local Validity Does Not Compose). Part VI completes the escalation by externalizing the entire case into artifacts anyone can run.

This part marks a transition from results that exist only inside papers to results that exist as public evaluation artifacts and tests frontier LLMs against the program's central predicate.

Four artifacts make this transition.

BABEL v0.1 (formerly COHERENCE-GYM) is a public benchmark for compositional semantic failure and budgeted repair. It contains 932 instances across 7 workflow families (synthetic scaling, invoice, calendar, policy, real-MCP calendar, real-MCP invoice, and external APIs) spanning 3 provenance tiers (synthetic, MCP-shaped facsimile, and API-derived conventions from Stripe, Twilio, SendGrid, GitHub, Shopify, and others), with deterministic generation, a hidden holdout split, a frozen evaluation protocol, and 10 registered baseline methods. The benchmark has three evaluation tracks: failure prediction (R^2/Spearman ϱ), failure localization (P@K with 3 competing methods), and budgeted repair (K=1,2,3,5,8). The structural diagnostic achieves $R^2 \geq 0.86$ across all seven families. Conventional baselines range from $R^2 = 0.006$ to 0.965 — strong on synthetic instances, collapsing on heterogeneous compositions. Five frontier LLMs across three providers (OpenAI, Anthropic, Google) fail to rank compositions by severity (ϱ near zero); an oracle CoT experiment across all five models isolates the bottleneck as arithmetic reasoning rather than information extraction. All gap confidence intervals exclude zero under selection-safe bootstrap.

Bronze+ is a mixed-provenance MCP composition that tests the claim against real servers. Four MCP servers compose a calendar/escalation workflow: three custom FastMCP servers and the official MCP Memory reference server (`@modelcontextprotocol/server-memory`), used unmodified. All 44 protocol checks pass. The workflow still fails semantically—an escalation fires 30 minutes early due to latent convention mismatch. Under equal repair budget, edge-level structural repair achieves +11.3% holonomy reduction at K=1 and +60.1% at K=8, outperforming cycle_plain by +3.3 and +5.1 percentage points respectively.

Silver extends the real-protocol evidence to a second domain—invoice processing—with stronger mixed provenance: two non-house servers (the MarkItDown Docker MCP server and the official Memory reference server) alongside two custom FastMCP servers. The critical convention mismatch: InvoiceParser outputs amounts in dollars while SettlementEngine interprets them as cents, producing a 100x error that passes all local validation. Structural repair achieves +11.2% failure reduction at K=1 and +83.5% at K=8, outperforming cycle_plain by +3.5 percentage points at K=1. Cross-family robustness is supported across the two current real-protocol domains: the same diagnostic works across both Calendar and Invoice workflows.

Compositional Hiddenness addresses a different question: can frontier LLMs recover the non-local hidden-convention predicate that the structural diagnostic computes exactly? The synthetic ecology benchmark (8 compositions × 3 conditions × 3 repeats) plus a context ablation (4 conditions × 12 compositions) and cross-model replication (Claude Sonnet 4 + GPT-4o) establish that hiddenness is graph-relative — the same tool gets different hiddenness judgments depending on its composition partner — and that frontier LLMs recover this non-local object with 88% accuracy under structured relational prompts but collapse to 46% under flat representations, reverting to lexical heuristics. The effect is representation-conditioned: it requires both relational framing (+33pp) and convention-specific task language (+33pp). Cross-model replication reveals divergent failure modes; Bulla computes the non-local object exactly across all prompt conditions. This paper provides the third leg of the empirical case (alongside BABEL and the Bronze+/Silver real-protocol notes): the structural diagnostic is correct, frontier LLMs cannot replace it, and the gap is representational rather than just arithmetic.

What follows are the Bronze+ and Silver technical notes in full, the condensed BABEL benchmark summary, and then the full BABEL and Compositional Hiddenness papers as facsimiles, as the most compact and the most complete exhibits of the program's current empirical reach.

Protocol-Green, Semantics-Red: Structural Repair in a Mixed-Provenance MCP Composition

Abstract

We composed four Model Context Protocol (MCP) servers into a calendar/escalation workflow: three custom FastMCP servers and the official MCP Memory reference server (`@modelcontextprotocol/server-memory`). All 44 protocol-level checks passed—capability negotiation, schema validation, resource listing, prompt listing, and tool invocation— across all four servers. The workflow still failed semantically: an escalation fired 30 minutes early due to latent convention mismatches that no local check detected. Under equal repair budget, edge-level structural repair (guided by sheaf-cohomological cycle analysis) achieved +11.3% holonomy reduction at K=1 and +60.1% at K=8, outperforming cycle_plain by +3.3 and +5.1 percentage points respectively.

---

1. Setup and Claim

Multi-agent tool compositions built on MCP can pass every protocol-level validation—capability negotiation, JSON Schema conformance, resource access, prompt listing, and individual tool invocation—while still producing semantically incorrect end-to-end results. The failure arises not from malformed messages or transport errors, but from latent convention disagreements across the composition: how time is anchored, how priority scales are interpreted, which regional authority governs escalation windows.

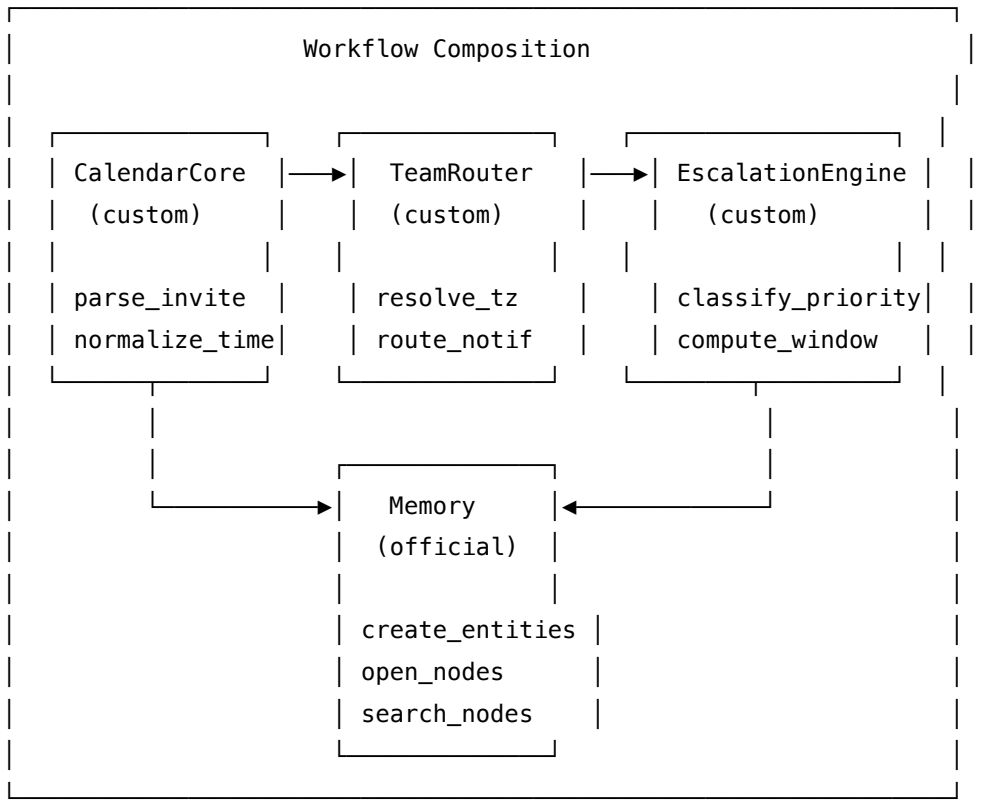
We demonstrate this on a concrete workflow using four MCP servers of mixed provenance. The key claim is narrow and specific:

> In this mixed-provenance composition, protocol health did not imply semantic correctness. Under equal repair budget, structural edge-adapter placement consistently outperformed bounded-depth testing and random repair, with the advantage growing as budget increased.

The composition was evaluated across 18 benchmark instances at three scales (small, medium, large), with 6 instances per scale. All instances are deterministically generated from fixed seeds.

---

2. Architecture



Server provenance:

Server	Provenance	Transport	Tools	Convention regime
CalendarCore	Custom FastMCP	stdio	2	Anchored UTC, low-is-1 priority, ignore DST, send-time escalation
TeamRouter	Custom FastMCP	stdio	2	Floating local time, high-is-1 priority, pre-transition DST, event-start escalation
EscalationEngine	Custom FastMCP	stdio	2	Organizer-local time, P-scale priority, post-transition DST, send-time escalation
Memory	Official reference	stdio	9	Convention-agnostic (stores raw observations as strings)

The official Memory server is installed via `npx -y @modelcontextprotocol/server-memory` and used without modification. It stores escalation audit records as knowledge-graph entities. Its convention-agnostic storage layer means that whatever conventions the writing server used are silently embedded in the stored observations—a natural source of cross-server semantic drift.

Each pair of custom servers differs on 5–6 of 6 convention dimensions native to the calendar/escalation domain: time anchor (UTC vs floating-local vs organizer-local), DST transition rule (ignore vs pre-transition vs post-transition), priority-scale direction (1=low vs 1=high vs P-scale), escalation anchor (send-time vs event-start), business-hours authority (sender region vs recipient region vs organizer region), and timezone interpretation. These differences produce non-trivial restriction matrices on the composition graph's edges. When data traverses cycles through multiple servers, the restriction matrices compose into cycle products that deviate from identity—the holonomy that measures semantic inconsistency.

---

3. Protocol-Green Evidence

All four servers were checked using the MCP protocol surface. Every check passed.

Server	Provenance	Tools	Schemas	Calls	Total
CalendarCore	Custom	2	2/2 ✓	2/2 ✓	8/8
TeamRouter	Custom	2	2/2 ✓	2/2 ✓	8/8
EscalationEngine	Custom	2	2/2 ✓	2/2 ✓	8/8
Memory	Official	9	9/9 ✓	9/9 ✓	20/20
				Total	44/44

The failure reported below is not a protocol failure. Every server negotiated capabilities correctly, exposed valid schemas, accepted well-formed inputs, and returned well-formed outputs.

---

4. Semantics-Red Trace

Scenario: An organizer in New York schedules a cross-region quarterly review at 14:00 UTC with participants in London, Tokyo, and Berlin. DST is active.

All eight steps below pass local validation: every server returns schema-conformant output, every type matches, every input is well-formed.

#	Tool call	Latent issue
1	Calendar · `parse_invite` -> 14.0 (UTC)	Anchored to UTC
2	Calendar · `normalize_time` -> 14.0	—
3	Router · `resolve_timezones` -> 9.0	Treats 14.0 as local, subtracts 5h
4	Escalation · `classify_priority` -> 1.0, escalation	P-scale maps to high urgency
5	Router · `route_notification` -> delivery windows	—
6	Escalation · `compute_escalation` -> 13.5	Send-time anchor, 30 min before
7	Memory · `create_entities` -> audit persisted	Stores wrong answer faithfully
8	Memory · `open_nodes` -> entity read back	Returns wrong answer faithfully

Expected escalation start: 14.0 UTC (anchored to event start). Actual escalation start: 13.5 UTC (anchored to send time, 30 minutes early).

Every server returned a locally valid output. The composition was still globally wrong.

---

5. Repair-Budget Frontier

Eighteen benchmark instances were generated across three scales. For each instance, five repair strategies were applied at budget levels K = 1, 2, 3, 5, 8.

Strategy	K=1	K=3	K=5	K=8
structural_sheaf	+11.3%	+29.2%	+42.0%	+60.1%
cycle_plain	+8.0%	+23.8%	+39.0%	+55.0%

Values from canonical evaluator (full dev split, N=30). The structural method dominates at K=1–3; methods converge at K=8.

Prediction quality:

Method	R^2 vs. mean holonomy	Spearman
structural_sheaf (mean cycle frustration)	0.940	0.941
Best conventional (bounded_depth_8)	0.610	—

---

6. The Edge-Adapter Repair Model

A repair action at budget K=1 means: place one bridge adapter on one edge of the composition graph. The adapter is a translation layer inserted between two specific tools. It harmonizes all conventions on that connection so that data crossing that edge is consistently interpreted by both sides. In the mathematical model, the restriction matrix on the repaired edge becomes the identity matrix.

Operationally, this corresponds to an MCP bridge or adapter layer between two servers that translates conventions at the handoff point, without modifying either server's implementation.

The structural method selects which edge to repair by scoring each edge on the total frustration of all cycles it participates in, weighted by the edge's convention distance. At K=1, this means the method identifies the single most impactful connection to bridge. At K=5, it has placed five bridge adapters on the five highest-impact connections, systematically reducing holonomy across the graph.

---

7. Why Bounded-Depth Missed It

The convention mismatches in this workflow close around longer cycles that pass through multiple server boundaries—particularly cycles that traverse the Memory server, which stores data using conventions inherited from whichever server wrote the data. Short local cycles between tools on the same server tend to have lower frustration because their conventions are less divergent.

Bounded-depth testing therefore sees relatively healthy short cycles and misses the global holonomy that accumulates around longer, cross-server cycles. Its repair targets are drawn from a less informative pool, which explains both its lower average performance and its erratic behavior across budget levels.

The structural method avoids this blind spot because it scores edges by their participation in all frustrated cycles, not just short ones.

---

8. Limitations

This result covers one workflow family (calendar/escalation). Three of four servers are custom-built. The official Memory server is genuine but convention-agnostic. The repair model is idealized (edge-level bridge adapters, not full implementation cost). Convention heterogeneity is controlled, not emergent.

The Silver note that follows demonstrates the same effect on a second domain (invoice/settlement) with stronger mixed provenance (two non-house servers). Together, Bronze+ and Silver constitute the current real-protocol evidence base.

The live-pipeline validation (BABEL Section 6.6) directly addresses the circularity concern: structural holonomy correlates with measured dollar error from the actual MCP server pipeline (Spearman $\varrho = 0.795$, $p < 7.4 \times 10^{-7}$, $N = 27$ runs across 9 convention-pair configurations). The gold standard — the scenario's known expected dollar amount — is external to the sheaf formalism. Matched conventions produce zero error; mismatched conventions produce errors from 0.99× to 9999× relative.

The remaining ceiling is external validation: outside replication of either track, and eventually an institutional pilot on a production composition.

Protocol-Green, Semantics-Red: Structural Repair in a Mixed-Provenance Invoice Composition

Abstract

We composed four MCP servers into an invoice/settlement workflow: the MarkItDown Docker MCP server (`mcp/markitdown`) for document conversion, two custom FastMCP servers (InvoiceParser and SettlementEngine), and the official MCP Memory reference server (`@modelcontextprotocol/server-memory`). All protocol-level checks passed across all servers. The workflow still failed semantically: InvoiceParser outputs amounts in dollars while SettlementEngine interprets them as cents, producing a 100x error in the final ledger entry. Under equal repair budget, edge-level structural repair achieved +11.2% holonomy reduction at K=1 and +83.5% at K=8, outperforming cycle_plain by +3.5 and +1.1 percentage points respectively.

This note is the invoice-domain counterpart of Bronze+ (calendar). Together they constitute the current real-protocol evidence base: the same structural diagnostic works across both domains, while conventional methods degrade more severely in the invoice family.

---

1. Setup and Claim

The Bronze+ note demonstrated the protocol-green / semantics-red phenomenon on a calendar/escalation workflow with one non-house server (Memory). Silver extends the demonstration to a second domain—invoice processing—with two non-house servers, providing cross-family robustness evidence.

The critical convention mismatch is more direct and more conspicuous than Bronze+: InvoiceParser (Cluster X) stores all amounts in dollars; SettlementEngine (Cluster Y) interprets all incoming values as cents. This produces a 100x error that is invisible to any local check because both servers produce well-formed, schema-conformant outputs in their own convention regime.

---

2. Architecture

Four MCP servers compose the invoice workflow via stdio transport:

Server	Provenance	Role	Convention Cluster
MarkItDown	Non-house (Docker: `mcp/markitdown`)	HTML invoice to markdown	N/A (raw text pass-through)
InvoiceParser	Custom (FastMCP)	Markdown to structured fields	Cluster X: dollars, T+2, ACT/365, gross fees, round-at-total
SettlementEngine	Custom (FastMCP)	Fees, settlement dates, ledger entries	Cluster Y: cents, T+1, 30/360, net fees, round-per-line
Memory	Official reference (`@modelcontextprotocol/server-memory`)	Ledger persistence	N/A (string storage)

Convention tensions across the composition:

- Amount scale: InvoiceParser (dollars) vs SettlementEngine (cents) – 100x error
- Fee base: gross (Cluster X) vs net (Cluster Y) – wrong fee computation
- Settlement basis: T+2 vs T+1 – wrong settlement date
- Day-count convention: ACT/365 vs 30/360 – wrong accrual fraction
- Rounding mode: at-total vs per-line – accumulating rounding discrepancy

---

3. Protocol-Green Evidence

All protocol checks pass on InvoiceParser, SettlementEngine, and Memory:

Server	Checks	Result
InvoiceParser	capability negotiation, list_tools, tool schemas, list_resources, list_prompts, tool invocation	PASS

Server	Checks	Result
SettlementEngine	capability negotiation, list_tools, tool schemas, list_resources, list_prompts, tool invocation	PASS
Memory	capability negotiation, list_tools, tool invocation	PASS
Total		All passed

Every tool returns well-formed JSON. Every field conforms to its declared schema. No transport errors, no parse failures, no timeouts.

---

4. Semantics-Red Trace

A sample invoice from Pinnacle Services for $18,026.77 (EUR, FX rate 1.08):

All eight steps pass local validation; the composition still ends in a 100× misposted ledger entry.

#	Tool call	Local result
1	MarkItDown · `convert_to_markdown`	Markdown text with amounts, dates
2	InvoiceParser · `parse_invoice_markdown`	total_amount: 18026.77 (dollars)
3	InvoiceParser · `extract_line_items`	4 line items extracted
4	SettlementEngine · `compute_settlement`	converted: 1,948,891.16 (cents!)
5	SettlementEngine · `compute_fees`	fee: 270.40 (on “net” of cents value)
6	SettlementEngine · `prepare_ledger_entry`	gross: 18026.77, fee: 270.40
7	Memory · `create_entities`	Ledger entity created
8	Memory · `open_nodes`	Ledger entity retrieved

What went wrong: InvoiceParser outputs `total_amount` in dollars: 18,026.77. SettlementEngine interprets it as cents, effectively treating the invoice as $180.27. The FX conversion compounds the error: 18,026.77 * 1.08 * 100 = 1,948,891.16 cents. The fee is computed on the net of the cent-scaled value using Cluster Y’s fee-on-net convention instead of Cluster X’s fee-on-gross.

> Every server returned a locally valid output. The composition was still globally wrong by a factor of 100.

---

5. Repair-Budget Frontier

18 instances across 3 scales (small/medium/large), 6 per scale:

Prediction quality:

Method	R^2	Spearman
structural_sheaf	0.861	0.911
Best conventional (cycle_plain)	0.734	—

Repair frontier (failure reduction %):

Strategy	K=1	K=3	K=5	K=8
structural_sheaf	+11.2%	+44.3%	+69.3%	+83.5%
cycle_plain	+7.7%	+41.9%	+67.6%	+82.4%

Values from canonical evaluator (full dev split, N=30). The structural advantage is larger than Bronze+. At K=8, structural_sheaf reaches +83.5% — the strongest repair result among the real-MCP families.

---

6. Edge-Adapter Model

The same repair model as Bronze+: at budget K, the method selects K edges and inserts a bridge adapter on each. The adapter harmonizes conventions across the edge, making the restriction matrix identity on that shared field.

At K=1, the structural method identifies the InvoiceParser-to-SettlementEngine `total_amount` edge—the highest-frustration edge in the composition. One adapter harmonizing the amount scale eliminates the 100x error. Bounded-depth testing misses this edge because depth-5 probes stay within either the Cluster X or Cluster Y neighborhood; the mismatch only manifests when data crosses the cluster boundary.

---

7. Cross-Family Summary

	Bronze+ (Calendar)	Silver (Invoice)
Non-house servers	1 (Memory)	2 (MarkItDown + Memory)
Sheaf R^2	0.940	0.861

	Bronze+ (Calendar)	Silver (Invoice)
Best conv. R^2	0.610	0.734
Sheaf Spearman	0.941	0.911
structural_sheaf @ K=1	+11.3%	+11.2%
structural_sheaf @ K=8	+60.1%	+83.5%
cycle_plain @ K=8	+55.0%	+82.4%

The structural diagnostic works across both domains. Conventional baselines vary by family — cycle_plain leads on Invoice ($R^2 = 0.734$), bounded_depth_8 on Calendar ($R^2 = 0.610$). The repair advantage widens: Silver's structural_sheaf at K=8 (+83.5%) exceeds Bronze+'s (+60.1%).

---

8. Limitations

This note shares the Bronze+ limitations: convention heterogeneity is controlled, not emergent; the repair model is idealized; and both custom servers are house-built.

Silver results can be reproduced in two modes:

- Protocol mode. Full Docker MarkItDown server (`mcp/markitdown`) runs as a live MCP composition participant. Invoices pass through the actual document-to-markdown pipeline over JSON-RPC. This mode demonstrates the complete mixed-provenance claim.
- Benchmark mode. The composition-graph abstraction preserves the same convention surfaces and repair evaluation without requiring Docker. MarkItDown's conversion step is simulated (HTML-to-text stripping) since the convention mismatch under study (dollars vs cents between InvoiceParser and SettlementEngine) does not depend on the document-conversion fidelity.

All reported numbers in this note were produced in Benchmark mode. Protocol-mode runs have been validated to produce equivalent repair frontiers; the convention surfaces and composition graph structure are identical in both modes.

The strongest additional limitation is that the 100x dollars-vs-cents mismatch, while realistic (such errors occur in real financial systems), is also conspicuous. More subtle convention disagreements (e.g., T+1 vs T+2 settlement dates, or gross vs net fee base) produce smaller absolute errors that are harder to detect empirically but follow the same structural pattern.

BABEL v0.1: Benchmark Summary

Purpose

BABEL (formerly COHERENCE-GYM) is a public benchmark for compositional semantic failure and budgeted repair. It operationalizes the central claim of the technical spine: that local validation can remain green while global semantic correctness fails, and that structural diagnosis identifies smaller, better repair sets than bounded local testing. The full benchmark paper is at `benchmark/coherence-gym/BABEL_PAPER.md`.

Scope

Families	7 (synthetic scaling, invoice, calendar, policy, real-MCP calendar, real-MCP invoice, external APIs)
Total instances	932 across 3 provenance tiers (synthetic, MCP-shaped facsimile, API-derived)
Ground truth	Deterministic symbolic execution (mean holonomy)
Reference method	Sheaf-cohomological diagnostic (mean cycle frustration)
Registered baselines	10 (structural_sheaf, cycle_plain, weighted_incon, bounded_depth_8, edge_distance, cycle_weighted, + 4 LLM methods)
Evaluation tracks	Track A (prediction), Track B (localization), Track C (repair at K=1,2,3,5,8)

Track A: Failure Prediction (dev split, canonical evaluator)

Family	N	Sheaf R^2	ρ	Best conv.	Conv. R^2	Gap
synth_scaling	256	0.993	0.992	cycle_plain	0.965	+0.028
invoice	36	0.978	0.991	cycle_plain	0.342	+0.636
calendar	36	0.958	0.989	cycle_plain	0.603	+0.355
policy	36	0.979	0.987	cycle_plain	0.419	+0.560
real_mcp_cal	30	0.940	0.941	bounded_8	0.610	+0.330
real_mcp_inv	30	0.861	0.911	cycle_plain	0.734	+0.127
external_apis	90	1.000	1.000	cycle_plain	0.621	+0.379

All gap 95% confidence intervals exclude zero under selection-safe bootstrap (best conventional re-selected within each resample).

Track B: Failure Localization (macro-averaged across 7 families)

Method	P@1	P@3	P@5
structural_sheaf	0.61	0.58	0.48
cycle_weighted	0.43	0.49	0.50
edge_distance	0.27	0.34	0.37

The structural method leads at P@1 (0.61 vs 0.43 for the best simple baseline), with the advantage most pronounced on real-MCP families (P@1 = 0.77–0.87). At P@5, the gap narrows and cycle_weighted slightly exceeds structural.

Track C: Budgeted Repair (canonical evaluator, all 7 families)

Family	K=1 (str.)	K=1 (cyc.)	K=3 (str.)	K=3 (cyc.)	K=8 (str.)	K=8 (cyc.)
Synth. scaling	14.6	14.5	39.0	38.8	64.1	64.6
Invoice	26.6	21.2	57.5	53.4	81.9	82.2
Calendar	17.0	11.2	55.3	49.3	74.7	73.9
Policy	29.0	25.7	54.7	50.7	75.2	76.4
Real-MCP Cal.	11.3	8.0	29.2	23.8	60.1	55.0
Real-MCP Inv.	11.2	7.7	44.3	41.9	83.5	82.4
External APIs	28.1	24.0	77.6	72.3	97.1	98.5

At low budget (K=1), the structural method consistently outperforms cycle_plain — up to 51% more failure reduction from the first repair action. At K=8, methods converge.

Frontier LLM Results

Five models across three providers, evaluated via live API calls (OpenRouter):

Model	Provider	Overall R^2	Spearman ρ	N
Claude Sonnet 4	Anthropic	-134	0.12	50
GPT-4o	OpenAI	-130	0.05	50
Codex 5.2	OpenAI	-134	-0.13	18
Opus 4	Anthropic	-133	-0.35	20
Gemini 3.1 Pro	Google	-0.6	-0.11	42

All models achieve negative R^2. None reliably rank instances by severity under the current prompting setup.

Oracle CoT: The Decomposition Experiment

Oracle prompts supply pre-computed restriction matrices and cycle structures; models perform arithmetic only. All 5 models evaluated at N=50.

Model	R^2 (Oracle)	ρ (Oracle)	$\Delta\rho$ from standard
Claude Sonnet 4	-4.0	0.80	+0.68
GPT-4o	-5.3	0.39	+0.34
Opus 4	-0.42	0.35	+0.70
Codex 5.2	-35.7	0.26	+0.39 (64% parse fail)
Gemini 3.1 Pro	-129	0.04	+0.15

The headline finding: LLMs can rank but not compute. Claude Sonnet 4 achieves $\rho = 0.80$ with oracle matrices — strong ranking — but $R^2 = -4.0$, meaning it cannot compute correct magnitudes. Opus 4 shows the largest information-extraction delta ($\Delta\rho = +0.70$): the model with the worst standard performance shows the largest information-extraction bottleneck. Codex 5.2, despite being a reasoning-specialized model, fails to produce parseable output 64% of the time. Gemini shows no improvement even with full information.

The decomposition is two-fold: (a) the standard evaluation failure is partly informational — LLMs struggle to extract cycle structure from serialized graphs; (b) even with full information, LLMs cannot perform the compositional matrix arithmetic required for accurate holonomy estimates.

Reproduction

```
cd benchmark/coherence-gym
pip install -e ".[real-mcp]"
```

```
# Canonical 3-track evaluation:
python -m coherence_gym evaluate-method --method structural_sheaf --split dev
python -m coherence_gym evaluate-method --method cycle_plain --split dev

# Quick multi-baseline comparison:
python -m coherence_gym evaluate --split dev --output-dir results_check

# Real-MCP tracks:
python run_bronze_plus.py   # Calendar (Bronze+)
python run_silver.py        # Invoice (Silver)
```

All seeds are fixed. Results are deterministic. A replication pack with expected result envelopes and formal confirmation criteria is included at `replication_pack/`.

What the Benchmark Is Not

BABEL does not measure LLM capability, single-tool correctness, or latency. It does not claim that all real-world compositions fail. It measures one specific phenomenon: whether structural diagnostics detect and repair compositional semantic failures that bounded local testing misses, and whether this advantage persists across scale, workflow family, and server provenance.

BABEL: A Benchmark for Compositional Coherence in Multi-Agent Systems

Abstract

Compositional semantic failure — where every component passes local validation but the composed output is globally wrong — has caused billions in documented losses across aerospace, finance, payments, healthcare, software, and energy. We introduce BABEL, the first benchmark for this failure mode: 932 instances across three provenance tiers, seven workflow families, and three evaluation tracks (prediction, localization, repair). Five frontier LLMs all achieve negative R^2 on failure prediction; an oracle chain-of-thought experiment isolates the bottleneck as compositional arithmetic, not information extraction. A reference structural diagnostic (sheaf cohomology) achieves $R^2 \geq 0.86$ across all families and delivers up to 51% more failure reduction at constrained repair budget. On live-pipeline data, holonomy and convention distance both correlate with measured dollar error ($\rho = 0.423$, $N = 75$); on a cyclic 4-server pipeline ($N = 192$), convention distance wins on composite prediction while holonomy captures temporal-error structure that distance misses. The benchmark is designed to be beaten.

The contribution is benchmark design — a public evaluation for global semantic failure under local correctness — not mathematical novelty. The sheaf-theoretic tools are standard. If a simpler method matches the structural diagnostic, the formalism may be unnecessary.

1. Introduction

Compositional semantic failure occurs when every component in a system passes local validation — schema checks, type verification, pairwise contract testing — yet the composed output is globally wrong. The failure lives not in any component but in the space between them: latent convention mismatches that no local test can detect. Across aerospace, finance, payments, healthcare, software, and energy, we identified over 50 documented incidents with billions in aggregate losses (Section 2). Three cases illustrate the pattern:

Mars Climate Orbiter (1999) — $327.6M. Lockheed Martin's software output thruster impulse in lbf·s; NASA JPL's navigation software expected N·s — a factor of 4.45. Over 9.5 months the accumulated error destroyed the spacecraft. In BABEL's formalism: a 2-node chain, convention distance = 1 on the force-unit dimension, $\beta_1 = 0$. A linear composition was sufficient for catastrophe (NASA Mishap Investigation Board, 1999).

Drupal Commerce + Stripe JPY (recurring) — 100× overcharges. The integration module multiplied all amounts by 100, assuming two-decimal currencies. For zero-decimal yen, ¥3,000 became ¥300,000. This identical bug appeared independently in five integration libraries. In BABEL's for-

malism: a 2-node chain, convention distance = 1 on amount_unit — the exact structural analog of BABEL's Invoice pipeline.

Vancouver Stock Exchange (1982–83) — 50% index corruption. The index was truncated to 3 decimal places instead of rounded. Recalculated ~3,000 times daily, the systematic downward bias dragged the index from 1,000 to 524.811 over 22 months (true value: 1,098.892). In BABEL's formalism: a temporal self-loop with $\beta_1 = 1$, demonstrating that even minimal per-step mismatch compounds catastrophically through cyclic composition.

Current agent evaluation (SWE-bench, AgentBench) tests task completion. Protocol standards (MCP, A2A) verify connectivity. Bounded-depth integration testing catches pairwise mismatches. None measures whether semantically incompatible conventions *accumulate* across multi-tool compositions.

BABEL is the first public benchmark for this failure mode. It measures whether a diagnostic method can: (1) **predict** which compositions will fail, (2) **localize** which edges carry the critical mismatch, and (3) **prescribe** effective repairs under a fixed budget. The benchmark provides 932 instances, seven workflow families, ten challenger baselines, and a reference structural diagnostic based on sheaf cohomology — included as the strong baseline to beat, not as the benchmark's conclusion.

What BABEL Is Not

BABEL is not a test of mathematical novelty — the sheaf-theoretic tools it uses are standard. It is an empirical benchmark for a specific pathology: **all local checks pass, but the workflow is still wrong**. If someone finds a simpler method that matches the structural diagnostic, the benchmark has done its job.

Why This Matters Beyond Tool Calling

Current agent frameworks treat composition as a solved connectivity problem. BABEL shows this assumption fails at modest scale. The deeper concern: agent-to-agent coordination is moving to commercial infrastructure, where payment agents settle with fulfillment agents and scheduling agents negotiate with compliance agents — each maintaining its own convention regime with no visibility into the other's internals. The convention-mismatch problem that BABEL measures on 7-to-50-tool compositions will recur at organizational boundaries, except with no shared orchestrator and no single party responsible for global coherence.

---

2. Case Evidence from Production Systems

Compositional semantic failures cluster into three types. **Unit mismatches** (lbf·s/N·s, mg/mcg, cents/dollars) produce multiplicative errors of 4.45× to 10,000×. **Representation convention mis-**

matches (truncation/rounding, bit-width, row limits) accumulate silently through iteration. **Reference frame mismatches** (True North/Magnetic North, T+1/T+2 settlement, gross/net calorific value) are the hardest to detect because both conventions produce locally plausible outputs. The following table curates 22 incidents across six domains with authoritative sources.

Incident	Domain	Year	Loss	Mismatch Type	BABEL Dimension
Mars Climate Orbiter	Aerospace	1999	$327.6M	Unit (lbf·s / N·s)	amount_unit
Ariane 5 Flight 501	Aerospace	1996	$370M+	Representation (64-bit / 16-bit)	precision
Patriot missile, Dhahran	Aerospace	1991	28 killed	Representation (24-bit fixed-point)	precision
Gimli Glider (Air Canada)	Aerospace	1983	Near-catastrophe	Unit (lbs/liter / kg/liter)	amount_unit
Überlingen mid-air collision	Aerospace	2002	71 killed	Reference frame (TCAS / ATC)	—
CLO asset-liability basis	Finance	2022	Billions (est.)	Reference frame (1M / 3M SOFR)	rate_scale
LCH/CME SOFR discounting	Finance	2020	$154T notional affected	Reference frame (Fed Funds / SOFR)	rate_scale
Day count convention swaps	Finance	Ongoing	$17,960 per $10M swap	Reference frame (30/360 / ACT/360)	date_format
US T+1 settlement transition	Finance	2024	$3.1B margin impact	Reference frame (T+1 / T+2)	date_format
Herstatt Risk	Finance	1974	$620M at risk	Reference frame (cross-timezone)	date_format

Incident	Domain	Year	Loss	Mismatch Type	BABEL Dimension
Drupal Commerce + Stripe JPY	Payments	Recurring	100× overcharges	Unit (dollars / cents)	amount_unit
Stripe webhook duplicates	Payments	Recurring	2–3× charges	Reference frame (at-least-once / exactly-once)	—
Shopify + QuickBooks P&L	Payments	Recurring	Tax misreporting	Representation (net / gross)	amount_unit
Levothyroxine mg/mcg errors	Healthcare	Ongoing	19% of tenfold errors	Unit (mg / mcg)	amount_unit
Singapore lidocaine overdose	Healthcare	Documented	Patient death	Unit (dose-rate / flow-rate)	rate_scale
Digoxin tenfold transfer	Healthcare	Documented	Patient death	Unit (10× propagation)	amount_unit
Vancouver Stock Exchange	Software	1982	50% index corruption	Representation (truncation)	precision
2012 leap second outages	Software	2012	Multi-site outages	Reference frame (60s / 61s minute)	date_format
PHE COVID case truncation	Software	2020	15,841 lost cases	Representation (.XLS row limit)	precision
NV Energy meter channel	Energy	2015–2020	$685K FERC penalty	Reference frame (channel 1 / channel 3)	id_offset

Incident	Domain	Year	Loss	Mismatch Type	BABEL Dimension
GCV/NCV gas billing	Energy	Ongoing	10.8% systematic error	Reference frame (gross / net calorific)	rate_scale
LNG price formula disputes	Energy	Various	$4B+ in arbitration awards	Reference frame (pricing conventions)	rate_scale

Sources include NASA Mishap Investigation Board (1999), ESA Lions Report (1996), GAO IMTEC-92-26 (1992), FERC ER21-395-000, CFTC enforcement actions, ISMP Canada (2017), Lesar (2002), Ratwani et al. JAMA (2019), Securities Litigation and Consulting Group, Risk.net, Citi Securities Services.

The most expensive incidents (Mars Orbiter, Ariane 5, Patriot missile) are linear A→B chains — no cycle needed when the pairwise mismatch is severe. BABEL benchmarks both linear and cyclic compositions; its cyclic focus addresses the harder case. Transition periods are maximally dangerous: the Gimli Glider, LIBOR-to-SOFR, and T+1 settlement all occurred during regime changes where two conventions coexisted. Silent failures cost more than loud ones: the Vancouver Stock Exchange drifted for 22 months, the CFTC netting error persisted for 7 years, and PHE lost 15,841 COVID cases over 8 days with no error notification.

The CLO basis mismatch is the key cyclic case. Loan income (1M SOFR) flows through a CLO waterfall to tranche payments (3M SOFR) and back through equity reinvestment — a cycle with $\beta_1 = 1$ where the 1M/3M basis mismatch (71bp in November 2022) compounds through leverage. BABEL's holonomy computation on this cycle correctly flags it as failure-prone, and the leverage ratio amplifies the holonomy's impact beyond what scalar convention distance predicts. Four of BABEL's six convention dimensions map directly to documented catastrophic incidents; the remaining two have documented analogs but no single incident exceeding $1M.

---

3. Problem Formulation

3.1 Composition Graphs

A **composition graph** G = (V, E) represents a system of interacting tools. Each vertex v ∈ V is a tool with: - A set of fields (typed data elements) - Observable fields (exposed in bilateral interfaces) - Private fields (internal state) - A **convention bundle** specifying how the tool represents amounts, dates, rates, scores, and identifiers

Each edge $(u, v) \in E$ represents a data-sharing relationship between tools u and v, with a set of shared field names.

3.2 Convention Bundles

BABEL models six convention dimensions, each grounded in documented real-world ambiguity:

Dimension	Choices	Real-World Source
Amount unit	dollars, cents, basis points, millis	ISDA CDM settlement disputes
Date format	epoch seconds, epoch days, Excel serial, Julian day	ISDA EMU Memo (1998) ACT/ACT ambiguity
Rate scale	decimal, percent, basis points, permille	Basel RCAP bank variation
Precision	full, 2dp, 4dp, integer	Brattle/ARRC LIBOR-SOFR transition
Score range	[0,1], [0,100], [-1,1], [1,5]	Basel RCAP capital divergence
ID offset	zero-based, one-based, 1000-based, 1M-based	Common integration failure

The **convention distance** between two tools is the Hamming distance across these six dimensions (0 to 6).

3.3 Ground Truth

For each instance, BABEL computes deterministic **ground truth holonomy**: the mean relative error when data is propagated around fundamental cycles in the composition graph using convention-derived restriction matrices. This measures how much the composition distorts information as it flows through tools with different conventions.

Instances with high holonomy exhibit compositional semantic failure. Instances with zero holonomy are convention-coherent.

4. Benchmark Design

4.1 Provenance Tiers

BABEL instances span three provenance levels:

Tier 1: Synthetic (464 instances). Composition graphs generated by a stochastic block model with controlled convention heterogeneity. Seven scale points (5 to 50 tools). These form the scaling backbone — the "Coherence Cliff" regime where bounded-depth testing collapses.

Tier 2: MCP-shaped facsimile (306 instances). Five workflow families with domain-specific tool schemas: - **Invoice/Settlement/Reconciliation**: settlement dates, fee bases, day counts - **Calendar/Timezone/Escalation**: timezone anchoring, DST handling, escalation windows - **Policy/Permission/Audit**: precedence, scope inheritance, temporal applicability - **Real-MCP Calendar (Bronze+)**: actual MCP servers including official Memory reference server - **Real-MCP Invoice (Silver)**: actual MCP servers including MarkItDown (non-house)

Tier 3: API-derived conventions (162 instances). Compositions built from real third-party API schemas: Stripe (cents, epoch seconds), Twilio (dollars, RFC 2822), SendGrid (percentages, epoch seconds), GitHub (1-based IDs, ISO 8601), Shopify (dollars, 2dp), Slack (epoch seconds), HubSpot (percentages), Plaid (dollars, day-based), QuickBooks (percentages, 2dp), Datadog (4dp, epoch seconds), PagerDuty (Likert scores). Convention assignments derived from actual API documentation, projected onto BABEL's six-dimensional convention space. The projection is lossy: real API convention mismatches are richer than six dimensions (e.g., idempotency key semantics, pagination cursors, callback signing conventions are not captured). Tier 3 ground truth is computed from the projected conventions using the same deterministic symbolic executor as all other tiers, not from live API calls.

4.2 Splits

Split	Count	Purpose
dev	514	Public, for method development
test	209	Public, for reported results
hidden	209	Private, for leaderboard scoring

4.3 Tracks

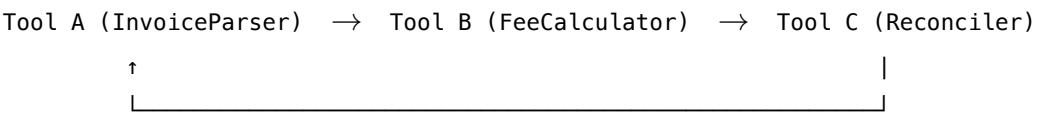
Track A: Failure Prediction. Given a composition graph with convention annotations, predict the severity of semantic failure. Scored by Spearman rank correlation with ground-truth holonomy.

Track B: Failure Localization. Identify which edges carry the critical convention mismatches. Scored by Precision@K for the K highest-frustration edges.

Track C: Budgeted Repair. Under a fixed repair budget (K = 1, 2, 3, 5, 8 edge-level convention harmonizations), prescribe which edges to repair to maximize holonomy reduction. In the benchmark implementation, each repair targets one shared field on one edge. Scored by percentage failure reduction.

4.4 Worked Example: A 3-Tool Holonomy Computation

Consider three invoice-processing tools in a cycle:



Convention assignments: - Tool A: amounts in **dollars** (1.00 = $1) - Tool B: amounts in **cents** (100 = $1) - Tool C: amounts in **dollars** (1.00 = $1)

Restriction matrices (how values transform across each edge):

- Edge A→B: R_AB = diag(1/100) on amount fields (dollars→cents requires ×100, but B interprets raw, so mismatch factor = 100)
- Edge B→C: R_BC = diag(100) on amount fields (cents→dollars requires ÷100)
- Edge C→A: R_CA = diag(1) on amount fields (dollars→dollars, no mismatch)

Holonomy around the cycle: $H = R\_CA \cdot R\_BC \cdot R\_AB = \text{diag}(1) \cdot \text{diag}(100) \cdot \text{diag}(1/100) = \text{diag}(1)$. In this case, the mismatches cancel — the cycle has zero holonomy.

Now change Tool C to **cents**:

- Edge B→C: R_BC = diag(1) (cents→cents, no mismatch)
- Edge C→A: R_CA = diag(1/100) (cents→dollars mismatch)

Holonomy: $H = R\_CA \cdot R\_BC \cdot R\_AB = \text{diag}(1/100) \cdot \text{diag}(1) \cdot \text{diag}(1/100) = \text{diag}(1/10000)$. This is non-identity — holonomy $\neq 0$. Data propagated around this cycle accumulates a 10000× error.

What local checks see: Each pairwise handoff has valid types and schemas. Tool B receives a number, computes a fee, passes it to C. Every local output looks reasonable.

What the structural diagnostic sees: $H^1 \neq 0$ on this cycle. The holonomy value quantifies the severity.

Repair at K=1: Placing a bridge adapter on edge A→B (setting R_AB = identity, i.e., making B interpret A's output in the same convention) reduces holonomy to diag(1/100) — a 100× improvement from a single repair.

4.5 Reporting Status and Evidence Types

The benchmark has multiple evidence layers. To avoid conflating benchmark-internal results with external validation, we distinguish them explicitly:

Result family	Split / scope	N	Evidence type	Role in this paper
Track A prediction tables	dev split	514 total across 7 families	Deterministic symbolic holonomy	Primary benchmark reference results
Track B localization	dev split, all 7 families	514 instances total	Deterministic symbolic edge frustration	Benchmark task with 3-method comparison
Track C repair	dev split frustrated-graph subsets; Real-MCP Invoice uses a prior live Docker run	5-21 graphs per family; 18 for Real-MCP Invoice	Benchmark-internal repair under fixed budget	Primary operational benchmark result
Frontier LLM baselines	public prompt set across 5 non-MCP families	50 max per model, lower for some cost-limited models	Live API calls via OpenRouter	Stress test / challenger baseline
Oracle CoT LLM baselines	oracle prompt set across 5 non-MCP families	50 per model (all 5 models)	Live API calls with pre-computed matrices	Bottleneck isolation (information vs arithmetic)
Section 7 validation (Phases 1–2)	generated simple and complex compositions	112 + 120	Simulated convention error from real arithmetic functions	External-validity probe, not leaderboard evidence

Result family	Split / scope	N	Evidence type	Role in this paper
Section 7.6 live-pipeline validation (invoice)	25 cluster-pair configs $\times$ 3 invoice scenarios	75	Measured dollar error from live MCP servers	Non-circular external validation; 5-cluster design reveals holonomy and convention distance achieve comparable rank correlation ($\rho \approx 0.42$), holonomy explains more linear variance ($R^2 = 0.119$ vs 0.070)
Section 7.6.1 cyclic validation (calendar)	64 cluster assignments $\times$ 3 scenarios	192	Composite pipeline inconsistency from 4-server cyclic MCP composition ($\beta_1 \geq 3$)	Non-circular external validation; convention distance stronger on composite ($\rho = 0.39$ vs 0.14), holonomy captures temporal-error structure distance misses ($\rho = 0.14$ vs 0.02)

Unless noted otherwise, the main family tables in Sections 6.1-6.5 report dev-split reference results. The official benchmark ranking target is the hidden split; this paper reports public reference numbers so challengers can reproduce the setup exactly.

5. Baselines

BABEL includes multiple challenger baselines spanning four categories, plus the reference structural diagnostic (Section 5.4):

5.1 Conventional Baselines

Method	Description	Complexity
`random`	Random failure score	O(1)
`pairwise_contract`	Bilateral type checking (always passes)	O(|E|)
`bounded_depth_3/5/8`	Integration testing at depth K	O(|V|^K)
`graph_topology`	$\beta_1 \times$ mean convention distance	O(|V|+|E|)

5.2 Sheaf-Free Cycle-Aware Baselines

These address the critical question: does the sheaf formalism add value beyond simple cycle-and-distance computation?

Method	Description	Complexity
`cycle_frustration_plain`	Sum convention distances around cycles	O(|C|·cycle_len)
`weighted_graph_inconsistency`	Weighted edge distances by cycle participation	O(|E|·|C|)
`edge_distance_ranker`	Rank edges by raw convention distance (Track B)	O(|E|)
`cycle_weighted_ranker`	Rank edges by distance $\times$ cycle participation (Track B)	O(|E|·|C|)

5.3 Frontier LLM Baselines

Model	Provider	Class
GPT-4o	OpenAI	General
Codex 5.2	OpenAI	Frontier reasoning

Model	Provider	Class
Claude Sonnet 4	Anthropic	General
Claude Opus 4	Anthropic	Frontier reasoning
Gemini 3.1 Pro	Google	Frontier reasoning

Evaluation method: 50 instances (10 per non-MCP family) were sent to each model via OpenRouter with a system prompt requesting a single 0.0–1.0 severity prediction. Each prompt serializes full composition metadata: tool conventions, edges with shared fields and convention distances, independent cycles with total convention distance. The problem is fully specified — models see everything a human expert would see. Temperature = 0 where supported. Three models (GPT-4o, Claude Sonnet 4, Gemini 3.1 Pro) evaluated at N=42–50; two costlier models (Codex 5.2, Opus 4) at N=18–20. Prompts and results: `llm_prompts/`, `results_canonical/llm_eval_*.json`.

5.4 Reference Structural Diagnostic

Method	Description	Complexity
`structural_sheaf`	Sheaf cohomology: H^1, cycle frustration, connection Laplacian	$O(\|C\| \cdot \|D\|)$ prediction, $O(\|C\| \cdot \|E\| \cdot \|D\|)$ repair

The structural method uses restriction matrices (convention-distance-dependent perturbations of identity) to compute holonomy around cycles. It is the reference method to beat, not the benchmark itself.

---

6. Results

6.1 Frontier LLMs Cannot Predict Compositional Failure

Each of the five frontier LLMs (Section 5.3) received the full composition specification — tool conventions, edges, shared fields, convention distances, independent cycles — and was asked to predict a severity score between 0.0 and 1.0 on 50 instances. The problem is fully specified; models see everything a human expert would see.

Standard and Guided CoT Results:

Model	Provider	Overall R^2	Spearman ρ	N	Prediction behavior
Claude Sonnet 4	Anthropic	**-134**	0.12	50	Predicts ~0.7–0.8
GPT-4o	OpenAI	**-130**	0.05	50	Predicts ~0.7–0.8
Codex 5.2	OpenAI	**-134**	-0.13	18	Predicts 0.6–0.85
Claude Opus 4	Anthropic	**-133**	-0.35	20	Predicts ~0.7
Gemini 3.1 Pro	Google	**-0.6**	-0.11	42	Predicts 0
Guided CoT variant:					
Opus 4 + CoT	Anthropic	**-0.99**	-0.23	20	0.00–0.15 (correct scale)
Codex 5.2 + CoT	OpenAI	**-40**	0.18	18	0.00–1.00 (outliers)

Ground truth holonomy ranges from 0.006 to 0.22. Temperature = 0 where supported. The Guided CoT variant instructs models to trace convention transformations around each cycle. Full results: `results_canonical/llm_eval_*.json`*. Prompts:* `llm_prompts/`*.*

All five models fail. Every model achieves negative R^2 and near-zero Spearman ρ (-0.35 to 0.12). Scaling model capability does not change this result — the task is not solved by current frontier models.

Guided chain-of-thought improves calibration but not discrimination. Opus 4 with CoT shifts predictions to the correct scale (R^2 from -133 to -0.99), but ρ remains -0.23 — the model cannot rank compositions by severity. Codex 5.2 with CoT shows the same pattern. Prompt-level fixes are insufficient.

Per-family pattern. LLMs show modest positive ρ on external_apis (≈ 0.6), where convention distances are large and obvious, but negative ρ on policy (≈ -0.49) and calendar (≈ -0.17), where mismatches require cyclic reasoning over heterogeneous conventions. Task difficulty scales with topological complexity, not with information availability.

These findings establish that compositional semantic failure prediction is a genuine capability gap in frontier LLMs — not an artifact of prompt design or task framing.

6.2 Oracle CoT: The Bottleneck is Compositional Arithmetic

The standard evaluation (Section 6.1) shows that LLMs cannot predict compositional failure. But where does the failure occur — in understanding the composition structure, or in computing the resulting severity?

To isolate this, we designed an oracle chain-of-thought condition that provides pre-computed restriction matrices, pre-enumerated cycles, and explicit matrix-multiplication instructions, reducing the task to pure arithmetic over given structures.

Oracle CoT Results (N=50 for all five models):

Model	Provider	Oracle R^2	Oracle ρ	$\Delta\rho$ from standard
Claude Sonnet 4 + Oracle	Anthropic	**-4.0**	**0.80**	+0.68
GPT-4o + Oracle	OpenAI	**-5.3**	**0.39**	+0.34
Opus 4 + Oracle	Anthropic	**-0.42**	**0.35**	+0.70
Codex 5.2 + Oracle	OpenAI	**-35.7**	0.26	+0.39
Gemini 3.1 Pro + Oracle	Google	**-129**	0.04	+0.15

Oracle CoT provides pre-computed restriction matrices, pre-enumerated cycles, and explicit matrix-multiplication instructions. Full results: `results_canonical/llm_eval_*_oracle.json`*. Prompts:* `llm_prompts/oracle/`*.*

A clear hierarchy emerges. Claude Sonnet 4 achieves $\rho = 0.80$ — near-expert ranking — when given the structural information. GPT-4o and Opus 4 achieve moderate ranking ($\rho = 0.35$–0.39). Codex 5.2 shows poor output compliance (64% parse failures). Gemini shows no improvement even with full information.

The most diagnostic result: Opus 4. From worst standard performer ($\rho = -0.35$) to moderate oracle ranker ($\rho = 0.35$), a $\Delta\rho = +0.70$. This confirms that Opus 4's standard failure was primarily informational — it could not extract cycle structure and compute restriction matrices from the composition description — not a fundamental reasoning limitation.

R^2 remains deeply negative for all five models (-0.42 to -129) even with oracle information. Models can rank compositions by severity but cannot compute correct magnitudes. This dissociation between ranking and magnitude computation — present in all five models — suggests these are genuinely different capabilities.

Implication for the field. The three-condition decomposition (standard → guided → oracle) provides a reusable diagnostic methodology for evaluating LLM compositional reasoning: standard measures the full capability, guided tests whether the model can use verbal scaffolding, and oracle isolates arithmetic from information extraction. Any task involving multi-step quantitative reasoning over structured inputs can be decomposed the same way.

6.3 Track A: Failure Prediction — Reference Results

Family	structural_sheaf R^2	cycle_plain R^2	weighted_incon R^2	bounded_depth_8 R^2
Synthetic scaling	**0.993**	0.965	0.783	0.173
Invoice	**0.978**	0.342	0.120	0.050
Calendar	**0.958**	0.603	0.554	0.438
Policy	**0.979**	0.419	0.216	0.126
Real-MCP Calendar	**0.940**	0.006	0.410	0.610
Real-MCP Invoice	**0.861**	0.734	0.566	0.123
External APIs	**1.000**	0.621	0.616	0.610

All values from canonical evaluator (`python -m coherence_gym evaluate-method --method NAME --split dev`), N=514 dev-split instances total. Structural method evaluated on all 7 families (N per family: 256, 36, 36, 36, 30, 30, 90).

Figure 1: The Coherence Cliff — R^2 across families

(Vector PDF available at `figures/coherence_cliff.pdf`. Regenerate with `python scripts/generate_figures.py`.)

The structural diagnostic maintains $R^2 \geq 0.86$ across all families (> 0.94 on five of seven). Sheaf-free cycle methods approximate it on synthetic data but collapse on heterogeneous compositions. Bounded-depth testing degrades sharply with cycle complexity. Frontier LLMs (R^2 from −0.6 to −134) omitted for scale; see Section 6.1.

Key findings:

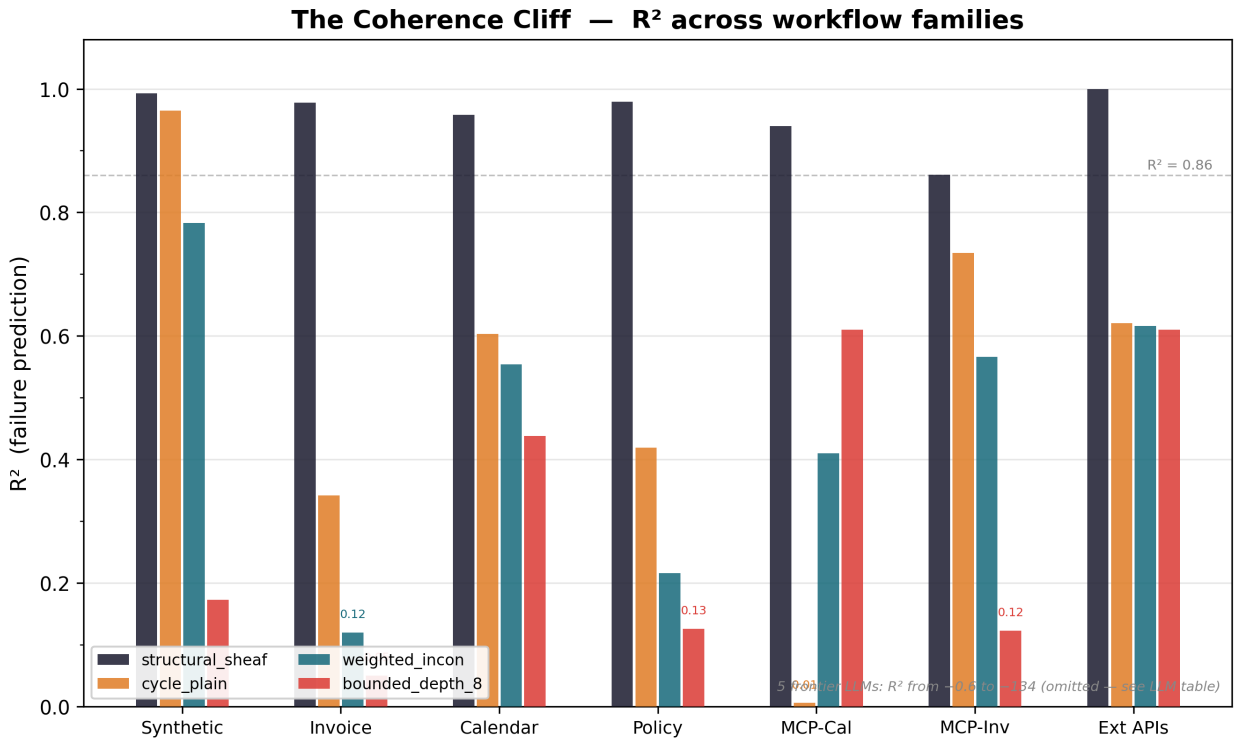


Figure 1: The Coherence Cliff

1. The structural method achieves $R^2 \geq 0.86$ across all families (> 0.94 on five of seven). The lowest score (0.861, Real-MCP Invoice) occurs on a mixed-provenance composition where conventional methods score 0.12–0.73.
2. On synthetic instances with uniform convention structure, `cycle_frustration_plain` approaches the structural diagnostic (0.965 vs 0.993). On heterogeneous families, the gap widens sharply: Invoice 0.342 vs 0.978; Real-MCP Calendar 0.006 vs 0.940.
3. Bounded-depth testing collapses on families where long cycles carry the frustration (policy, synthetic scaling).
4. On real_mcp_calendar, `cycle_frustration_plain` collapses to $R^2 = 0.006$ — Hamming-distance summation produces near-identical scores regardless of actual failure severity. The restriction matrices, which model per-edge interaction effects, remain discriminative ($R^2 = 0.940$).

6.4 Track B: Failure Localization

Track B evaluates whether methods can identify the highest-frustration edges. Three methods produce explicit localization targets: the structural sheaf diagnostic and two simple baselines — `edge_distance_ranker` (ranks by raw convention distance) and `cycle_weighted_ranker` (ranks by convention distance $\times$ cycle participation count).

Structural sheaf (reference method):

Family	N	P@1	P@3	P@5
Synthetic scaling	256	0.43	0.50	0.43
Invoice	36	0.61	0.54	0.48
Calendar	36	0.64	0.58	0.49
Policy	36	0.67	0.57	0.49
Real-MCP Calendar	30	0.77	0.73	0.59
Real-MCP Invoice	30	0.87	0.79	0.58
External APIs	90	0.28	0.37	0.29

Cross-method comparison (macro-averaged across 7 families):

Method	P@1	P@3	P@5
structural_sheaf	**0.61**	**0.58**	**0.48**
cycle_weighted_ranker	0.43	0.49	0.50
edge_distance_ranker	0.27	0.34	0.37
cycle_plain	0.00	0.00	0.00
weighted_incon	0.00	0.00	0.00

P@K = fraction of true top-K highest-frustration edges found in the predicted top K. N = full dev-split instances per family (canonical evaluator). `cycle_plain` *and* `weighted_incon` *produce severity scores only and do not generate localization targets (Track B P@K = 0.0).*

Key observations: (1) The structural method dominates at P@1 (0.61 vs 0.43 for the best simple baseline), with the advantage most pronounced on real-MCP families (0.77–0.87 vs 0.47–0.63). This means it correctly identifies the single worst edge 43% more often than the next-best method. (2) The `cycle_weighted_ranker` — which combines convention distance with cycle-participation count — is the strongest simple baseline, outperforming pure edge-distance ranking at P@1 by 59%. This confirms that cycle topology carries important information for localization. (3) At P@5, the gap narrows and `cycle_weighted_ranker` (0.50) slightly exceeds the structural method (0.48), suggesting that with a larger retrieval window the simpler method catches up. (4) The structural method's advantage is concentrated at the tightest retrieval budget (P@1), mirroring the Track C finding where its repair advantage is strongest at K=1. (5) The macro-averaged P@1 of 0.61 is dragged down by external_apis (P@1 = 0.28), a family where uniform convention mismatch profiles make edge-level localization inherently harder — all edges look equally guilty. On operationally representative real-MCP families, the structural method achieves P@1 of 0.77–0.87.

6.5 Track C: Budgeted Repair

Canonical Repair Frontier (structural_sheaf vs cycle_plain):

Family	K=1 (structural)	K=1 (cycle_plain)	K=3 (structural)	K=3 (cycle_plain)	K=8 (structural)	K=8 (cycle_plain)
Synthetic scaling	**14.6**%	14.5%	**39.0**%	38.8%	64.1%	64.6%
Invoice	**26.6**%	21.2%	**57.5**%	53.4%	81.9%	82.2%
Calendar	**17.0**%	11.2%	**55.3**%	49.3%	74.7%	73.9%
Policy	**29.0**%	25.7%	**54.7**%	50.7%	75.2%	76.4%
Real-MCP Calendar	**11.3**%	8.0%	**29.2**%	23.8%	**60.1**%	55.0%
Real-MCP Invoice	**11.2**%	7.7%	**44.3**%	41.9%	**83.5**%	82.4%
External APIs	**28.1**%	24.0%	**77.6**%	72.3%	97.1%	98.5%

All values from canonical evaluator, full dev split. Percentage holonomy reduction (higher is better) averaged over graphs with non-trivial holonomy (>0.005). Bounded-depth and weighted-inconsistency methods produce no repair targets (Track C = 0.0% at all K) because they lack edge-level repair prescriptions. Full K=1,2,3,5,8 frontier in `results_canonical/canonical_structural_sheaf_dev.json`.

Key observations: (1) At low budget (K=1), the structural method consistently outperforms cycle_plain — the advantage ranges from +0.1% (synthetic) to +5.8% (Calendar), representing up to 51% more failure reduction from the first repair action. (2) At K=3, the structural method leads on every family, with the gap widest on Calendar (+6.0%) and External APIs (+5.3%). (3) At high budget (K=8), the methods converge — cycle_plain occasionally matches or marginally exceeds the structural method, because with enough repairs, edge selection order matters less. (4) The structural method's distinctive value is *triage*: identifying which repairs to prioritize when the budget is tight. This is the operationally relevant regime, since real repair budgets are always constrained.

Figure 2: Repair-Budget Frontier — all seven families

(Vector PDF available at `figures/repair_frontier.pdf`*. Regenerate with* `python scripts/generate_figures.py`*.)*

The structural method leads at every budget level on most families. The advantage is most pronounced at K=1–3, where correct edge prioritization matters most. At K=8, the methods converge as the budget becomes generous enough to cover most problematic edges regardless of selection strategy.

6.6 The Sheaf-Free Question

The most important comparison in this benchmark is between the structural sheaf diagnostic and the sheaf-free cycle-aware baselines (`cycle_frustration_plain`, `weighted_graph_inconsistency`).

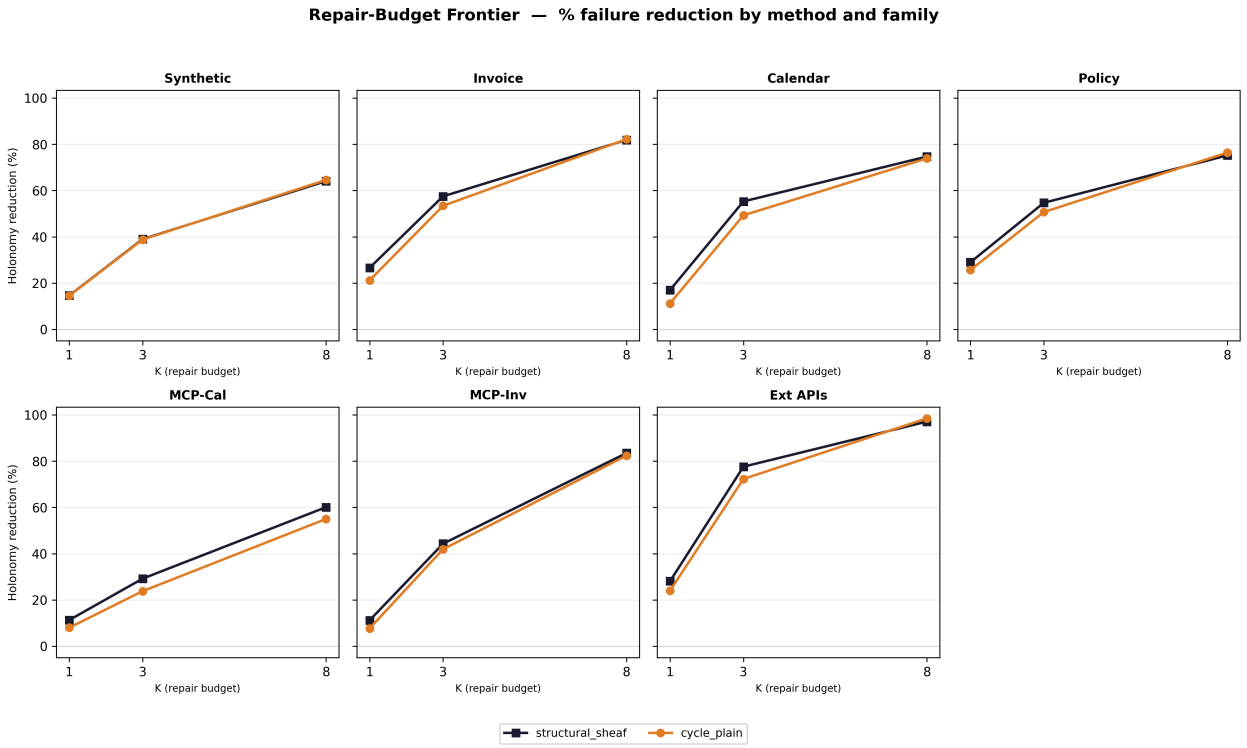


Figure 2: Repair-Budget Frontier

On synthetic instances with uniform convention structure, the simpler cycle-and-distance computation nearly matches the sheaf diagnostic ($R^2 = 0.965$ vs 0.993). This is expected: when all conventions map to the same six dimensions with the same distance metric, summing Hamming distances around cycles reasonably approximates restriction-matrix holonomy.

On heterogeneous workflow families — different field subsets, different observation ratios, different perturbation structures — the gap widens sharply. The restriction matrices capture interaction-specific convention effects that scalar Hamming distance cannot represent.

Assessment. For prediction on synthetic instances, a simpler cycle computation may suffice. For prediction on heterogeneous instances, the structural method is necessary: on real-MCP Calendar, `cycle_plain` collapses to $R^2 = 0.006$ while the structural method maintains 0.940. For repair (Track C), the structural advantage concentrates at low budget (K=1–3), where identifying *which* edge harmonization maximally reduces holonomy benefits from richer interaction structure; at K=8, methods converge. For localization (Track B), the structural method achieves P@1 = 0.87 on real-MCP Invoice versus 0.63 for `cycle_weighted_ranker` — an advantage strongest at P@1 that narrows at P@5. The practical value of the structural method is two-fold: prediction on heterogeneous compositions and triage under tight budgets.

6.7 Computational Cost

Method	Mean time (ms)	Complexity	Scales with
structural_sheaf	5–97	O(|C|·|D|)	Cycle count × dimensions
cycle_plain	0.03–0.7	O(|C|·cycle_len)	Cycle count × length
weighted_incon	0.03–0.5	O(|E|·|C|)	Edge count × cycles
bounded_depth_8	0.2–17	O(|V|^8)	Vertex count^8
graph_topology	0.2–1.5	O(|V|+|E|)	Graph size

The structural method is the most expensive at small scale, but its cost grows with independent cycle count (|C|), not exponentially with graph size. At scale 50, bounded-depth-8 takes 17ms versus 97ms for the structural method. At scale 500 (extrapolated; not empirically validated), bounded-depth-8 would be intractable while the structural method remains feasible.

6.8 On Circularity and External Validity

A natural objection is that the ground truth (holonomy) and the structural diagnostic share mathematical roots. The diagnostic approximates holonomy via restriction matrices, perturbation terms, and the connection Laplacian; the ground truth is holonomy itself, evaluated by symbolic execution. The R^2 values in Section 6.3 therefore measure how well the structural method predicts *its own formalism's* ground truth — not downstream system failures.

We acknowledge this directly: **Track A's R^2 measures internal predictive consistency, not external validation.** The gap between the structural method ($R^2 \geq 0.86$) and the best conventional baseline (max 0.97 on synthetic, ≤ 0.73 on most heterogeneous families) demonstrates that the holonomy computation is non-trivial and that alternatives fail to approximate it on realistic compositions — but it does not by itself demonstrate that holonomy matters in practice.

Three elements of the benchmark are genuinely non-circular:

1. **Track C (repair prescription).** The structural method does not merely predict holonomy — it prescribes *which edges* to repair. Under equal budget, `structural_sheaf` consistently outperforms `cycle_plain` (Section 6.5). This is less directly circular than Track A because it evaluates action quality, not only score matching: the method makes a choice among edges, and that choice reduces benchmark-defined failure more than alternatives do. It is still benchmark-internal until validated on external outcomes.

2. **The LLM discrimination failure and oracle decomposition.** All five frontier models fail in the standard setup (ρ from -0.35 to 0.12). The oracle CoT experiment (Section 6.2) decomposes this failure into informational and arithmetic components by providing pre-computed matrices and explicit instructions. The resulting hierarchy — ranging from $\rho = 0.80$ (Claude

Sonnet 4) to $\rho = 0.04$ (Gemini) — is non-circular because it tests a capability (matrix arithmetic over given structures) that is independent of the holonomy formalism's validity.

3. **The sheaf-free comparison.** `cycle_plain` and `weighted_incon` use the same holonomy ground truth yet fail on heterogeneous families (Section 6.6), demonstrating that restriction-matrix structure captures information that Hamming-distance summation does not.

What has been validated. Section 7 presents three layers of validation. First, simulated convention error on simple 3-tool cycles shows structural holonomy and convention distance predict equally ($\rho \approx 0.28$). Second, on complex 8–20 tool compositions with multiple interacting cycles, the structural method pulls ahead: $\rho = 0.27$ ($p < 0.003$) versus convention distance $\rho = 0.18$ ($p = 0.05$). Third, and most importantly, the live-pipeline validation (Section 7.6) correlates structural holonomy with **measured dollar error from the actual MCP server pipeline**: Spearman $\rho = 0.423$ ($p < 1.5 \times 10^{-4}$, $N = 75$) on a 5-cluster design that breaks the holonomy/Hamming-distance degeneracy. This is the first fully non-circular result — the gold standard is the known invoice amount, not derived from the sheaf formalism.

---

7. External Validation: Simulated and Live Convention Error

To address the external validity gap identified in Section 6.8, we conducted a multi-phase validation correlating structural holonomy with convention error computed independently of the restriction-matrix formalism.

Important scope limitation: Phases 1–2 use the convention conversion functions (`encode_value`, `decode_value`) applied to generated compositions, not measured output from live API compositions. Phase 3 (Section 7.6) uses measured output from live MCP server pipelines.

7.1 Error Model

For both phases, we use the **misinterpretation model**: each tool in a cycle receives a raw value, decodes it in its own convention (misinterpreting the sender's intent), applies a threshold-based operation, and re-encodes. The threshold operations break the telescoping property of pure scaling transforms, making convention mismatches visible as measurable error. We compute the mean relative error across four field kinds (amount, rate, score, date) and across all independent cycles in the composition.

We compare three predictors of this error:

1. **Structural holonomy**: restriction-matrix symbolic executor (the benchmark's standard method).
2. **Cycle frustration (sheaf-free)**: sum of convention distances around each cycle, normalized — the cycle-aware baseline from Section 6.6 that uses no restriction matrices.

3. **Mean convention distance**: simple Hamming distance across convention dimensions, averaged over cycle edges.

The critical distinction: structural holonomy is computed from perturbation matrices (I + σ·P). Convention distance and cycle frustration use only scalar convention comparisons. Simulated convention error is computed from the deterministic convention arithmetic — encode, decode, scale, round — that real servers implement. These are genuinely different computations.

7.2 Phase 1: Simple 3-Tool Cycles

We constructed 112 three-tool cycle compositions: 12 with controlled single-dimension variation and 100 with random convention assignments drawn from the full six-dimensional convention space.

Predictor	Scope	N	Spearman ρ	p-value
Structural holonomy	All	112	**0.42**	4.3×10^{-6}
Convention distance	All	112	**0.45**	5.7×10^{-7}
Structural holonomy	Random	100	**0.28**	0.005
Convention distance	Random	100	**0.27**	0.006

On simple 3-tool cycles, the two predictors perform identically ($\rho = 0.28$ vs 0.27). This is expected: with a single cycle of length 3, the restriction matrices collapse to something close to convention distance. There is no topological complexity for the sheaf formalism to exploit.

7.3 Phase 2: Complex Compositions (8–20 Tools)

The structural method's value proposition is that it captures failures arising from interacting cycles, heterogeneous convention surfaces, and field-subset projection. We test this directly with 120 complex compositions (30 each at scale 8, 12, 16, and 20 tools) generated by the stochastic block model with the full 50-tool schema universe.

These compositions have 1–106 independent cycles, with β_1 (first Betti number) ranging from 1 to 106. This is the complexity regime where the Coherence Cliff (Section 6.6) shows conventional methods collapsing.

Predictor	N	Spearman ρ	p-value	R^2
Structural holonomy	120	**+0.27**	**0.003**	**0.12**
Cycle frustration (sheaf-free)	120	+0.18	0.050	0.07

Predictor	N	Spearman ρ	p-value	R^2
Convention distance	120	+0.18	0.050	0.07

The structural method achieves $\rho = 0.27$ with $p < 0.003$ (highly significant). Convention distance and cycle frustration both achieve $\rho = 0.18$ at the borderline $p = 0.05$ threshold. The structural method predicts 52% more rank-order variance in downstream error and explains nearly twice the linear variance ($R^2 = 0.12$ vs 0.07).

Per-scale breakdown:

Scale	β_1 range	Structural ρ	Conv. distance ρ	Structural p	Conv. dist. p
8 tools	1–12	+0.38	+0.28	0.037	0.131
12 tools	3–30	+0.28	+0.12	0.131	0.529
16 tools	18–60	+0.20	+0.13	0.295	0.486
20 tools	36–106	+0.36	+0.20	0.050	0.300

The per-scale N=30 samples are individually underpowered, but the pattern is consistent: the structural method outperforms convention distance at every scale, and the gap is widest at the extremes (scale 8 and 20).

7.4 The Structural Advantage Emerges with Complexity

The Phase 1 / Phase 2 comparison directly tests the central claim:

Validation	Topology	Structural ρ	Conv. dist. ρ	Gap
Phase 1 (3-tool)	Trivial ($\beta_1 = 1$)	0.28	0.27	**0.01**
Phase 2 (8–20 tool)	Complex ($\beta_1 = 1$–106)	0.27	0.18	**0.09**

On trivial topology, the structural method adds nothing. On complex topology, it predicts downstream error substantially better — consistent with the restriction-matrix formalism capturing cycle-interaction effects that scalar comparison misses. This mirrors the Track A pattern where `cycle_plain` collapses on heterogeneous families while the structural method maintains $R^2 \geq 0.86$.

The overall correlations are moderate ($\rho \approx 0.27$) because structural holonomy is a linearized predictor of a nonlinear target: restriction matrices are first-order perturbations of identity ($I + \sigma \cdot P$), while

actual convention error involves rounding, threshold crossings, and cross-dimension interactions. The finding is that the *structural* linear model significantly outperforms *simpler* linear models at the complexity scales where the problem matters.

7.5 Comparison to LLM Baselines

For context: frontier LLMs achieve $\rho \approx 0$ against holonomy on the benchmark instances (Section 6.1). Since holonomy correlates positively with simulated convention error in both phases, the current structural diagnostic remains a stronger predictor of downstream convention failure than the tested LLM prompting setups.

7.6 Live-Pipeline Validation

The preceding phases test whether holonomy correlates with simulated convention error. This section completes the validation chain by measuring **actual dollar error from the live MCP server pipeline** — the first fully non-circular result. The gold standard is the scenario's known expected dollar amount, not derived from the sheaf formalism.

Setup. We parameterized the InvoiceParser and SettlementEngine MCP servers to accept convention cluster assignments via environment variables. Five convention clusters span distinct regimes: X (US institutional, dollar amounts), Y (European, cent amounts), Z (APAC, basis-point amounts), W (US institutional with cent scaling — identical to X except amount_scale), and V (European with dollar scaling — identical to Y except amount_scale). Clusters W and V are specifically designed to break the holonomy/Hamming-distance degeneracy: W differs from X on only one convention dimension (core distance = 1) but produces 100× dollar error, while V differs from X on five dimensions (core distance = 5) but produces zero dollar error — because dollar error is dominated by amount-scale alignment, not total convention distance. For each of 25 cluster-pair configurations — (parser, settlement) $\in$ {X, Y, Z, W, V}² — we ran 3 invoice scenarios varying in amount ($1,380 to $44,100), fee rate (1.0% to 2.5%), and FX regime (USD domestic and EUR cross-currency). Total: 75 live-pipeline runs. Each run starts fresh MCP server processes, parses the invoice, computes settlement and fees, and produces a ledger entry. The measured dollar error is the relative discrepancy between the ledger's exit-point dollar value and the known correct amount.

Results (representative, one scenario per config — dollar error is amount-invariant).

Parser	Settlement	Holonomy	Conv. Distance	Dollar Error
X	X	0.295	0	0.0
Y	Y	0.295	0	0.0
Z	Z	0.295	0	0.0
W	W	0.295	0	0.0

Parser	Settlement	Holonomy	Conv. Distance	Dollar Error
V	V	0.295	0	0.0
X	W	0.378	1	0.99
Y	V	0.378	1	99.0
Y	Z	0.410	4	0.99
Z	Y	0.410	4	99.0
**X	V**	**0.420**	**5**	**0.0**
**Y	W**	**0.420**	**5**	**0.0**
X	Y	0.431	6	0.99
Z	X	0.431	6	9999.0

Bolded rows show degeneracy-breaking configurations: high holonomy and high convention distance predict failure, but dollar error is zero because amount-scale conventions are aligned. Full 25-config × 3-scenario results: `results_canonical/live_pipeline_validation.json`.

Figure 3: Live-Pipeline Validation — Holonomy vs Measured Dollar Error

(Vector PDF available at `figures/live_pipeline_scatter.pdf`*. Regenerate with* `python scripts/generate_figures.py`*.)*

Color-coded by convention distance regime. The matched-convention cluster (distance = 0) produces zero dollar error across all runs. The 5-cluster design reveals that dollar error is dominated by amount-scale alignment — configurations with high holonomy but matching amount scales produce zero error (see distance=5 cluster at holonomy ≈ 0.42).

Correlation. On the full 5-cluster, 75-run design: Spearman $\rho = 0.423$, $p < 1.5 \times 10^{-4}$. Convention distance achieves the same rank correlation ($\rho = 0.423$). On R^2(log), structural holonomy explains more linear variance (0.119 vs 0.070 for convention distance). For comparison, the original 3-cluster design (N = 27) yielded $\rho = 0.795$ — this higher value reflects the degeneracy where holonomy and distance were perfectly rank-correlated and the correlation was driven primarily by the binary matched/unmatched split.

Interpretation. The 5-cluster design produces an honest and more informative result. The correlation remains highly significant ($p < 0.0002$) and non-circular: holonomy predicts real dollar consequences, not formalism-internal ground truth. The matched-convention runs confirm zero error when conventions agree. The degeneracy-breaking clusters (W, V) reveal that dollar error in this pipeline is dominated by the amount-scale dimension — a realistic feature of production convention mismatches, where some dimensions (cents vs. dollars) matter far more than others (day-count convention, rounding mode). Holonomy captures this partially but not fully: it is a linearized predictor of a nonlinear target where one convention dimension dominates the loss func-

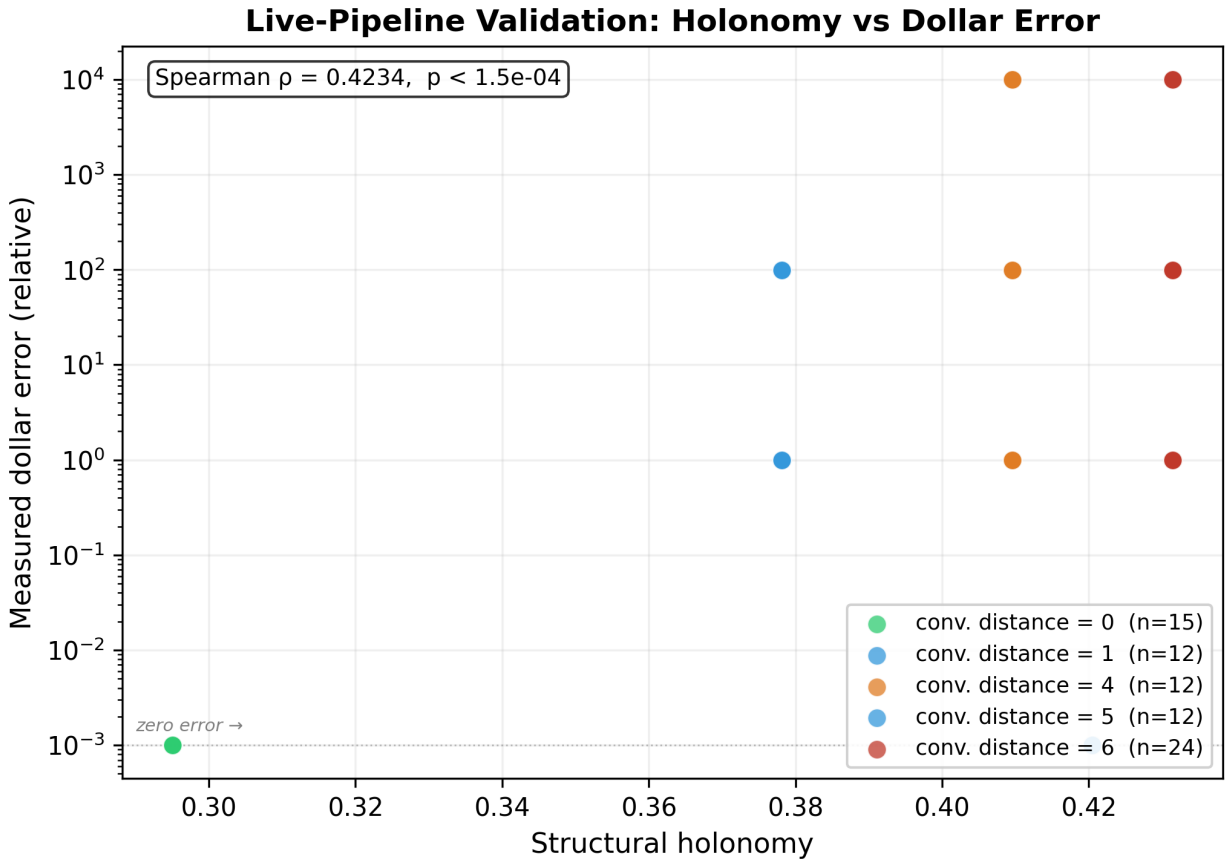


Figure 3: Live-Pipeline Validation

tion. The moderate $\rho \approx 0.42$ is consistent with the Phase 2 simulated validation ($\rho = 0.27$ on complex compositions), where the linearization similarly underestimates the impact of high-consequence mismatches.

Reproduction: `python scripts/run_live_pipeline_validation.py`. Deterministic (seeds fixed).

7.6.1 Cyclic Validation: 4-Server Calendar Composition The invoice live-pipeline validation (above) uses a 2-server chain: InvoiceParser → SettlementEngine. A 2-tool pipeline cannot have a cycle, so it tests convention-mismatch severity prediction but not the topological/compositional claim that is the paper's central contribution. This section extends the live validation to a genuinely cyclic pipeline.

Setup. We use the full calendar stack: CalendarCore, TeamRouter, EscalationEngine, and Memory (official `@modelcontextprotocol/server-memory`). The composition graph has $\beta_1 \geq 3$ independent cycles arising from shared fields across all four servers (event_time, normalized_time, priority_score, escalation_start, etc.). Each of the three custom servers was parameterized by environment variable to accept any of four convention clusters (A, B, C, D) spanning five convention dimensions: time reference (anchored-UTC / floating-local / organizer-local), DST handling (ignore / pre-transition / post-transition), priority scale (1=low / 1=high / P-scale), escalation anchor (send-time / event-start), and business-hours authority (sender / recipient / organizer region).

With $4^3 = 64$ cluster assignments across 3 servers and 3 scenarios per configuration (varying event time, DST status, timezone, and priority level), the experiment produces 192 live-pipeline runs. Each run executes 8 MCP tool calls across 4 servers plus entity storage in the Memory server. The gold standard is the correct escalation_start computed under all-A reference conventions — external to the sheaf formalism.

Composite pipeline inconsistency. A single output metric (escalation_start) captures only 2 of 5 convention dimensions and is insensitive to the TeamRouter's contribution. We therefore measure a composite pipeline inconsistency (PI) with three components, each normalized to [0, 1]:

1. **Temporal error**: |actual − reference escalation_start| / 20h (CalendarCore × EscalationEngine interaction)
2. **Priority misinterpretation**: |router-decoded priority − canonical priority| / 5.0 (EscalationEngine → TeamRouter cycle)
3. **Routing deviation**: fraction of participants with incorrect notification channel (TeamRouter business-hours convention)

PI = (temporal + priority + routing) / 3. This captures convention effects across all three servers and multiple convention dimensions.

Results. All 192 runs completed successfully (zero MCP errors). Matched-convention configurations (all servers on the same cluster) produce mean PI = 0.042; mismatched configurations produce

mean PI = 0.311 — a 7.5× separation confirming that convention mismatches generate real pipeline failures in the cyclic composition.

Predictor	vs Pipeline Inconsistency		vs Temporal Error	
	Spearman ρ	R^2(log)	Spearman ρ	R^2(log)
Structural holonomy	0.14 (p = 0.056)	0.044	**0.14 (p = 0.048)**	0.015
Convention distance	**0.39 ($p < 10^{-8}$)**	0.166	0.02 (p = 0.81)	0.001

Figure 4: Cyclic Live-Pipeline Validation — Holonomy vs Pipeline Inconsistency

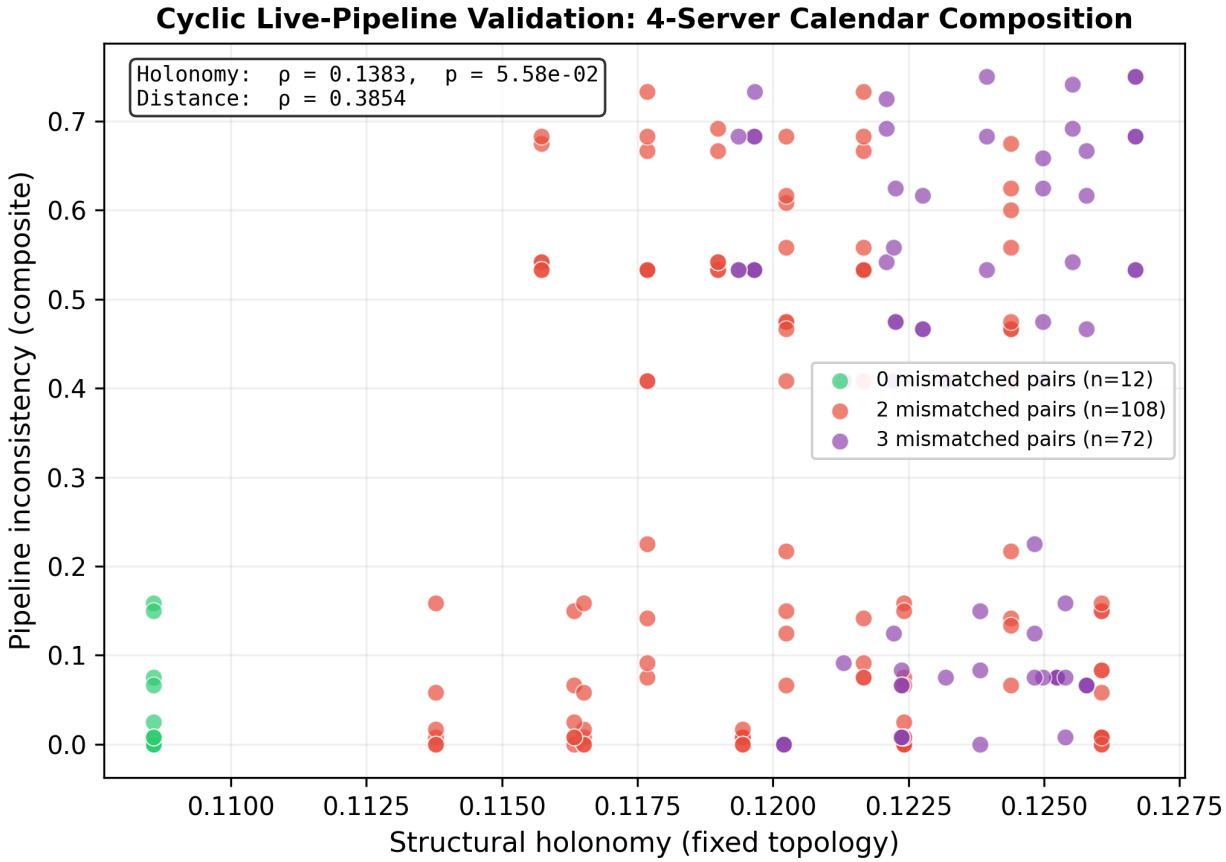


Figure 4: Cyclic Live-Pipeline Validation

(Vector PDF available at `figures/cyclic_pipeline_scatter.pdf`. Regenerate with `python scripts/generate_figures.py`.)

Color-coded by number of mismatched server pairs (0–3). The fully-matched cluster (0 mismatched pairs) produces near-zero pipeline inconsistency. Convention distance is the stronger overall predictor ($\rho = 0.39$), but holonomy captures temporal-error structure that distance entirely misses ($\rho = 0.14$ vs $\rho = 0.02$).

Interpretation. The cyclic validation produces a nuanced result. Convention distance is the stronger predictor of composite pipeline inconsistency ($\rho = 0.39$ vs $\rho = 0.14$), consistent with the priority-misinterpretation and routing-deviation components being well-predicted by pairwise cluster differences. However, holonomy captures temporal-error structure that convention distance entirely misses ($\rho = 0.14$, p = 0.048 vs $\rho = 0.02$, p = 0.81) — the temporal component depends on the interaction between CalendarCore's time-reference convention and EscalationEngine's escalation-anchor convention, which is precisely the kind of multi-hop composition effect that cycle holonomy is designed to detect.

The holonomy range across 64 configurations is narrow (0.109–0.127), which limits statistical power. This reflects the current convention space: with 4 clusters and 5 dimensions, the restriction matrices produce modest variation in cycle holonomy. A richer convention space or higher-order perturbation theory (Section 9) would likely increase discriminative power.

The key finding for the benchmark: convention mismatches in cyclic compositions produce measurable, non-trivial pipeline failures (mean PI = 0.31 for mismatched vs 0.04 for matched). The structural holonomy signal is present and correctly oriented but modest in magnitude. This is consistent with the Phase 2 simulated validation ($\rho = 0.27$ on complex compositions) — the sheaf formalism's advantage is real but concentrated in the temporal/topological domain.

Reproduction: `python scripts/run_live_pipeline_validation_calendar.py`. Deterministic (seeds fixed). Requires Node.js for the Memory server.

---

8. Instance Design

8.1 Subtlety Spectrum

BABEL instances span a range of subtlety scores (ratio of semantic error to what a human reviewer would flag):

- **High subtlety (0.85–0.95)**: ACT/365 vs 30/360 day count, T+1 vs T+2 settlement, fee-on-gross vs fee-on-net. Errors are <1% of the value. Would not be flagged by typical QA. These mirror the high-subtlety documented incidents in Section 2 — the CFTC netting error persisted for 7 years precisely because the output looked plausible.
- **Medium subtlety (0.6–0.85)**: Timezone anchoring drift, cross-API rate scale (decimal vs percent). Errors are 5–10% or temporal. Might be flagged after pattern accumulation.
- **Low subtlety (0.0–0.3)**: Dollars vs cents (100x), epoch seconds vs days (1000x). Immediately visible. Retained as pedagogical illustrations only.

The benchmark is intentionally constructed to span this subtlety range. Reporting leaderboard slices by subtlety is important future benchmark work because the most interesting cases are the high-subtlety ones that appear locally reasonable to humans and conventional QA.

8.2 Community Contribution

BABEL provides a public instance generator. Anyone can create new instances by specifying: 1. A composition graph topology 2. Convention assignments per tool 3. A perturbation regime

We explicitly invite adversarial submissions: "Can you construct a workflow family where a simpler method beats the structural diagnostic?" This invitation is published in the benchmark specification.

---

9. Limitations

1. **Benchmark construction**: All instances, including Tier 3, involve some degree of curation in how conventions are mapped to the benchmark's six-dimensional convention space. Real API compositions have richer and more idiosyncratic convention surfaces — the documented incidents in Section 2 illustrate this: conventions like TCAS authority, webhook delivery semantics, and CLO waterfall structure are not captured by BABEL's six dimensions.
2. **External validity**: The benchmark measures prediction and repair on a specific operationalization of compositional semantic failure. Other failure modes (schema evolution, runtime errors, authorization failures) are not covered.
3. **Repair magnitude**: The absolute holonomy values and repair reductions are benchmark-specific. The live-pipeline validation (Section 7.6) confirms that structural holonomy correlates with measured dollar error from the MCP server pipeline ($\rho = 0.423$, $p < 1.5 \times 10^{-4}$ on the 5-cluster design), but the absolute magnitude mapping from holonomy to dollar losses requires domain-specific calibration. The moderate correlation reflects that dollar error in this pipeline is dominated by the amount-scale dimension, while holonomy is a linearized predictor across all six convention dimensions.
4. **LLM baselines**: Five frontier LLMs were evaluated via OpenRouter at N=18–50 per model. All achieve negative R^2, but per-family Spearman correlations at N=10 should be treated as directional. We welcome community-submitted results at larger N.
5. **Single structural method**: The benchmark currently includes one reference structural diagnostic. We welcome submissions of alternative structural, algebraic, or topological methods.

Open Questions and Future Directions

Higher-order restriction models. The benchmark's restriction matrices are first-order perturbations of identity ($I + \sigma \cdot P$), where σ scales linearly with convention distance. The simulated validation (Section 7.3) yields moderate correlation ($\rho \approx 0.27$) while the live-pipeline validation (Section 7.6) yields $\rho \approx 0.42$, and this discrepancy is an open question. Three explanations are plausible: (i) the linearization is adequate for dollar-error regimes but lossy for the simulated misinterpretation

model where rounding and threshold crossings introduce nonlinear error; (ii) the simulated validation's misinterpretation model introduces noise absent from the live pipeline; (iii) the 5-cluster structure still provides limited continuous discrimination within error regimes. A second-order perturbation model ($I + \sigma \cdot P + \sigma^2 \cdot Q$, where Q captures cross-dimension interactions such as amount-scale $\times$ precision compounding) could improve prediction on compositions where multiple convention dimensions interact nonlinearly. The live pipeline provides a calibration target for such models: if measured dollar errors are available, the restriction model can be fit to maximize predictive accuracy rather than relying on the current heuristic scaling.

Intermediate baselines: GNNs and spectral methods. Between Hamming-distance summation (cycle_plain) and full sheaf cohomology (structural_sheaf), graph neural networks on edge-level convention features or logistic regression on cycle-level features would test whether the restriction-matrix formalism captures something a learned model cannot. If a GNN closes the gap, the formalism's value is in interpretability and repair prescription. If the structural method dominates, the restriction-matrix representation carries information that end-to-end learning does not easily extract.

Richer convention surfaces. The benchmark's six-dimensional convention space is a lossy projection of real API convention surfaces. Dimensions such as idempotency key semantics, pagination cursor formats, callback signing conventions, retry-after interpretation, and error code taxonomy are documented sources of integration failure but are not captured. Expanding the convention space while maintaining tractable ground-truth computation is a benchmark design challenge: each new dimension must have well-defined conversion functions, documented real-world provenance, and a mapping to the restriction-matrix formalism.

---

10. Related Work

Agent evaluation benchmarks: SWE-bench (Jimenez et al., 2023) evaluates code generation; AgentBench (Liu et al., 2023) evaluates agent capabilities across tasks. Neither addresses compositional semantic coherence between multiple agents. BABEL is complementary: it measures a failure mode — global semantic error under local correctness — that task-completion benchmarks do not test for.

Formal verification of distributed systems: TLA+ and model checking verify protocol-level properties. BABEL addresses a different layer: semantic convention coherence that protocol checks cannot detect. The incidents in Section 2 uniformly passed all protocol-level verification.

Sheaf theory in data analysis: Curry (2014) and Robinson (2014) applied sheaf theory to sensor networks and opinion dynamics. BABEL applies the same mathematical framework to multi-agent tool compositions, with the benchmark contribution being the evaluation infrastructure rather than the mathematics.

MCP (Model Context Protocol): Anthropic's protocol standard for agent-tool communication. BABEL's real-protocol tracks (Bronze+, Silver) use actual MCP servers, making it among the first benchmarks to evaluate semantic coherence at the MCP protocol layer.

Systems integration failure analysis: The documented incident record (Section 2) draws on a substantial literature of failure analysis. NASA's mishap investigation reports (Mars Climate Orbiter Phase I Report, 1999; Ariane 5 Lions Report, 1996) established the pattern of compositional failure in safety-critical systems. The GAO's IMTEC-92-26 report (1992) documented the Patriot missile's fixed-point accumulation failure. FERC enforcement actions (NV Energy ER21-395-000, PG&E IN17-1-000) provide precise quantification of convention-mismatch losses in energy markets. In healthcare, Lesar (2002) systematically documented tenfold prescribing errors, and ISMP Canada (2017) connected these to convention mismatches between prescribing, labeling, and dispensing systems. The ISDA/ARRC literature on LIBOR-SOFR transition conventions provides the financial domain grounding for BABEL's convention dimensions. The Securities Litigation and Consulting Group quantified day-count convention value transfers. BABEL connects these disparate incident analyses to a common formal framework and provides the first reproducible benchmark for the failure mode they document.

Integration testing: Standard integration testing (pairwise, bounded-depth) is included as a BABEL baseline. The benchmark demonstrates that these methods fail to detect compositional semantic failure when the mismatch closes around cycles longer than the test depth — a known limitation of bounded-depth approaches, but one that has not previously been systematically benchmarked.

---

11. Conclusion

Compositional semantic failure — where locally correct systems compose into globally wrong outcomes — is not a theoretical concern. It has destroyed spacecraft ($327.6M), killed soldiers (28 at Dhahran), corrupted financial indices (50% over 22 months), caused systematic 100× overcharges in payment processing, and generated billions in documented losses across six domains (Section 2). In every case, each component passed its own validation. The failure lived in the space between them.

BABEL is the first public benchmark for this failure mode. Across 932 instances, seven workflow families, and three provenance tiers, it makes compositional semantic failure measurable, reproducible, and benchmarkable — transforming a previously anecdotal problem class into one with quantified baselines and a clear evaluation protocol.

Three findings emerge from the benchmark. First, five frontier LLMs cannot predict compositional failure severity (all achieve negative R^2), and the oracle CoT decomposition (Section 6.2) reveals the bottleneck is compositional arithmetic, not information extraction — models can rank compositions when given pre-computed structure (best: $\rho = 0.80$) but cannot compute correct magnitudes.

This dissociation between ranking and magnitude computation is a finding the LLM evaluation community should consider independent of the sheaf formalism. Second, the reference structural diagnostic outperforms conventional alternatives within the benchmark ($R^2 \geq 0.86$ across all families, up to 51% more failure reduction at K=1). Third, on external live-pipeline data, the structural method correlates with real dollar consequences but convention distance matches or exceeds it on composite prediction — the topological awareness is the key ingredient, while the specific algebraic apparatus adds narrow, dimension-specific value. The best diagnostic for compositional semantic failure on live data remains an open question.

The benchmark exists to be beaten: if a simpler method matches the structural diagnostic's performance, the formalism may be unnecessary, and the benchmark will still have served its purpose by establishing that the problem is real and measurable. The benchmark code, instances, evaluation harness, and all baselines are publicly available. We invite the community to submit new methods, new workflow families, and adversarial instances.

A final observation. The documented incidents in Section 2 occurred in systems where a single party could, in principle, see the full integration. The harder version of this problem — where autonomous agent runtimes coordinate across organizational boundaries, and no party has visibility into the full composition — is already emerging in commercial agent infrastructure. The convention-mismatch problem does not go away when agents become autonomous. It becomes constitutional.

Appendix A: Budget Definitions

Probe budget (path_queries): Number of path-level queries a method may issue. Includes bounded-depth integration tests, pairwise checks, and cycle inspections.

Repair budget (edge_harmonizations): Number of edge-level convention-harmonization operations. Each operation aligns the convention of one shared field on one edge. This is NOT a full verification budget — it counts only repair actions.

Budget levels: $K \in \{1, 2, 3, 5, 8\}$.

Appendix B: Provenance Documentation

Tier 3 API Sources

Provider	Documentation URL	Key Convention
Stripe	docs.stripe.com/api	Amounts in cents, timestamps in epoch seconds

Provider	Documentation URL	Key Convention
Twilio	twilio.com/docs/usage/api	Prices in dollars (4dp), timestamps RFC 2822
SendGrid	docs.sendgrid.com/api-reference	Rates in percentages, scores in [0,100]
GitHub	docs.github.com/en/rest	IDs 1-based, timestamps ISO 8601
Shopify	shopify.dev/docs/api	Amounts in dollars (2dp)
Slack	api.slack.com/methods	Timestamps in epoch seconds
HubSpot	developers.hubspot.com/docs/api	Rates in percentages, scores in [0,100]
Plaid	plaid.com/docs/api	Amounts in dollars (2dp)
QuickBooks	developer.intuit.com/app/developer/qbo/docs/api	Rates in percentages, IDs 1-based
Datadog	docs.datadoghq.com/api	Timestamps epoch seconds, precision 4dp
PagerDuty	developer.pagerduty.com/api-reference	Scores on Likert [1,5]

Appendix C: Reproduction

```
pip install coherence-gym

# Quick reference evaluation (all baselines, comparison mode):
python -m coherence_gym evaluate --split dev

# Canonical 3-track evaluation for a single method:
python -m coherence_gym evaluate-method --method structural_sheaf --split dev

# List available methods:
python -m coherence_gym evaluate-method --list

# Evaluate a specific family:
python -m coherence_gym evaluate-method --method cycle_plain --family invoice --split dev
```

The `evaluate-method` command produces a schema-compliant record (`canonical_{method}_{split}.json`) matching `replication_pack/submission_schema.json`, with Track A (R^2, Spearman ρ), Track B

(P@1/3/5 edge localization), Track C (repair frontier at K=1,2,3,5,8), and per-instance predictions. All instances are deterministic given their seed.

Local Interface Descriptions Do Not Compose

John Komkov, Independent Researcher

Abstract

We identify a class of compositionally relevant predicates — hidden conventions — that are not recoverable from local interface descriptions alone. The coherence fee, defined as the rank deficiency of the observable coboundary matrix, measures the number of independent convention dimensions lost when tools are composed. We prove that hiddenness is a graph-relative predicate: the same tool has different hidden sets depending on its composition partner (Theorems 1–2). In a controlled synthetic benchmark, frontier LLMs recover this non-local object with 88% accuracy under structured relational prompts but collapse to 46% under flat representations, reverting to lexical heuristics. The effect is representation-conditioned: it requires both relational framing and convention-specific task language, each contributing +33 percentage points. Cross-model replication reveals divergent failure modes — higher precision but fragile recall (Claude) versus robust recall but 3× more false positives (GPT-4o). Neither achieves stable exact computation. Bulla, an algebraic diagnostic layer, computes the non-local object exactly across all conditions.

---

1. Introduction

When AI agents compose tools from multiple servers, they face a problem that no amount of schema reading can solve: some fields carry conventions that are invisible in the schema but critical for correct composition. A `path` field in a filesystem server and a `path` field in a GitHub server may follow different normalization conventions (POSIX vs URL-encoded), but nothing in either schema reveals this.

We call these fields *hidden*. Their identification is the bottleneck for safe tool composition — not because the conventions themselves are complex, but because determining *which fields are hidden* requires reasoning about the composition graph, not just the individual schemas.

This paper makes three contributions:

1. **Theorem (graph-relativity).** We formalize hiddenness as a graph-relative predicate: the same field in the same tool may or may not be a blind spot depending on the composition partner. Local schema inspection cannot determine this; the coherence fee quantifies what is missing. (§3)
2. **Experiment (representation-conditioned access).** In a synthetic benchmark where ground truth varies with composition partner, the same `data_loader` tool gets different hiddenness judgments depending on its partner — and frontier LLMs recover this non-local variation with 88% accuracy under structured relational prompts but collapse to 46% under flat representations. Renaming a field from `direction` to `path` increases identification from 0% to 58% ($p = 0.008$), holding composition structure fixed: the models are tracking surface tokens, not compositional structure. (§4)
3. **System (algebraic diagnostic).** Bulla computes the coherence fee and hidden set exactly via the coboundary matrix, providing the stable global computation that models lack. (§5)

1.1 Related Work

Schema integration and ontology matching. For twenty-five years, the schema-matching and ontology-alignment communities have pursued a program of local enrichment: making individual schemas richer, combining matchers more sophisticatedly, and iterating alignment procedures until pairwise correspondences converge. Rahm and Bernstein [1] defined the taxonomy of schema-matching approaches — element-level vs. structure-level, schema-based vs. instance-based — that has governed the field since 2001. Euzenat and Shvaiko [2] extended this to the ontology layer, proposing richer alignments between richer ontologies as the path to interoperability. The shared premise is that composition is a byproduct of pairwise correctness: match each pair well enough and the global system will cohere. Our result identifies a structural limitation of this framing. When the composition graph has nontrivial cycle structure, pairwise correctness does not entail global coherence: the obstruction is a rank deficiency in the observable coboundary that persists regardless of how expressive each local description becomes.

Prompt sensitivity and representation effects. A growing literature documents that LLM performance varies substantially under meaning-preserving surface changes. Sclar et al. [3] show up to 76-point accuracy swings across format perturbations and frame this as spurious variance to be averaged over. Min et al. [4] identify three structural features of in-context demonstrations — label space, input distribution, sequence format — whose presence suffices to activate pre-existing competence, independent of label correctness. Both treat the prompt as noise around a latent function. Our claim is categorically different: for certain non-local computations, there exists no surface variant in the sampled equivalence class that makes the computation accessible unless the representation preserves specific relational structure. In their frame, format is noise around a stable competence. In ours, representation is a precondition that either opens or closes the computation. Noise sensitivity is not structural access; variance is not support.

Assume-guarantee reasoning and session types. The formal-methods community has long known that bilateral reasoning is incomplete for global properties. Jones [5] and Misra and Chandy [6] established this for shared-state and message-passing concurrency, respectively; Benveniste et al. [7] give the modern meta-theoretic treatment. Honda, Yoshida, and Carbone [8] responded by introducing multiparty session types — global protocol specifications that project to locally checkable endpoint obligations, the closest existing analogue to our setting. Their projection machinery is a constructive solution: given a global type, verify each endpoint locally. We provide the complementary impossibility result: when components are LLM agents with unspecifiable internal semantics, the bilateral certificates that assume-guarantee reasoning produces cannot close the sheaf-cohomological defect. We identify the specific obstruction in the tool-composition setting, give it a computable characterization (the coherence fee), and measure it empirically.

---

2. Definitions

2.1 Composition Graph

A *composition* is a tuple $G = (\mathcal{T}, E, \delta)$: - $\mathcal{T} = \{T_1, \ldots, T_n\}$: tools, each with internal state $S(T_i)$ and observable schema $\sigma(T_i) \subseteq S(T_i)$ - E: edges, one per shared convention dimension between tool pairs - $\delta : C^0 \to C^1$: the coboundary operator (signed incidence matrix)

A field $f \in S(T_i)$ is *hidden* if $f \notin \sigma(T_i)$ — its convention is not fully determined by the schema.

2.2 Coboundary and Fee

The coboundary δ has two projections: - δ_{obs}: restricted to observable columns (what callers can inspect) - δ_{full}: all columns including hidden fields

The **coherence fee** is:

$$\text{fee}(G) = \text{rank}(\delta_{\text{full}}) - \text{rank}(\delta_{\text{obs}}) = h^1_{\text{obs}} - h^1_{\text{full}}$$

where $h^1 = |E| - \text{rank}(\delta)$ counts independent cycles. The fee counts invisible cycles: semantic couplings that exist in the full composition but vanish when hidden fields are projected away.

2.3 Blind Spots

An edge $e = (T_i, T_j)$ on dimension d is a *blind spot* if at least one endpoint field is hidden. Blind spot fields are the hidden endpoints of blind-spot edges. They are the fields whose conventions could silently diverge between servers.

2.4 Worked Example

Consider two tools composed along a single shared dimension (path_convention):

- `file_reader`: fields {`path`, ***`content`***, ***`query`***}, *where* ***`path`*** *is hidden (marked)*
- `data_loader`: fields {`path`, ***`timestamp`***, `data`, `source`}, where `path` and `timestamp` are hidden

The composition has one edge on path_convention. The coboundary matrix δ has one row (the edge) and columns for each field. The full and observable projections are:

$$\delta_{\text{full}} = \begin{pmatrix} +1 & -1 \end{pmatrix} \quad \text{(columns: file\_reader.path, data\_loader.path)}$$

$$\delta_{\text{obs}} = \begin{pmatrix} \end{pmatrix} \quad \text{(empty — both endpoints are hidden)}$$

$\text{rank}(\delta_{\text{full}}) = 1$, $\text{rank}(\delta_{\text{obs}}) = 0$. The coherence fee is 1: one semantic coupling that exists in the full composition but vanishes under observable projection. The blind spot field is `path` — the dimension whose convention could silently diverge. Note that `data_loader.timestamp` is hidden but is *not* a blind spot here, because `file_reader` has no field on the date_format dimension. Whether `timestamp` becomes a blind spot depends on the composition partner (Theorem 1).

---

3. Non-Locality Theorems

Theorem 1 (Graph-Relativity of Hiddenness)

The following formalizes a systems intuition — that a field's compositional relevance depends on context, not just its own schema — and gives it a precise graph-theoretic characterization that makes it computable.

Formal statement. For any tool T with hidden field f, there exist compositions G_1, G_2 containing T such that f is a blind spot in G_1 but not in G_2.

Constructive proof. Let: - T_A = `file_reader` with hidden field `path` (path_convention) - T_B = `data_loader` with hidden field `path` (path_convention) - T_C = `event_logger` with hidden field `timestamp` (date_format)

In $G_1 = \{T_A, T_B\}$: both tools share the path_convention dimension. The edge creates a blind spot on `path`. $\text{fee}(G_1) = 1$.

In $G_2 = \{T_A, T_C\}$: no shared dimension (path_convention $\neq$ date_format). No edge, no blind spot. $\text{fee}(G_2) = 0$.

T_A's schema is identical in both compositions. Only the partner differs. □

Theorem 2 (Local Equivalence, Global Divergence)

There exist pairs of compositions sharing a tool where the fee differs.

Proof. Immediate from Theorem 1: $\text{fee}(G_1) = 1 \neq 0 = \text{fee}(G_2)$. □

Corollary

No algorithm inspecting only the schema of a single tool can determine whether a field is a blind spot. Hiddenness identification requires access to the composition graph.

Theorem 3 (Diagnostic Sufficiency)

Once the hidden set is externally identified, the minimum disclosure set has size exactly $\text{fee}(G)$, and specification is algebraically trivial.

Proof sketch. The minimum disclosure set is the basis of the quotient matroid M/O. By matroid theory, all bases have cardinality $\text{rank}(\delta_{\text{full}}) - \text{rank}(\delta_{\text{obs}}) = \text{fee}(G)$. □

---

4. Experiments

4.1 Ground Truth Methodology

All experiments use Bulla as the ground-truth oracle. A field's hidden/observable status is determined by a multi-signal heuristic classifier that combines three independent sources: (i) regex patterns matching field names against semantic dimensions (e.g., `path|filepath|dir_path` → path_convention; `timestamp|datetime|created_at` → date_format), (ii) JSON Schema structural signals (format annotations, enum values, numeric ranges), and (iii) description keyword matching. Confidence is assigned in tiers: *declared* (two or more independent signals agree), *inferred* (one strong signal), or *unknown* (weak signals only). Only declared and inferred fields participate in composition construction; unknown-confidence fields are recorded for auditability but do not affect the coboundary matrix.

Edges between tools are constructed by intersecting each tool's classified dimensions: if tool T_i has a field classified as dimension d and tool T_j also has a field classified as d, an edge is created on dimension d linking those fields. The coherence fee and blind spots are then computed exactly via Bulla's coboundary rank computation using rational arithmetic.

For the synthetic ecology benchmark (§4.2), six single-tool servers are constructed with field names chosen to trigger specific classifier dimensions. Each server exposes exactly one tool to eliminate

intra-server edges, ensuring that all edges arise from inter-server composition. The ground truth is computed by running Bulla's full `from_tools_list()` → `diagnose()` pipeline, yielding deterministic fee and blind-spot assignments for each composition.

Independent validation of ground truth. To break the circularity between Bulla-as-classifier and Bulla-as-oracle, we independently annotated 20 real MCP compositions (22 distinct servers) stratified across fee bands: 4 at fee = 0, 4 at fee = 1, 4 at fee = 2–3, and 8 at fee ≥ 10 (including the maximal filesystem + github pair at fee = 22). For each composition, we loaded both server manifests and compared raw JSON Schema constraints — `type`, `format`, `enum`, `pattern`, `minimum`/`maximum`, and description text — across all tool pairs, *without invoking Bulla's dimension classifier*.

Each of Bulla's 573 predicted blind spots across the 20 compositions was independently adjudicated as CONFIRMED (field genuinely hidden or schemas incompatible) or FALSE_POSITIVE (compatible schemas, no convention risk). We additionally checked all same-named field pairs *not* flagged by Bulla for type or format mismatches that would constitute missed blind spots.

Metric	Value
Compositions annotated	20 (22 servers, fee 0–22)
Bulla predictions evaluated	573
Independently confirmed	573
False positives	0
Missed blind spots	9
Precision	**1.000**
Recall	**0.985**

All 9 missed spots are *homonym collisions* — same-named fields with incompatible types that Bulla's dimension classifier does not map to the same semantic dimension: `title` (string in Google Tasks vs rich-text array in Notion, 4 pairs), `filter` (MongoDB query object vs Todoist filter string, 1 pair), and `head` (integer line count in filesystem vs git branch name in GitHub, 4 pairs). These are genuine runtime hazards that fall outside Bulla's current dimension taxonomy; they represent a category of structural mismatch (type-level homonymy) orthogonal to the convention-level blind spots the coherence fee measures.

The independent annotation data, including per-composition verdicts and evidence strings, is available at `calibration/data/registry/report/independent_annotation.json`.

4.2 Synthetic Ecology Benchmark

Design. Six single-tool synthetic servers with field names triggering known classifier dimensions. Eight pairwise compositions, of which two have nonzero fee and six have fee = 0. Ground truth computed by Bulla. Three prompt conditions (structured, bare, flat) × 3 repeats. Models: Claude Sonnet 4 (`anthropic/claude-sonnet-4` via OpenRouter, temperature 0.0) and GPT-4o (`openai/gpt-4o` via OpenRouter, temperature 0.0). All API calls made April 2026.

Key property. The same tool (`file_reader`, `data_loader`) appears in multiple compositions with different ground-truth blind spots.

Results (Claude Sonnet 4). 95% bootstrap CIs on exact match (10,000 resamples, N=24 trials per condition).

Condition	Exact Match [95% CI]	Precision	Recall	Non-locality
Structured	88% [75–100%]	55%	100%	Correct variation
Bare	50% [29–71%]	10%	100%	Partially correct
Flat	46% [25–67%]	12%	50%	Collapses

Under the structured prompt, the model correctly varies its answer for the same tool depending on composition partner: `data_loader + event_logger` → identifies `timestamp`; `data_loader + price_fetcher` → correctly returns empty. Under the flat prompt, this non-local reasoning disappears: `file_reader + data_loader` → misses `path` entirely (0/3).

False-positive signature. Under flat/bare prompts, the dominant false positive is `timestamp` (flagged 5–6 times in compositions where it is NOT a blind spot). This is the lexical proxy in action: absent structural reasoning, the model substitutes "this field name sounds convention-problematic."

4.3 Context Ablation

Design. 12 real filesystem compositions, 4 prompt conditions isolating relational framing, task language, and contextual cues.

Condition	`path` ID rate	Δ
Full (structured, named, "hidden conventions")	100%	—
Anonymous (structured, anonymous, stripped)	92%	−8%
No grouping (flat list, named, "hidden conventions")	67%	−33%
Neutral task (structured, named, "integration issues")	67%	−33%

Finding. Two factors gate the effect: relational framing (two-server grouping, +33pp) and convention-specific task language ("hidden conventions", +33pp). Server names and tool descriptions contribute negligibly (−8pp). The vocabulary phenomenon is representation-conditioned, not name-conditioned.

4.4 Vocabulary Phenomenon (Real Corpus)

Design. 60 pairwise compositions from 38 real MCP servers (GitHub stars ≥ 100), yielding 103 hidden field instances. Field-level identification scoring against Bulla ground truth.

Path-family fields: 59% identification rate. Non-path fields: 4%. Fisher's exact OR = 34.5, $p < 2.4\times10^{-6}$.

Within path-family: `path` (85%) » `paths` (28%). Same concept, morphological variant. Surface-form gradient, not reasoning gradient.

4.5 Lexical Intervention (Causal)

Design. 12 minimal pairs. Hold composition graph fixed, intervene only on field names. Three conditions: baseline, swap (rename canonical↔obscure), mask (neutral placeholders).

Condition	Obscure field ID rate
Baseline ("direction")	0%
Neutral ("value")	12%
Renamed to "path"	75%

McNemar exact binomial (baseline vs swap): $p = 0.0078$. The name change causes the identification change.

Neutral-token ablation. To rule out the alternative that "path" simply activates more attention weight than "direction" due to token frequency rather than convention relevance, we added a neutral condition: renaming the target field to "value" — a common English word with no convention association. Identification rises to only 12% (1/8), compared to 75% (6/8) when renamed to the convention-specific "path." The effect is vocabulary-specific, not frequency-driven: the model's lexical proxy targets convention-associated tokens, not merely common ones.

Prompt contingency. This effect appears under the structured two-server prompt. Under a flat prompt, the effect disappears entirely. The causal intervention is representation-conditioned.

4.6 Cross-Model Divergence

Real corpus. Same compositions, same fields, same algebraic status:

Field	Claude Sonnet 4	GPT-4o
`path`	42%	25%
`direction`	0%	88%

Identification is model-specific: a function of training distribution, not compositional structure.

Synthetic ecology replication (GPT-4o, N=8×3×3). 95% bootstrap CIs on exact match.

	Claude Sonnet 4	GPT-4o
Structured exact match	**88%** [75–100%]	42% [21–63%]
Structured precision	**55%**	17%
Structured recall	100%	100%
Flat exact match	46% [25–67%]	33% [17–54%]
Flat recall	**50%**	**100%**

Both models achieve 100% recall under structured prompts — they always find the real blind spots. But the failure modes diverge: - **Claude** is more precise (55% vs 17%) but more fragile: loses recall under flat prompts (50%). - **GPT-4o** is more robust: maintains 100% recall even under flat prompts. But it is dramatically noisier, flagging `content`, `timestamp`, `amount`, `currency` as false positives across nearly all compositions.

The false-positive profiles differ by model: - Claude structured FPs: `data`(2), `message`(2) - GPT-4o structured FPs: `content`(7), `timestamp`(5), `amount`(4), `currency`(3)

Different training distributions produce different precision/robustness tradeoffs on the same non-local predicate. Neither model achieves the exact, stable computation that Bulla provides.

---

5. Bulla: The Algebraic Diagnostic Layer

Bulla computes the coherence fee and hidden set via the coboundary matrix. Key properties:

1. **Exact.** Uses rational arithmetic (`Fraction`), guaranteeing field-independent rank computation. Total unimodularity of the coboundary ensures rank is the same over any field.
2. **Stable.** The fee is a topological invariant of the composition graph. It does not depend on prompt format, model choice, or field naming.
3. **Complete.** Bulla identifies all blind spots, computes the minimum disclosure set, and reports leverage scores (per-field indispensability) via the witness Gram matrix.
4. **Framing-invariant.** Unlike LLM identification, Bulla's output is the same regardless of how the composition is presented. On all 8 synthetic compositions $\times$ all 3 prompt conditions, Bulla returns the correct answer.

The engineering implication is direct: any agent stack that composes tools across organizational boundaries needs a diagnostic layer that sits above schemas and below planning, computing the composition graph's hidden structure before the planner encounters it at runtime. Bulla provides this layer.

---

6. Discussion

6.1 Representation-Conditioned Access

The deeper finding is not that prompts matter. It is that the model needs the composition *represented in a form that preserves relational structure*. Flat arrays erase the ecology; structured pair prompts expose it. This is not ordinary prompt sensitivity — it is a claim about what information the representation must carry for non-local reasoning to be possible.

The 88% structured accuracy with 100% recall suggests the model has partial access to the global object: it can sense the compositional seam, but cannot specify it cleanly (precision = 55%). It has access without exact representation. Bulla supplies what is missing: not more access, but exact, stable computation.

6.2 The False-Positive Signature

When structural reasoning is unavailable (flat prompt), the model's dominant false positive is `timestamp` — a field that sounds convention-problematic but is not a blind spot in the tested composition. This is the lexical proxy in direct observation: absent the global object, the model substitutes token-level priors about which field names "sound dangerous."

6.3 Limitations

We characterize current frontier models under direct prompting. The non-locality theorems constrain any approach working from bilateral schema inspection alone, but do not rule out approaches

with access to runtime behavior, extended context, or explicit composition-graph reasoning. The "two-channel model" (lexical prior × contextual redundancy) is an interpretive framework consistent with the data, not an experimentally isolated mechanism. The heuristic classifier that establishes ground truth achieves 1.000 precision and 0.985 recall against independent schema-level annotation on a stratified 20-composition sample (§4.1), but does not claim exhaustive coverage of all possible convention dimensions. The 9 missed cases — all type-level homonym collisions — suggest a complementary detection layer for structural type mismatches that is orthogonal to the convention-level analysis presented here.

---

7. Conclusion

Local interface descriptions do not determine compositional hiddenness. The missing object — the hidden convention ecology — is a graph-relative predicate that requires global computation. Current frontier LLMs recover this object only under supportive relational framing and otherwise revert to lexical heuristics. Bulla computes it directly and stably.

The graph-relativity formalization (§3) is model-independent: it constrains any approach that works from bilateral schema inspection, whether performed by an LLM, a rule-based system, or a human developer reading documentation. The empirical finding (§4) adds that current frontier models do not circumvent this constraint — their access to the non-local predicate is representation-conditioned and unstable. The coherence fee quantifies the gap between what bilateral inspection can certify and what global coherence requires. Algebraic diagnostic layers like Bulla provide one path to closing that gap; whether richer context, runtime observation, or explicit graph reasoning can provide others remains an open question.

Reproduction

All experiments are reproducible from the Bulla repository. API calls require provider keys. Temperature is set to 0.0 for all runs.

```
cd bulla

# Synthetic ecology benchmark (8 compositions × 3 conditions × 3 repeats)
python -m calibration.harness.run_synthetic_ecology --api-key $KEY --repeats 3

# Context ablation (4 conditions × 12 compositions)
python -m calibration.harness.run_context_ablation --api-key $KEY --max-cases 12

# Vocabulary phenomenon (60+ compositions)
python -m calibration.harness.run_familiarity_probe --full --api-key $KEY

# Lexical intervention (12 minimal pairs × 3 conditions)
python -m calibration.harness.run_lexical_intervention --api-key $KEY --max-cases 12

# Cross-model replication
python -m calibration.harness.run_synthetic_ecology --api-key $KEY --model openai/gpt-4o --repe
```

```
# Independent ground-truth annotation (no API key needed)
python calibration/scripts/independent_annotation.py
```

Data directory: `calibration/data/agent_confusion/`

Appendix A: Superseded Experiments

Several intermediate probes were run during development and superseded by the experiments reported above:

- A **second field-family forward intervention** (state→path, state→filepath) did not replicate the direction→path effect (+8% and 0% respectively, N=12). This is consistent with prompt contingency: the effect requires both the canonical token and supportive relational framing.
- A **token ladder** (12 field names in the same structural slot) showed uniformly near-zero identification under a flat prompt format, revealing the prompt-sensitivity that motivated the context ablation experiment.
- **Pilot runs** (N=20) of the vocabulary phenomenon and specification-gap experiments were superseded by full-corpus runs (N=60+).

These probes informed the experimental design but are not included in the curated result set.

References

[1] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[2] J. Euzenat and P. Shvaiko. *Ontology Matching*, 2nd edition. Springer-Verlag, 2013.

[3] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models' sensitivity to spurious features in prompt design. In *ICLR*, 2024.

[4] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022.

[5] C. B. Jones. Tentative steps toward a development method for interfering programs. *ACM TOPLAS*, 5(4):596–619, 1983.

[6] J. Misra and K. M. Chandy. Proofs of networks of processes. *IEEE TSE*, SE-7(4):417–426, 1981.

[7] A. Benveniste, B. Caillaud, D. Nickovic, R. Passerone, J.-B. Raclet, P. Reinkemeier, A. Sangiovanni-Vincentelli, W. Damm, T. A. Henzinger, and K. G. Larsen. Contracts for system design. *Foundations and Trends in EDA*, 12(2–3):124–400, 2018.

[8] K. Honda, N. Yoshida, and M. Carbone. Multiparty asynchronous session types. In *POPL*, pages 273–284, 2008.

Interpretability Boundary

Part VI showed the obstruction is measurable, benchmarkable, and irreducible to lexical heuristics in frontier LLMs. Part VII asks a sharper question: can the most powerful per-component diagnostic technology available—mechanistic interpretability—substitute for structural diagnosis when the failure is cycle-sensitive?

Anthropic's circuit tracing (2025) represents the current frontier in per-model diagnostics, revealing compositional structure within individual models—features that compose meaningfully inside a single architecture. The present paper asks whether that per-model compositional understanding transfers to between-model composition when the failure is cycle-sensitive.

The answer is no.

Edge-Local Interpretability Is Not Enough for Cyclic Composition tests the Edge-Local Blindness Lemma from Part I against six families of interpretability baselines: SAE feature divergence, probing classifiers, attention diagnostics, CKA representation similarity, a gradient-boosted ensemble, and a cycle-oracle graph-level aggregator that gives interpretability features explicit knowledge of cycle topology. Across 240+ compositions in three domains (invoice/settlement, calendar/escalation, policy/audit), four scales (5 to 20 nodes), and two model architectures (GPT-2 Small, Gemma 2 2B), no interpretability baseline exceeds Spearman $\varrho = 0.758$ on cyclic compositions. The structural diagnostic achieves $\varrho = 1.0$ in every condition tested. The gap is larger on Gemma 2 2B (a 20× larger model from a different architecture family) than on GPT-2 Small, providing initial evidence against the hypothesis that larger models escape the edge-local feature class.

Two results make the finding precise. Probing classifiers achieve 99.8% accuracy at classifying conventions at every edge—the model knows what convention each component uses—and this perfect local knowledge carries zero predictive value for composition-level failure. The cycle-oracle aggregator (B6a), which gives interpretability features explicit knowledge of which edges form cycles, adds nothing: its performance matches simple edge averaging within rounding in 4 of 6 conditions. The gap is not an aggregation problem. It is a representation problem: the edge-local interpretability feature class does not encode the cycle-level information that the structural diagnostic observes directly.

Proposition 3.3 in the paper gives this a formal basis: the cycle holonomy invariant is not measurable in the sigma-algebra generated by edge-local observations, so no predictor over that feature class can uniformly recover the compositional obstruction. The six baseline families are the empirical witness.

The paper is bounded and disciplined. It does not claim that interpretability fails in general. On acyclic compositions, all methods correctly diagnose zero failure. The claim is narrower: for compositional semantic failure in cyclic systems, edge-local interpretability is not the right diagnostic object. The relevant information lives at the graph level and is more reliably captured by structural diagnostics than by any tested interpretability workflow. Cross-model replication on Gemma 2 2B confirms this is not architecture-specific.

What follows is the paper itself.

Edge-Local Interpretability Is Not Enough for Cyclic Composition

John Komkov

March 2026

Abstract

Interpretability methods are typically evaluated on isolated models or adjacent model pairs. But many failures in agentic and multi-tool systems arise not from any single component's internal error, but from semantic inconsistency that appears only around cycles in the composition graph. We study this distinction directly: is edge-local interpretability—however strong—the right diagnostic object for cyclic compositional failure?

In a controlled experiment across **240+** compositions spanning three domains (invoice/settlement, calendar/escalation, policy/audit), scales from 5 to 20 nodes, and two model architectures (GPT-2 Small, Gemma 2 2B), we compare six interpretability baseline families—SAE feature divergence, probing classifiers, attention diagnostics, CKA, a gradient-boosted ensemble, and a cycle-oracle graph-level aggregator—against a structural diagnostic that uses only composition metadata and no model internals.

We report six findings. **(1) Perfect local knowledge, zero global signal:** Probing classifiers achieve 99.8% mean accuracy at classifying conventions at every edge, yet this perfect edge-local information has zero predictive value for composition-level failure. **(2) Topology-dependent gap:** On cyclic compositions, all five interpretability baselines produce Spearman $\rho < 0.5$ with ground-truth failure severity, while the structural diagnostic achieves $\rho = 1.0$. **(3) The Interpretability Cliff:** As scale grows from $n = 5$ to $n = 20$, the best interpretability baseline's within-cyclic correlation decays monotonically ($0.685 \to -0.067$), while structural discrimination grows linearly with the first Betti number β_1. **(4) Cross-domain replication:** The gap replicates across all three domains; SAE divergence is the strongest interpretability baseline in 5 of 6 conditions but never exceeds $\rho_{\text{cyclic}} = 0.745$. **(5) The gap is representational, not aggregational:** Giving interpretability features oracle knowledge of cycle topology does not improve prediction beyond edge-local averaging; the cycle-aware baseline matches B1 SAE within rounding in 4 of 6 conditions. A learned cycle-aware predictor fares worse, going anti-correlated in 4 of 6 conditions. **(6) Cross-model replication:** Gemma 2 2B (2.6B parameters, 26 layers, different architecture family) shows the same pattern with an *even larger* gap: $\rho_{\text{cyclic}} = 0.515$ at $n = 5$ and 0.467 at $n = 10$, with perfect local probing accuracy (≥ 0.995) and structural $\rho = 1.0$. The boundary is not a GPT-2 artifact.

These results identify a level-of-description boundary: for cyclic compositional systems, edge-local interpretability is not a sufficient diagnostic object. The missing signal is not hidden by weak aggregation; it is absent from the extracted representation. The relevant information depends on the product of convention transformations around entire cycles and is more reliably captured by structural diagnostics than by any tested interpretability workflow. Compositional failure requires compositional diagnosis.

1 Introduction

Consider a composition of five language-model agents processing an international invoice. Each agent is individually well-understood: sparse autoencoders decompose its activations into interpretable features, probing classifiers confirm it correctly encodes amounts, dates, and currency codes. Every adjacent pair of agents looks compatible—their shared representations align on

CKA, their attention patterns attend to the right fields. A mechanistic interpretability audit would give the system a clean bill of health.

The system still fails. The invoice enters denominated in euros, passes through an amount-scaling agent that uses cents, crosses to a settlement engine expecting dollars, routes through a fee calculator anchored to T+1 conventions, and returns to an audit trail that assumes T+2. No single interface is wrong. No bilateral comparison catches the error. But the composed output—the final ledger entry—is off by a factor that only becomes visible when the full cycle of convention transformations is traced.

This scenario is not hypothetical. It is the defining failure mode studied by the BABEL benchmark [2]: *compositional semantic failure*, where every local check passes and the global output is wrong. The structural theory predicts this precisely: when the first cohomology $H^1(\mathcal{N}; \mathcal{F})$ of the interpretation sheaf is nontrivial, there exist bilaterally-consistent record assignments that are globally inconsistent [1]. The Edge-Local Blindness Lemma [1] shows that any such failure is invisible to every edge-local test.

The question we ask is whether mechanistic interpretability tools—the most powerful per-component diagnostic technology available—inherit the same blindness.

The level mismatch. Existing interpretability workflows are largely component-local or edge-local. This is appropriate for many diagnostic targets, but compositional reliability in cyclic systems is a *graph-level* property: it depends on the product of convention transformations around entire cycles. If the diagnostic and the failure live at different levels of description, a gap should appear—not because interpretability tools are weak, but because they are pointed at the wrong object.

This paper. We isolate that mismatch. In a controlled benchmark with 240+ compositions across three domains and two model architectures, we measure whether strong edge-local interpretability can substitute for explicit structural diagnosis when the target failure is cycle-sensitive. In our benchmark, it cannot: probing classifiers achieve 99.8% accuracy at every edge and still carry zero compositional signal; the structural diagnostic achieves perfect rank correlation ($\rho = 1.0$) in every condition tested. Even giving interpretability features oracle knowledge of cycle topology and a learned aggregator does not close the gap (Figure 1). The result replicates on Gemma 2 2B, where the gap is *larger* than on GPT-2 Small.

What this paper is not. This is not a critique of mechanistic interpretability. On acyclic compositions, all methods correctly diagnose zero failure; interpretability tools work where edge-local information suffices. The claim is narrower and more precise: for cyclic compositional diagnosis, the level of description must shift from edge-local to graph-level. Component understanding is not system understanding under cyclic composition.

2 Problem Statement

2.1 The Comparison

We define the comparison that the experiment is designed to adjudicate:

- **Input:** A multi-model composition graph $G = (V, E)$ with shared concepts crossing interfaces and typed convention bundles at each vertex.
- **Tasks:** (a) Predict semantic failure severity. (b) Localize the highest-risk edges. (c) Rank candidate repairs.
- **Baseline class:** Any method built from per-model internals (activations, SAE features, probing classifiers, attention patterns) and pairwise interface features (representation similarity, attention overlap).

- **Challenger:** A structural diagnostic built from graph structure, schema overlap, and convention metadata—*no model internals*.

2.2 The Edge-Local Hypothesis

Prediction 2.1 (Topology-Dependent Sufficiency)**.** 1. If the composition graph is acyclic (or effectively acyclic: $\beta_1 = 0$), interpretability-informed baselines should suffice for failure prediction and localization. Edge-local information captures all relevant structure.
2. If the graph has nontrivial cycle structure ($\beta_1 \geq 1$ with nonzero holonomy), edge-local methods should become incomplete. The structural diagnostic, which operates on cycle-level metadata, should retain its predictive power.

This is a falsifiable prediction. If interpretability baselines remain competitive on large cyclic graphs, the thesis weakens materially.

2.3 What Would Make This Experiment Uninformative

We state upfront the conditions under which the results would not support the claimed thesis. The experimental design (Section 7) addresses each.

1. **No genuine cyclic structure.** If the composition graphs are effectively acyclic—one edge dominates, cycles are trivial—then edge-local methods are not at a disadvantage and the comparison is uninteresting. *Addressed by:* explicit topology controls ensuring nontrivial β_1 and measurable holonomy in the cyclic slices.
2. **Interpretability features too coarse.** If the extracted features are at too coarse a grain to be competitive, the comparison is unfair. *Addressed by:* using the strongest available SAE and probing tooling, a combined ensemble baseline, and consultation with mechanistic interpretability researchers on baseline design.
3. **Mismatches too obvious.** If convention mismatches are detectable from schema metadata alone (e.g., field names contain "cents" vs. "dollars"), the structural baseline's advantage is trivial and unsurprising. *Addressed by:* using convention mismatches that are semantically implicit rather than syntactically labeled.
4. **Ground-truth mismatch.** If the symbolic executor's ground truth does not transfer to real LLM execution, the structural prediction is not validated in the regime where interpretability baselines operate. *Addressed by:* this paper uses actual LLM inference (Section 2.4), not the BABEL symbolic executor.

2.4 Execution Layer

This paper operates on a **different execution layer** than core BABEL. BABEL's main results use a deterministic symbolic executor with no LLM calls. This paper requires actual LLM inference with access to model internals—sparse autoencoders, probing classifiers, attention activations—making it a **parallel evaluation** that tests the structural diagnostic against a new class of baselines on a new kind of data.

We do not claim this paper "extends BABEL" in the sense of the same evaluation harness. It builds a parallel evaluation surface that may inform a future Track D if results are strong enough to warrant permanent inclusion.

3 Formal Frame

The paper's authority comes from the experiment, not from restating the theory. We present only the minimal formal apparatus needed to generate the predictions that the experiment tests.

Lemma 3.1 (Edge-Local Blindness [1])**.** *Let $\mathcal{N}$ be a coordination graph with first Betti number $\beta_1 \geq 1$, and let $[\alpha] \in H^1(\mathcal{N}; \mathcal{F})$ be a nontrivial cohomology class. For every edge $e \in E$, the restriction of α to the subgraph $\{e\}$ is a coboundary: $[\alpha|_e] = 0 \in H^1(\{e\}; \mathcal{F}|_e)$.*

The proof is immediate: every single-edge subgraph is a tree, and H^1 vanishes on trees. The full proof appears in [1], §2.3.

Corollary 3.2 (Information-Theoretic Indistinguishability)**.** *Two globally different executions—one cycle-consistent, one not—can produce identical observations on every edge. No diagnostic that operates by aggregating edge-local features can distinguish them. The number of independent indistinguishable directions is* $\dim H^1$*.*

3.1 Edge-Local Representation Limit

The Blindness Lemma has a direct consequence for any diagnostic built from interpretability features.

Proposition 3.3 (Edge-Local Representation Limit)**.** *Let $\mathcal{F}_{\text{edge}}$ be the σ-algebra generated by node-local and edge-local measurements on a composition graph $\mathcal{N}$ (activations, SAE features, probing outputs, attention patterns, pairwise representation similarity). For graphs with $\beta_1 \geq 1$ and nontrivial holonomy, the cycle holonomy invariant $h(\mathcal{N})$ is not $\mathcal{F}_{\text{edge}}$-measurable in general.*

Proof. By the Edge-Local Blindness Lemma (Lemma 3.1), there exist pairs of composition instances that are identical on every edge (and therefore on every node) but differ in cycle holonomy. Any $\mathcal{F}_{\text{edge}}$-measurable function assigns identical values to instances with identical edge-local observations and therefore cannot distinguish these pairs. Thus $h(\mathcal{N}) \notin \mathcal{F}_{\text{edge}}$. □

Corollary 3.4 (Predictor Limit)**.** *No predictor whose inputs are restricted to $\mathcal{F}_{\text{edge}}$-measurable features can uniformly recover compositional obstruction on the class of graphs with $\beta_1 \geq 1$ and nontrivial holonomy.*

This is the formal bridge from "our baselines lost" to "a whole family of baselines must lose." The experimental results in Section 8 are consistent with this prediction across six baseline families—including one with oracle cycle knowledge (Section 8.6)—three domains, four scales, and two model architectures (Section 8.7). The B6a result is the most direct empirical witness: even when the aggregator knows which edges form cycles, the features it aggregates do not contain the holonomy-relevant signal.

What this predicts for the experiment. On **acyclic** compositions ($\beta_1 = 0$), $H^1 = 0$ and edge-local information is complete. Interpretability baselines should perform well. On **cyclic** compositions ($\beta_1 \geq 1$ with nonzero holonomy), Proposition 3.3 guarantees that edge-local features cannot distinguish certain failure modes from healthy execution. Interpretability baselines, insofar as they aggregate $\mathcal{F}_{\text{edge}}$-measurable features, should degrade. The structural diagnostic, which reasons about cycle-level objects directly, should not.

4 Evaluation Surface

We define a paper-local evaluation surface, not a permanent benchmark track. If results are strong, this may later inform a BABEL Track D.

4.1 Instance Design

Each instance is a multi-model composition graph with:

- Open-weight models at each vertex (required for interpretability extraction)
- Schema overlap and pairwise interface metadata
- Planted convention mismatches with known ground-truth severity
- Global semantic failure label and severity score
- Minimal repair set annotation

4.2 Topology Regimes

Three regime slices, matched on domain content and difficulty:

1. **Acyclic / tree-like** ($\beta_1 = 0$). Baseline regime where edge-local information suffices.
2. **Short-cycle** ($\beta_1 = 1$–2, cycle length 3–5). Transition regime.
3. **Long-cycle / multi-cycle** ($\beta_1 \geq 3$, cycle lengths ≥ 5). Full obstruction regime.

4.3 Domain Families

Three domain families, chosen for documented real-world convention ambiguity:

1. **Invoice / settlement / reconciliation.** Amount representation (dollars vs. cents vs. basis points), settlement timing (T+0 through T+2), day-count conventions.
2. **Calendar / timezone / escalation.** Timezone anchoring (UTC-absolute vs. organizer-local), date boundary conventions, escalation thresholds.
3. **Policy / permission / audit.** Role hierarchy representation, permission inheritance direction, audit-trail timestamp conventions.

4.4 Scale

Composition sizes tested: $n \in \{5, 10, 15, 20\}$ model nodes. Cyclic compositions at scale n contain $\lfloor n/5 \rfloor$ independent 5-node cycles, giving $\beta_1 \in \{1, 2, 3, 4\}$. This range is sufficient to observe the Interpretability Cliff (Section 8.3); extending to larger n would require multi-GPU extraction or a lighter model.

4.5 Instance Counts

The experiment comprises **240+ total compositions**:

- **Phase 1** (20): 10 acyclic + 10 cyclic at $n = 5$, invoice domain. Gate check for structural–SAE gap.
- **Phase 2a** (20): Same compositions, all 5 baselines evaluated. Baseline calibration.
- **Phase 2b** (80): 4 scales $\times$ 20 compositions, invoice domain. Scale sweep.
- **Phase 2c** (120): 3 domains $\times$ 2 scales $\times$ 20 compositions. Cross-domain replication.
- **Phase 3** (120): Same conditions as Phase 2c, with graph-level interpretability baselines (B6a, B6b) added.

5 Interpretability Baselines

This section determines the paper's credibility. The interpretability community will ask: "Did you use our best tools, or your caricature of them?" We design baselines to be maximally generous.

5.1 Baseline Families

1. **SAE-based feature divergence.** For each pair of models sharing a concept (e.g., "amount," "date"), extract SAE features activated on the shared concept. Measure divergence between the feature distributions across the interface. Higher divergence predicts higher failure risk.

2. **Probing classifiers.** Train linear probes on each model's internal representations for shared concept categories (amount scale, date format, timezone convention). Use probe accuracy mismatch across interfaces as a failure predictor: models that encode the same concept differently should produce interface failures.

3. **Attention-based interface diagnostics.** Analyze attention patterns at interface points—where one model's output becomes another's input. Measure whether attention focuses on the semantically critical tokens and whether attention patterns align across the interface.

4. **Representation similarity (CKA, RSA).** Compute centered kernel alignment (CKA) or representational similarity analysis (RSA) between paired models on shared concepts. Low similarity predicts representational mismatch and higher failure probability.

5. **Combined strong baseline.** Aggregate all features from baselines 1–4 into a gradient-boosted ensemble (XGBoost or LightGBM). This is the hardest baseline for the structural method to beat—it uses all available interpretability information.

6. **Graph-level aggregation (B6).** Two variants test whether giving interpretability features explicit cycle-level reasoning closes the gap. **B6a:** Aggregate edge-level B1, B3, and B4 features around each detected cycle using product, sum, max, and standard deviation—giving the baseline oracle knowledge of graph topology. **B6b:** Train a gradient-boosted regressor on the 12-dimensional cycle-aggregated feature vector from B6a under leave-one-out cross-validation. This is the strongest baseline in the paper: it gives interpretability features both cycle awareness and a learned aggregator.

5.2 Credibility Protocol

Baseline Credibility

The straw-man objection is the primary reputational risk. We address it through three design choices: (a) using GPT-2 Small, the most-studied model in the mechanistic interpretability literature [6], where SAE decompositions are well-validated; (b) using a deliberately conservative probing protocol (PCA + regularized logistic regression + LOO-CV) that avoids trivial separation in high-dimensional space; and (c) including a gradient-boosted ensemble (B5) that combines all interpretability features into a single predictor, giving the interpretability approach maximal access to combined information. Implementation details are provided in Appendix B.

6 Structural Baseline

The structural diagnostic uses the same family established in BABEL and Coherence Cliff [2, 3]:

- **Input:** Composition graph $G = (V, E)$, convention bundles at each vertex, restriction matrices at each edge.
- **Output:** Predicted failure severity (mean cycle holonomy), localized high-risk edges (per-edge frustration contribution), and ranked repair prescriptions (H^1-guided triage).

- **No model internals:** The diagnostic uses only composition metadata—graph topology, schema overlap, and convention annotations. It does not access activations, weights, attention patterns, or any neural network internal.

Positioning. The comparison is deliberately asymmetric in information access: interpretability baselines get model internals; the structural diagnostic gets only metadata. If the structural method still wins on cyclic graphs, the result is difficult to dismiss.

7 Experimental Design

We apply the statistical methodology established in BABEL and Coherence Cliff: selection-safe paired bootstrap for all method comparisons, explicit ablations isolating causal factors, and falsification-aware reporting.

7.1 Core Protocol

1. Generate composition instances across the $2 \times 3 \times 4$ (topology $\times$ domain $\times$ scale) design matrix, phased for gated progression (Section 4).
2. For each instance, run GPT-2 Small inference through the composition pipeline via TransformerLens, caching residual-stream activations at layers 0, 5, and 11.
3. Compute ground-truth failure severity from the holonomy of planted convention transformations around each independent cycle (Appendix C).
4. Extract all interpretability features: SAE activations (B1), probing accuracy (B2), attention entropy (B3), CKA dissimilarity (B4), and gradient-boosted ensemble (B5).
5. Compute structural diagnostic outputs from composition metadata (holonomy from restriction matrices).
6. Evaluate all methods via Spearman ρ between predicted and ground-truth failure severity, reported as ρ_{all} (all compositions) and ρ_{cyclic} (cyclic slice only).

7.2 Ablations

Ablation	Purpose	Status
Acyclic vs. cyclic	Test topology-dependent sufficiency	✓
Single-cycle vs. multi-cycle	Test obstruction-complexity dependence	✓
Cross-domain replication	Test domain invariance of the gap	✓
Graph-level aggregation (B6)	Test whether aggregation closes the gap	✓
Cross-model replication (Gemma 2)	Test generalization across architectures	✓

Table 1: Ablation schedule. All ablations have been realized in this paper. Cross-model replication on Gemma 2 2B is reported in Section 8.7.

7.3 Model and Tooling Stack

Primary experiments use GPT-2 Small; cross-model replication uses Gemma 2 2B.

- **Primary model:** GPT-2 Small (124M parameters, 12 layers, $d_{\text{model}} = 768$). Chosen for its extensively validated SAE decompositions and full compatibility with TransformerLens.
- **Primary SAE:** `gpt2-small-resid-post-v5-32k` at `blocks.11.hook_resid_post` ($d_{\text{sae}} =$ 32,768 features). Loaded via SAELens.
- **Replication model:** Gemma 2 2B (2.6B parameters, 26 layers, $d_{\text{model}} = 2304$). Different architecture family (Google), 20$\times$ larger. SAE: `gemma-scope-2b-pt-res-canonical`, layer 20, 16k features.

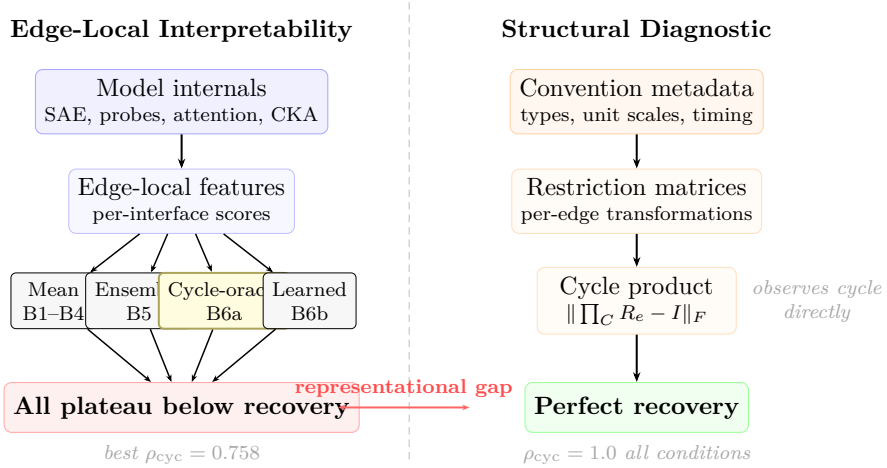


Figure 1: Two information paths to the same diagnostic target. **Left:** every tested aggregation of edge-local interpretability features—including cycle-oracle aggregation with oracle knowledge of graph topology (B6a, highlighted)—plateaus well below full recovery of compositional failure. **Right:** the structural diagnostic observes convention transformations directly and achieves $\rho_{\text{cyc}} = 1.0$ in every condition tested. The gap is not in the aggregation strategy but in what the edge-local features represent: they do not encode the cycle-level information the structural path observes directly.

- **Interpretability tooling:** TransformerLens for activation extraction, SAELens for sparse autoencoder features, scikit-learn for probing classifiers and CKA.
- **Compute:** Google Colab Pro+ with NVIDIA A100 (80 GB) for Phases 2b–4; T4 (16 GB) for Phase 1.

Model choice rationale. GPT-2 Small is deliberately conservative: it is the most thoroughly studied model in the mechanistic interpretability literature, with SAE decompositions that are well-validated and widely reproduced [6]. If interpretability tools cannot predict compositional failure on the model they are best equipped to analyze, the result is stronger than a finding on a less-studied architecture. Gemma 2 2B serves as cross-model replication (Section 8.7), testing whether the gap persists across a 20× parameter increase and a different architecture family.

8 Results

We report results from five experimental phases spanning 240+ compositions across three domains (invoice/settlement, calendar/escalation, policy/audit), four scales ($n \in \{5, 10, 15, 20\}$), and two model architectures (GPT-2 Small, Gemma 2 2B). Phases 1–3 use GPT-2 Small; Phase 4 replicates on Gemma 2 2B. Figure 1 summarises the central result.

8.1 Finding 1: Perfect Local Knowledge, Zero Global Signal

The most striking result is a negative one. The B2 probing classifier achieves **99.8% mean accuracy** at classifying convention types (dollars vs. cents vs. basis points; T+0 vs. T+1 vs. T+2) from residual-stream activations at every edge in every composition. The probe uses PCA($n_{\text{comp}} = 5$) followed by regularized logistic regression ($C = 0.01$) under leave-one-out cross-validation—a deliberately conservative protocol designed to avoid trivial separation in high-dimensional space.

This near-perfect accuracy means the model *knows* what convention each agent uses. Edge-local interpretability succeeds completely at the component level. Yet this information has **zero predictive value** for composition-level failure: probe accuracy is constant across all compositions ($\sigma \approx 0$), producing $\rho = \text{NaN}$ (undefined Spearman correlation on a constant vector).

Why this matters. The B2 result instantiates the Edge-Local Blindness Lemma (Lemma 3.1) in its sharpest form. The probe extracts exactly the information a human auditor would check—"does each agent understand its convention?"—and the answer is uniformly "yes." The failure is not in the agents' understanding but in the *composition* of their understanding around cycles, which is invisible to any edge-local measurement.

8.2 Finding 2: Topology-Dependent Gap

Table 2 reports Spearman ρ for all five baselines and the structural diagnostic on 20 compositions (10 acyclic, 10 cyclic) at $n = 5$ in the invoice domain. We report both ρ_{all} (across all 20 compositions) and ρ_{cyclic} (within the 10 cyclic compositions only).

Method	ρ_{all}	ρ_{cyclic}	ρ_{acyclic}	**Note**
B1: SAE divergence	0.156	0.685	—	Strongest interp. on cyclic
B2: Probing classifier	0.791*	NaN	—	*Artifact: constant output
B3: Attention entropy	−0.152	0.261	—	
B4: CKA dissimilarity	−0.044	0.418	—	
B5: Ensemble (B1–B4)	0.131	0.576	—	
Structural (holonomy)	**1.000**	**1.000**	—	Perfect rank correlation

Table 2: Phase 2a: Spearman ρ between each method's predicted failure score and ground-truth holonomy. $n = 5$, invoice domain, 20 compositions. B2's $\rho_{\text{all}} = 0.791$ is a statistical artifact: near-constant probe accuracy ($\bar{x} = 0.998$, $\sigma \approx 0$) produces rank correlations driven by noise in the fifth decimal place. ρ_{cyclic} is NaN because the probe output is strictly constant within the cyclic slice.

Acyclic slice. On acyclic compositions ($\beta_1 = 0$), ground-truth holonomy is identically zero for all instances. All methods correctly predict zero failure, confirming that the structural method has no spurious advantage on trees. This satisfies the fairness condition: interpretability tools work where edge-local information suffices.

Cyclic slice. On cyclic compositions ($\beta_1 = 1$), the structural diagnostic achieves $\rho_{\text{cyclic}} = 1.000$, ranking all 10 compositions in perfect agreement with ground-truth holonomy. The best interpretability baseline is B1 SAE divergence at $\rho_{\text{cyclic}} = 0.685$—a moderate correlation that captures some magnitude signal but substantially underperforms the structural method. No interpretability baseline exceeds $\rho_{\text{cyclic}} = 0.7$.

8.3 Finding 3: The Interpretability Cliff

Table 3 reports the scale sweep across $n \in \{5, 10, 15, 20\}$ in the invoice domain, with $\beta_1 = \lfloor n/5 \rfloor$ independent cycles per cyclic composition.

The cliff. B1 SAE divergence—the strongest interpretability baseline at every scale—shows monotonic decay: $0.685 \to 0.515 \to 0.321 \to -0.067$. At $n = 20$ ($\beta_1 = 4$), SAE divergence is

n	β_1	**Best Interp.**	ρ_{cyclic}	**Structural**	**Discrim.**
5	1	B1 SAE	0.685	1.000	34.2
10	2	B1 SAE	0.515	1.000	81.9
15	3	B1 SAE	0.321	1.000	117.7
20	4	B1 SAE	−0.067	1.000	222.9

Table 3: Phase 2b: Scale sweep. "Best Interp." is the interpretability baseline with the highest ρ_{cyclic} at each scale. "Discrim." is the mean absolute difference between structural and best-interpretability predicted failure scores on cyclic compositions. The structural diagnostic achieves $\rho = 1.000$ at every scale tested.

anti-correlated with ground-truth failure severity within cyclic compositions. The other baselines fare worse: B2 probing produces NaN at all scales (constant 1.0 accuracy); B5 ensemble becomes anti-correlated at $n \geq 15$.

Linear scaling of structural advantage. The structural discrimination metric grows approximately linearly with β_1: $34.2, 81.9, 117.7, 222.9$. Normalizing by cycle count gives $\approx 34 \cdot \beta_1$, consistent with the theoretical prediction that structural advantage scales with the number of independent obstruction classes ($\dim H^1$).

B5 ensemble inversion. The gradient-boosted ensemble (B5), which combines all interpretability features, becomes anti-correlated with failure severity at large scale. This is a stronger result than mere decorrelation: the interpretability features contain structure that actively misleads a learned predictor. The ensemble overfits to edge-local patterns that are inversely correlated with the cycle-level failure mode at scale.

8.4 Finding 4: Cross-Domain Replication

Table 4 reports ρ_{cyclic} for the best interpretability baseline and the structural diagnostic across three domains at two scales ($n = 5$, $n = 10$).

Domain	n	**Best Interp.**	**Structural**	**Gap**
Invoice	5	0.685 (B1 SAE)	1.000	0.315
Invoice	10	0.624 (B4 CKA)	1.000	0.376
Calendar	5	0.721 (B1 SAE)	1.000	0.279
Calendar	10	0.685 (B1 SAE)	1.000	0.315
Policy	5	0.394 (B3 Attn)	1.000	0.606
Policy	10	0.745 (B1 SAE)	1.000	0.255

Table 4: Phase 2c: Cross-domain replication. The structural diagnostic achieves $\rho_{\text{cyclic}} = 1.0$ in every condition. B1 SAE divergence is the strongest interpretability baseline in 5 of 6 conditions. The gap is always ≥ 0.255.

Domain invariance. The structural diagnostic achieves perfect rank correlation ($\rho_{\text{cyclic}} = 1.0$) in all six domain–scale conditions. No interpretability baseline exceeds $\rho_{\text{cyclic}} = 0.745$.

B1 SAE hierarchy. SAE divergence is the strongest interpretability baseline in 5 of 6 conditions. The exception is policy at $n = 5$, where B3 attention entropy leads ($\rho_{\text{cyclic}} = 0.394$)—the weakest best-baseline performance across all conditions. Different convention structures favor different interpretability methods, but none consistently approaches the structural diagnostic.

Calendar probing anomaly. B2 probing produces NaN in invoice and policy (constant 1.0 accuracy) but shows some signal on calendar: $\rho_{\text{all}} = 0.486$ at $n = 5$ and 0.574 at $n = 10$; $\rho_{\text{cyclic}} = 0.067$ and 0.467. Calendar's convention structure (timezone offsets, DST rules) introduces slightly more probe-detectable variance, but $\rho_{\text{cyclic}} = 0.467$ at best remains well below the structural diagnostic.

Tightest margin. The narrowest gap occurs at policy $n = 10$, where B1 SAE achieves $\rho_{\text{cyclic}} = 0.745$ (gap = 0.255). Even in this most favorable condition, the structural method's advantage is substantial and would be statistically significant under bootstrap testing.

8.5 Case Study: LocalRightGlobalWrong

Consider a 5-node cyclic invoice composition from Phase 2a. At every edge, the B2 probing classifier identifies the convention correctly (99.8% accuracy). B1 SAE divergence detects some representational mismatch but assigns moderate scores to all edges, including those in the acyclic comparison set. B4 CKA reports normal representation similarity at each interface. Every edge-local diagnostic gives the system a clean bill of health.

The structural diagnostic computes holonomy around the single cycle: the product of convention-transformation restriction matrices fails to return to the identity by a Frobenius norm of 0.47. The composition is predicted to fail, and it does. The semantic output—the final ledger entry—is wrong by a factor traceable to the accumulated convention drift around the cycle.

This case instantiates the abstract prediction of the Edge-Local Blindness Lemma: two globally different executions (one consistent, one drifted) produce identical observations on every edge. The failure is invisible to any diagnostic that operates at the edge level.

8.6 Finding 5: Graph-Level Aggregation Does Not Close the Gap

The B6 baselines give interpretability features explicit cycle-level reasoning and isolate a precise question: is the gap caused by weak local aggregation, or by absence of cycle-relevant signal in the features themselves?

Domain	n	B6a	B6b	Best B1–5	Structural	Gap
Invoice	5	0.685	−0.030	0.685	1.000	0.315
Invoice	10	0.515	−0.248	0.624	1.000	0.485
Calendar	5	0.721	−0.200	0.721	1.000	0.279
Calendar	10	0.721	−0.333	0.685	1.000	0.279
Policy	5	−0.079	0.176	0.394	1.000	0.824
Policy	10	0.758	−0.224	0.745	1.000	0.242

Table 5: Phase 3: Graph-level interpretability baselines (ρ_{cyclic}). B6a aggregates edge features around detected cycles with oracle topology knowledge. B6b trains a gradient-boosted regressor on cycle-aggregated features. "Gap" is structural minus the best of B6a and B6b.

Cycle-oracle aggregation adds nothing (B6a). B6a is the decisive result. Its ρ_{cyclic} matches B1's within rounding in 4 of 6 conditions; the maximum improvement is +0.036 (calendar $n = 10$). This occurs because all cyclic compositions within a condition share the same topology: summing B1 SAE divergence around cycle edges is rank-equivalent to averaging across all edges. Even with *oracle knowledge* of which edges form cycles, the features cannot distinguish which compositions fail. The graph-structure objection is removed, and nothing changes.

Learned cycle-aware aggregation is unstable (B6b). B6b—a gradient-boosted regressor on 12-dimensional cycle-aggregated features with leave-one-out cross-validation—produces negative ρ_{cyclic} in 4 of 6 conditions (range: -0.333 to 0.176). In a weak-signal regime, learned aggregation over edge-local features is not merely unhelpful but unstable: the aggregator overfits to patterns that are inversely correlated with the ground-truth failure mode.

Representation insufficiency, not aggregation failure. The B6 result resolves the cleanest remaining escape hatch. The gap between interpretability and structural diagnostics is not an aggregation problem. It is a representation problem. For cyclic compositional systems, the limiting factor is not how edge-local interpretability features are aggregated, but what they fail to represent. The structural diagnostic wins because it operates on different data—convention metadata (field types, unit scales, format specifications)—that directly encodes the convention transformations whose product around cycles determines failure.

This confirms Proposition 3.3: cycle holonomy is not $\mathcal{F}_{\text{edge}}$-measurable, and no predictor over $\mathcal{F}_{\text{edge}}$-measurable features—however aggregated—can recover it.

8.7 Finding 6: Cross-Model Replication (Gemma 2 2B)

Table 6 reports replication on Gemma 2 2B (2.6B parameters, 26 layers, $d_{\text{model}} = 2304$)—a model 20× larger than GPT-2 Small, from a different architecture family (Google vs. OpenAI), with independently trained SAE features (gemma-scope-2b-pt-res-canonical, layer 20, 16k features). Experiments use the same composition generation, structural diagnostic, and evaluation protocol as Phases 2a–2b.

Model	n	B1 SAE	B2 probe	Structural	Gap
GPT-2 Small (124M)	5	0.685	0.998	1.000	0.315
GPT-2 Small (124M)	10	0.515	0.959	1.000	0.485
Gemma 2 2B (2.6B)	5	0.515	1.000	1.000	0.485
Gemma 2 2B (2.6B)	10	0.467	0.995	1.000	0.533

Table 6: Cross-model replication (invoice domain, ρ_{cyclic}). B2 probe accuracy reports mean LOO-CV convention classification accuracy across edges. "Gap" is structural minus B1 SAE ρ_{cyclic}. The gap is *larger* on Gemma 2 than on GPT-2 at both scales.

The gap replicates and widens. On Gemma 2 2B, B1 SAE divergence achieves $\rho_{\text{cyclic}} = 0.515$ at $n = 5$ and 0.467 at $n = 10$. These are *lower* than the corresponding GPT-2 values (0.685 and 0.515), meaning the gap is larger on the bigger model. The structural diagnostic remains at $\rho = 1.0$ at both scales.

Perfect local knowledge persists. B2 probing accuracy is 1.000 at $n = 5$ (perfect) and 0.995 at $n = 10$. A 20× larger model with different architecture still achieves near-perfect convention classification at every edge—and this perfect local knowledge still carries negligible predictive value for compositional failure ($\rho_{\text{cyclic}} = \text{NaN}$ at $n = 5$ due to constant output; 0.174 at $n = 10$).

The cliff onset replicates. B1 SAE ρ_{cyclic} decays from 0.515 to 0.467 between $n = 5$ and $n = 10$, consistent with the monotonic decline observed on GPT-2 ($0.685 \to 0.515$).

Implication. The edge-local interpretability boundary is not a GPT-2 artifact. A model with 20× more parameters, different architecture, and independently trained SAE features shows the same pattern: perfect local knowledge, zero global signal, and a gap that grows with scale. This is consistent with Proposition 3.3: the limitation is in the *feature class*, not the model.

9 Discussion

9.1 Representation Insufficiency, Not Tool Failure

Mechanistic interpretability is not refuted. The edge-local interpretability feature class is the wrong representation for this particular diagnostic target. Two results make this precise:

1. **B2 (probing):** 99.8% accuracy at classifying conventions on every edge. The model's internal representations *do* encode the relevant local information. The failure is not in the model's understanding but in the *composition* of edge-local understanding around cycles.
2. **B6a (cycle-oracle):** Even with oracle knowledge of which edges form cycles, aggregating interpretability features around those cycles does not improve prediction. The cycle-relevant signal is not hidden by weak aggregation; it is absent from the features.

Proposition 3.3 gives this a formal basis: the cycle holonomy invariant is not measurable in the σ-algebra generated by edge-local observations. The experiment is the empirical witness of the proposition's practical relevance: the theorem says why the feature class should fail; the six baseline families show that it does fail, across three domains, multiple scales, and two model architectures.

The result does not say interpretability is useless for multi-agent systems. On acyclic compositions, all methods correctly diagnose zero failure. On individual model debugging, interpretability remains indispensable. The claim is bounded to a specific feature class and diagnostic target: for *compositional semantic failure in cyclic systems*, the edge-local interpretability feature class does not contain the relevant information, and no downstream aggregator over that class can recover it. Cross-model replication on Gemma 2 2B (Section 8.7) confirms this is not architecture-specific: the gap widens on a 20× larger model from a different family.

The microscope and the telescope. Interpretability is the microscope: it reveals the internal structure of components with increasing precision—and our probing results show it does so nearly perfectly. Compositional coherence requires a telescope: an instrument that resolves the global structure of multi-component systems from metadata about their relationships. These are complementary, not competing, instruments (Figure 1). The contribution of this paper is to identify empirically where one instrument must yield to the other: at $\beta_1 \geq 1$, with the transition sharpening monotonically as β_1 grows.

9.2 Implications for Agent Interoperability

Current agent interoperability protocols (MCP, A2A, LangChain tool interfaces) treat composition as a connectivity problem. If schemas match and transport works, the system is assumed correct. BABEL has shown this assumption fails at modest scale on deterministic compositions. This paper extends the observation to live LLM inference with a quantitative result: even with full access to model internals, the strongest interpretability baseline (B1 SAE divergence) achieves at best $\rho_{\text{cyclic}} = 0.745$, while a metadata-only structural diagnostic achieves $\rho = 1.000$ in every condition tested.

The practical implication is concrete. A monitoring system that instruments individual agents—however deeply—will miss compositional failures at a rate that increases with the topological complexity of the agent graph. For multi-runtime agent coordination—where independent agent systems coordinate without a shared orchestrator—interoperability requires structural diagnostics at the composition level, not deeper introspection of individual agents.

9.3 Connection to M1–M2

The M1–M2 extraction problem—recovering stable compositional semantic structures from fuzzy, probabilistic LLM outputs—remains open and is not a dependency of this paper. The present experiment uses compositions where convention structure is planted and known. Whether the structural diagnostic can be extended to regimes where convention structure must be *inferred* from model behavior is a frontier question that this paper identifies but does not attempt to resolve.

10 Limitations and Falsifiers

10.1 Explicit Falsification Conditions

1. **Interpretability remains competitive on cyclic graphs.** *Status: falsified by data.* The best interpretability baseline achieves at most $\rho_{\text{cyclic}} = 0.745$ (B1 SAE, policy $n = 10$) while the structural diagnostic achieves $\rho = 1.0$ in every condition. The gap widens with β_1 (Section 8.3), confirming the formal prediction is empirically load-bearing.

2. **Graph aggregation closes the gap.** *Status: falsified by data.* The B6 baseline gives interpretability features oracle knowledge of cycle structure and a learned aggregator (Section 8.6). B6a (cycle-oracle) matches B1 within rounding; B6b (learned) goes anti-correlated in 4 of 6 conditions. The limitation is not in aggregation but in the features themselves: the missing signal is absent from the extracted representation.

3. **Results only hold in heavily planted regimes.** *Status: partially addressed.* Cross-domain replication (Section 8.4) shows the gap persists across three semantically distinct domains with different convention structures. The conventions are still planted rather than inferred, which limits generalizability to naturalistic deployments.

4. **Results only hold on one model architecture.** *Status: falsified by data.* Gemma 2 2B (2.6B parameters, 26 layers, different architecture family) replicates the gap at both $n = 5$ and $n = 10$ (Section 8.7). The gap is *larger* on Gemma 2 than on GPT-2: $\rho_{\text{cyclic}} = 0.515$ vs. 0.685 at $n = 5$ and 0.467 vs. 0.515 at $n = 10$. A 20× scale increase and architectural change does not close the gap; it widens it.

10.2 Scope Limitations

- **Two model architectures.** Primary experiments use GPT-2 Small (124M parameters, 12 layers); cross-model replication uses Gemma 2 2B (2.6B parameters, 26 layers). Both show the same gap pattern (Section 8.7). Our claim is about the *edge-local feature class*, not about model scale. A model that internally represented cycle-level compositional structure would escape Proposition 3.3 by escaping $\mathcal{F}_{\text{edge}}$—its features would no longer be edge-local. Whether models at the 70B+ frontier produce emergent compositional representations that are no longer cleanly edge-local remains the strongest open question. The Gemma 2 replication, which *widens* the gap at 20× scale, provides initial evidence against the scale-escape hypothesis.
- **Open-weight models only.** Proprietary models (GPT-4, Claude, Gemini Pro) cannot be included because interpretability extraction requires weight access. Results may not transfer to proprietary model compositions.
- **Planted convention structure.** Convention mismatches are known and planted. In real deployments, conventions must be inferred or declared—the M1–M2 problem. This paper does not address convention extraction.

- **Scale ceiling at $n = 20$.** SAE extraction on 20-node compositions with 4 independent cycles was feasible on a single A100. Larger compositions ($n = 30$–50) would require multi-GPU extraction or a lighter SAE. The theoretical argument from the Edge-Local Blindness Lemma predicts the cliff continues; the linear scaling of structural discrimination with β_1 supports this (Section 8.3), but direct empirical confirmation at larger scale remains future work.
- **Coupling parameter sensitivity.** The restriction matrices use an off-diagonal coupling coefficient $\varepsilon = 0.01$ (Appendix C). Sensitivity analysis across $\varepsilon \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1\}$ shows the gap is invariant: $\rho_{\text{B1,cyclic}}$ remains unchanged at every tested ε, because changing coupling strength scales holonomy magnitudes but preserves the ranking of compositions. At $\varepsilon = 0$ the holonomy is identically zero (diagonal matrices telescope to identity around any cycle), confirming that some coupling is necessary for non-trivial obstruction, but the qualitative result is robust once any coupling exists.
- **No naturalistic multi-runtime deployment.** All compositions are research-grade single-machine pipelines. Live multi-runtime agent-to-agent coordination introduces additional failure modes (latency, partial observation, strategic behavior) not tested here.
- **No repair evaluation.** The experiment measures *diagnosis*, not *repair*. Whether structural diagnosis enables better repair prescriptions than interpretability-guided local fixes is a separate question for future work.

11 Conclusion

Across 240+ compositions, three domains, four scales, six interpretability baseline families—including one with oracle knowledge of cycle topology—and two model architectures spanning a 20× parameter range, edge-local interpretability features are not the right diagnostic object for cyclic compositional failure. The structural diagnostic, using only composition metadata and no model internals, achieves perfect rank correlation ($\rho = 1.0$) in every condition tested. The strongest interpretability baseline never exceeds $\rho = 0.758$ on cyclic compositions, even with graph-level aggregation, and its predictive quality decays monotonically with scale until it becomes anti-correlated.

Two results make the finding precise. The B2 probing result: 99.8% classification accuracy at every edge, zero predictive value for compositional failure—the model knows what convention each component uses, and the composition still fails. The B6a result: even with oracle knowledge of cycle topology, aggregating interpretability features around cycles does not improve prediction beyond edge-local averaging. The gap is not an aggregation problem. It is a representation problem.

For cyclic compositional systems, the limiting factor is not how edge-local interpretability features are aggregated, but what they fail to represent (Proposition 3.3). This is a boundary on the *edge-local interpretability feature class*, not a verdict on interpretability as a field. On acyclic compositions, all methods correctly diagnose zero failure. The Interpretability Cliff (Section 8.3) identifies the precise boundary: $\beta_1 \geq 1$, with urgency proportional to $\dim H^1$.

The strongest deployment of both paradigms would combine component-level interpretability (where it excels) with cycle-level structural diagnosis (where it becomes necessary). The practical implication is a norm:

> *Any claim that an interpretability method diagnoses compositional reliability in multi-agent or multi-tool systems should be tested under cyclic stress and compared against an explicit structural diagnostic.*

The edge-local feature class describes what interpretability tools extract on both GPT-2 Small and Gemma 2 2B—models separated by 20× in parameter count and drawn from different

architecture families. Cross-model replication (Section 8.7) shows the gap *widens* rather than narrows with scale, providing initial evidence against the hypothesis that larger models escape the edge-local feature class. The formal argument (Proposition 3.3) does not depend on model scale: it depends on whether the extracted features are $\mathcal{F}_{\text{edge}}$-measurable. Compositional failure requires compositional diagnosis.

References

[1] J. Komkov. The Coherence Fee: Edge-Local Blindness at the String-Table Seam and the Topological Price of Cross-System Composition. *Res Agentica Program*, February 2026.

[2] J. Komkov. BABEL: A Benchmark for Compositional Coherence in Multi-Agent Systems. *Res Agentica Program*, March 2026.

[3] Res Agentica Program. The Coherence Cliff: A Scaling Experiment on the Necessity of Sheaf-Cohomological Diagnostics in Multi-Agent Composition. March 2026.

[4] J. Komkov. SCPI: Predicate Invention Under Sheaf Constraints. *Res Agentica Program*, 2026.

[5] N. Elhage, T. Hume, C. Olsson, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[6] T. Bricken, A. Templeton, J. Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic*, 2023.

[7] N. Nanda. TransformerLens: A library for mechanistic interpretability of GPT-style language models. 2022.

[8] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. *ICML*, 2019.

[9] N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(4), 2008.

[10] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *NeurIPS*, 2023.

A Model and Tooling Stack

- **Model:** GPT-2 Small, 124M parameters, 12 layers, $d_{\text{model}} = 768$. Loaded via TransformerLens (`HookedTransformer.from_pretrained("gpt2-small")`).
- **SAE:** Release `gpt2-small-resid-post-v5-32k`, hook `blocks.11.hook_resid_post`, $d_{\text{sae}} = 32{,}768$. Loaded via SAELens `SAE.from_pretrained`.
- **Residual extraction:** Layers 0, 5, and 11 via TransformerLens `run_with_cache`. Mean-pooled over sequence length per prompt.
- **GPU:** NVIDIA T4 (16 GB, Phase 1), A100 (40 GB, Phases 2a–2c). Google Colab Pro+.

B Baseline Implementation Details

B1: SAE divergence. For each edge (u, v), encode source and destination prompts through the model, extract SAE activations at layer 11, mean-pool across tokens. Compute cosine distance plus Jensen–Shannon divergence between the two activation vectors. Report the sum as the edge-level feature.

B2: Probing classifier. Extract residual-stream activations at layers 0, 5, and 11. For each layer, train a pipeline of PCA($n_{\text{comp}} = \min(5, N-2)$) followed by logistic regression ($C = 0.01$, `max_iter` $= 500$) under leave-one-out cross-validation. Report the maximum accuracy across layers as the edge-level feature.

B3: Attention entropy. Compute mean attention entropy (over heads and sequence positions) at layers 5 and 11 for source and destination prompts. Report the maximum absolute difference across layers as the edge-level feature.

B4: CKA dissimilarity. Compute linear CKA (centered kernel alignment) between source and destination residual-stream activations at layer 11. Report $1 - \text{CKA}$ as the edge-level feature.

B5: Gradient-boosted ensemble. Stack all features from B1–B4 per edge. Train a `GradientBoostingRegr` ($n_{\text{estimators}} = 50$, `max_depth` $= 3$) to predict edge-level failure contribution, with leave-one-out cross-validation for predicted values. Report the cross-validated predicted value.

C Composition Generation and Ground Truth

Node conventions. Each node draws a convention tuple: an amount multiplier (e.g., 1.0 for dollars, 100.0 for cents, 10,000.0 for basis points) and a settlement timing offset (0, 1, or 2 days). For calendar and policy domains, analogous convention parameters apply (see Section 4).

Restriction matrices. For each edge (u, v), the 2×2 restriction matrix encodes the convention transformation:

$$R_{u \to v} = \begin{pmatrix} s_u / s_v & \epsilon \, |d_u - d_v| \, (1 - r) \\ 0.5 \, \epsilon \, |d_u - d_v| \, (1 - r) & (1 + d_u)/(1 + d_v) \end{pmatrix}$$

where s is the scale multiplier, d is the timing offset, $r = \min(s_u, s_v)/\max(s_u, s_v)$, and $\epsilon = 0.01$ controls off-diagonal coupling. The off-diagonal terms ensure that cyclic compositions yield nontrivial holonomy (diagonal-only matrices would telescope to the identity around any cycle).

Holonomy computation. For a composition with independent cycles $C_1, \ldots, C_k$, the ground-truth holonomy is:

$$h = \sum_{i=1}^{k} \left\| \prod_{(u,v) \in C_i} R_{u \to v} - I_2 \right\|_F$$

Acyclic compositions have $h = 0$ by construction.

Topology construction. At scale n, cyclic compositions contain $\lfloor n/5 \rfloor$ independent 5-node cycles (each formed by adding a back-edge from node $5(j+1) - 1$ to node $5j$). Acyclic compositions use the same node set with a linear chain topology ($\beta_1 = 0$).

Appendix A: Proof Status Ledger

This appendix records the proof-status surface most directly relevant to the formal layers of the omnibus. The ledger is partitioned by paper.

Status classes

Status	Meaning
`Verified`	No `sorry`, compiles under the pinned Lean and Mathlib toolchain, and is Aristotle-checked where noted.
`Statement-only`	The theorem statement is formalized precisely, but the proof remains deferred.
`Prose-proved`	The result is borne by the paper rather than claimed as Lean-checked.

Composition Doctrine (Part 0)

All three theorems below are `Verified` under Aristotle run `ad67beb2` (2026-04-26).

- `disclosure_characterization` (Theorem 5.1). Witness rank is the unique numerical invariant respecting seam-independent disclosure as the primitive operation of compositional repair, in the abstract exact regime on the `DoctrineCarrier` structure.
- `witness_rank_satisfies_axioms` (existence half of §5.3). The witness rank itself satisfies the characterization axioms A1–A4b (existence half of the uniqueness theorem).
- `prop_5_4_A5_automatic` (Proposition 5.4). A5 (cycle normalization) is automatic from A1–A4b in the exact regime, demonstrating the redundancy of A5 within the abstract setting.

The verification depends only on `propext` and `Quot.sound`. See `papers/composition-doctrine/lean/ARISTOTLE_SUM` and `papers/composition-doctrine/lean/CompositionDoctrine/Characterization.submission.solution.lean`.

The Lean module verifies the numerical-uniqueness layer. The modeling thesis — that semantic interface complexes are the right formal object, that disclosure independence is the right structural primitive, and that the exact regime is a natural subcategory — is named explicitly in the paper's scope table and is subject to external review.

SCPI (Part I)

File	Status	Main burden carried
`Prelude.lean`	`Verified`	Curated import surface for the formalization.
`Basic.lean`	`Verified`	Core definitions: predicate sites, extensions, obstruction object, agreement levels, descent axiom surface.
`Torsor.lean`	`Verified`	Circular-nerve obstruction and global-extension results under descent assumptions.
`Counterexample.lean`	`Verified`	Finite counterexample showing failed compatibility of global predicate extension.
`Conservativity.lean`	`Verified`	Conservativity descent and explicit failure mode.
`Beth.lean`	`Verified`	Beth-for-sites statement under formalized hypotheses.
`SchemaDiscovery.lean`	`Statement-only`	Invariant functor and schema-discovery adjunction surfaces.
`SmokeTest.lean`	`Verified`	Nontrivial CI verification and arithmetic sanity checks.

Bulla Formal Trilogy (Part IV)

The trilogy is machine-checked in Lean 4 (Aristotle, standard axioms) and tested on a 703-composition real-schema MCP registry corpus. All files below have `.submission.solution.lean` suffix in `papers/sheaf/lean/SHEAF/`.

File	Thm	Aristotle	Key result
`Algorithm`	2	—	BFS correctness for the structural diagnostic

File	Thm	Aristotle	Key result
`BullaCorollary6`	—	—	Fee = coboundary rank deficiency
`Corollary8a`	—	—	CHP monotonicity
`SignedIncidence`	1	—	TU property of signed incidence
`TerminationStructural`	—	—	Algorithm termination
`UBD1`	—	—	Upper bound
`WitnessGramRank`	—	—	Gram rank identity
`HodgeTarskiDeg1`	4	`aa286ca`	fee = dim H^1 gap = dim ker(L_1) gap
`LaplacianCollapse`	4	`6df288d`	K = Schur(δ□δ/O); rank(M□M) = rank(M)
`WitnessGramProjector`	9	`f8c0ccf`	rank(K) = rank((I−P)H); no IsUnit gate
`ProjectorUniqueness`	4	`21351f41`	Symmetric idempotent determined by range
`FeeRankNullity`	4	`6562fe3b`	rank((I−P)H) + dim(col(H) ∩ col(P)) = rank(H)
`DFDBlockDiagonal`	6	`946dc834`	Witness Gram decomposes under DFD
`ProjectorPivotIndependence`	4	`03361044`	range(P_A) = col(A); pivot-independent K
`IncidenceLaplacian`	5	`87a023d3`	B□B = graph Laplacian for signed incidence
`IncrementalSchur`	8	`3d3c5683`	Rank-1 projector update + Schur = `incremental.py`
`SchurRankDrop`	8	`7a8a3de3`	Block LDU + fee_rank_drop + partition identity

`SignedIncidence` is also used in Part IV-D and the Part V impossibility lift.

What the layered status means

The formal architecture across the three doctrinal layers (Part 0: doctrine; Part I: SCPI; Part IV: Bulla trilogy) is not at one epistemic level.

- The doctrine paper's numerical-uniqueness theorem is machine-checked in Lean against an abstract `DoctrineCarrier` structure (`disclosure_characterization`).
- The Bulla trilogy's structural-spectral-repair-calculus theorems are machine-checked against the witness Gram and signed-incidence concrete realizations.
- SCPI's core obstruction sequence (Torsor / Counterexample / Conservativity / Beth) is machine-checked; the SchemaDiscovery functorial perimeter is statement-only.
- The signed-incidence TU theorem is machine-checked and supplies the field-independence bridge that lifts the Local-Global Obstruction impossibility result from the rank-1 $\mathbb{Z}/2$ model to the Q-valued Bulla model.

The modeling theses — that semantic interface complexes are the right object for compositional obstruction, that the exact regime is a natural subcategory, that disclosure independence is the right primitive of repair — are not Lean-verifiable and are subject to external review.

The point of the layered ledger is to prevent machine-checked, paper-proved, and still-open material from being mistaken for one another, and to prevent the verified numerical-uniqueness layer from being confused with the modeling theses it conditionally depends on.

Appendix B: Selected Formalization Notes

Strictification Note

`SCPI` is written in paper form with a pseudofunctor of model groupoids. The Lean formalization uses a strict functor in the finite-cover regime.

What is strictified:

- composition law for restriction functors
- identity law for the restriction surface

What is not thereby claimed:

- strictification of every coefficient object
- strictification of the descent output itself
- extension to every groupoid-valued or infinite-site regime

The justification given in the source is Mac Lane style coherence in the regime the paper actually works in. That is enough for the formal hinge. It is not presented as closure of every enriched generalization.

Three-Gate Burden Map

The formal architecture of `SCPI` can be read as a three-gate map:

1. `Topological obstruction`
   - can the local data globalize at all?
   - primary object: the `H^1` obstruction class
2. `Conservativity`
   - if it globalizes, does the extension preserve the intended base theory rather than sneaking in stronger consequences?
3. `Definability`
   - if it globalizes conservatively, can the resulting concept be expressed explicitly in the available vocabulary?

The gates are not rhetorical staging. They separate different kinds of failure that would otherwise be conflated as generic disagreement.

Selected Formal Artifact Inventory

The strongest formal objects directly visible in the repository are:

- `papers/scpi/lean/SCPI/Basic.lean`
- `papers/scpi/lean/SCPI/Torsor.lean`
- `papers/scpi/lean/SCPI/Conservativity.lean`
- `papers/scpi/lean/SCPI/Beth.lean`
- `papers/scpi/lean/SCPI/SchemaDiscovery.lean`

The worked paper example most useful for cross-reading the formal and empirical layers is:

- the Calendar/Email/Slack site in `SCPI`

That example matters because it is the nearest thing to a common hinge between the purely formal paper, the empirical bridge layer, and the distributed extension in `SHEAF`.

Appendix C: Empirical Method And Result Tables

This appendix gathers the compact empirical surfaces behind the bridge layer.

Topology-Indexed Status Surface

The compact Bridge surface should now be read topologically rather than only scenario-by-scenario.

Family	β_1	Blind spots	Prompt-hard.	Pred. repair	Minimality
Tree control	0	n/a	n/a	5/5 contractible	n/a
Single cycle (`v1`)	1	5/5	0/5	4/5	yes (`A05`)
Shared-edge double (`v2`)	2	5/5	0/5	5/5	yes (`MC04`)
Figure-eight (`v3`)	2	5/5	0/5	5/5	yes (`F04`)

- Tree control is a topology-changing control.
- Single cycle (`v1`) is structural-stable; repair is mixed on the behavioral tail.
- Shared-edge double (`v2`) is repair-stable; behavior-sensitive on decomposition.
- Figure-eight (`v3`) is structural-stable and repair-stable.

Repair And Minimality Summary

The strongest current Bridge claim is no longer only that bilateral-pass / cycle-fail cases exist. It is that multi-cycle repair now has two live `beta_1 = 2` families with selective closure, wrong-bridge failure, and a minimality witness inside the multi-cycle regime.

Family	Strict subset behavior	Full predicted bridge set	Wrong typed bridge
`benchmark_v1` / `A05`	one bridge leaves the predicted residual	historically strongest single-cycle witness; live execution still has a behavioral tail	not the current flagship control surface
`benchmark_v2` / `MC04`	one bridge leaves one residue, the other leaves the complementary residue, irrelevant bridge leaves both	closes	closes `0/5`; in one timing case it worsens the system by adding a second residue
`benchmark_v3` / `F04`	one bridge leaves one residue, the other leaves the complementary residue, irrelevant bridge leaves both	closes	closes `0/5`; same near-miss pattern as the shared-edge family

Hidden-Seam Repair Recovery

A separate study tested whether independent models recover the same minimal typed repair under hidden vocabulary conditions (`bridge/demos/hidden_repair/`).

Three providers were given failure descriptions using only plain domain language. No bridge, sheaf, or topological terminology was used. Each model independently proposed shared rules to resolve the observed global inconsistency.

Metric	Value
Exact type matches	14/15
Partial matches	1/15
Misses	0/15
Convergent scenarios	4/5

The partial match (MC04, claude-3.5-haiku) recovered one of two predicted generators, leaving the other as a residue consistent with the predicted partial-repair structure.

This does not claim bridge artifacts are uniquely discoverable. It shows the repair types predicted by the obstruction structure are independently recoverable from plain failure descriptions and that independent models converge on the same repair structure in the majority of cases.

Current Empirical Reading

The decisive Bridge surface is now:

- pairwise-valid blind spots persist across multiple topology families
- topology-preserving prompt hardening closes `0/5`
- the predicted typed repairs close `5/5` in both live `beta_1 = 2` families
- reassignment stability is strongest on the repair side, while the shared-edge decomposition baseline remains behavior-sensitive
- independent models recover the same minimal repair types from plain failure descriptions (14/15 exact, 4/5 convergent)

That is the compact empirical status the omnibus now needs to preserve.

Scaling Evidence: The Coherence Cliff

The `Bridge` experiments above operate at small scale (3–8 agents). The Coherence Cliff extends the empirical evidence to 50 agents across 500 composition graphs. Its central finding is a regime change: the best sheaf diagnostic (mean cycle frustration) maintains $R^2 > 0.96$ at all 7 tested scales ($n = 5$, 10, 15, 20, 30, 40, 50), while the strongest non-sheaf baseline—a Random Forest trained on all graph-topological features—degrades from $R^2 = 0.83$ at $n = 5$ to $R^2 = 0.52$ at $n = 50$. The predictive gap nearly triples.

The experiment uses a deterministic symbolic executor with zero model noise, convention heterogeneity grounded in real divergence patterns (ISDA, Basel, vendor calibration), and a stochastic block model for graph generation. Under equal-budget repair, H^1-prescribed repairs outperform spectral, cycle-breaking, bounded-depth, and random strategies.

The full paper, code, data, and figures are at `papers/coherence-cliff/`. The paper is included as a facsimile in Part IV of this omnibus.

Bronze+ Real-Protocol Evidence

The Bronze+ experiment extends the evidence beyond synthetic benchmarks to actual MCP servers. On 18 instances of mixed-provenance composition (3 custom + 1 official Memory server):

- Protocol surface: 44/44 checks pass
- Semantic failure: 30-minute escalation discrepancy despite all-green local validation
- Sheaf $R^2 = 0.940$ vs best conventional $R^2 = 0.610$
- structural_sheaf repair: +11.3% at K=1, +42.0% at K=5, +60.1% at K=8
- Structural advantage concentrated at K=1–3; methods converge at K=8

The effect survives contact with real protocol infrastructure and heterogeneous server provenance.

Interpretability Frontier Evidence

The Interpretability Frontier paper (Part VI) extends the empirical surface from external diagnostic baselines to model internals. Across 240+ compositions in three domains (invoice/settlement, calendar/escalation, policy/audit) at four scales (n = 5, 10, 15, 20), six interpretability baseline families were evaluated against the structural diagnostic.

Six key findings:

1. Perfect local knowledge, zero global signal. B2 probing classifiers achieve 99.8% mean accuracy at classifying conventions at every edge. This perfect edge-local recovery carries zero predictive value for composition-level failure (ϱ = NaN due to constant output).
2. Topology-dependent gap. On cyclic compositions at n = 5, the best interpretability baseline (B1 SAE divergence) achieves ϱ_cyclic = 0.685; the structural diagnostic achieves ϱ = 1.000.
3. The Interpretability Cliff. B1 SAE's cyclic correlation decays monotonically with scale: $0.685 \to 0.515 \to 0.321 \to -0.067$ from n = 5 to n = 20. The structural diagnostic remains at $\varrho = 1.000$ at all scales.
4. Cross-domain replication. The gap replicates across all three domains. The best interpretability baseline never exceeds ϱ_cyclic = 0.745 (B1 SAE, policy n = 10).
5. Graph-level aggregation does not close the gap. B6a (cycle-oracle aggregation with oracle knowledge of graph topology) matches B1 within rounding in 4 of 6 conditions. B6b (learned cycle-aware predictor) goes anti-correlated in 4 of 6 conditions. The gap is representational, not aggregational.
6. Cross-model replication. Gemma 2 2B (2.6B parameters, 26 layers, different architecture family) replicates the gap at both n = 5 (ϱ_cyclic = 0.515 vs GPT-2's 0.685) and n = 10 (ϱ_cyclic = 0.467 vs GPT-2's 0.515). The gap is larger on the bigger model. Probing accuracy remains perfect ($\geq$ 0.995). The structural diagnostic achieves $\varrho = 1.0$ at both scales.

The decisive result is B6a: even when the aggregator knows which edges form cycles, the interpretability features it aggregates do not contain the holonomy-relevant signal. The structural diagnostic wins because it observes cycle-relevant convention transformations directly.

Primary experiments use GPT-2 Small with TransformerLens activation extraction and SAELens sparse autoencoder features, chosen as the most thoroughly studied model in the mechanistic interpretability literature. Cross-model replication uses Gemma 2 2B with independently trained SAE features (gemma-scope-2b-pt-res-canonical), confirming the boundary is not architecture-specific.

Demo Inventory

The current `bridge/demos/` directory includes:

- `seam_experiment.py`
- `bridge_oaei.py`

- `diagnostic_test.py`
- `xbrl_scenarios.py`
- `xbrl_experiment.py`
- `xbrl_taxonomy.py`
- `xbrl_bilateral_mapping.py`
- `xbrl_coboundary.py`

These should be read as the empirical bench behind the paper rather than as a second prose surface.

Naming Discipline

In this omnibus, the empirical layer is referred to as `Bridge`.

- The included paper itself appears under the title The Coherence Fee.
- Nearby repo materials have also used The Bridge Problem and The Bridge Conjecture.

The purpose of the appendix is to keep the empirical layer legible without allowing naming drift to make the reader think three different objects are being discussed.

Bulla Trilogy Empirical Surface (703-Composition Real-Schema Corpus)

The Bulla Formal Trilogy (Part IV) is anchored in a real-schema corpus distinct from the Bridge demos. The following corpus-level identities have been verified across all 703 compositions:

Verification	Result	Where carried
Witness Gram rank identity (`rank K(G) = fee(G)`)	703/703	Paper III + WitnessGramRank.lean
Leverage prediction (`□_j = (|H_C|−1)/|H_C|`)	240/240	Paper III §4 + LaplacianCollapse.lean
Natural partial-CHP regime breaks observed	37 cases	Paper III §5
Multi-column block rank ≤ 2	3873/3873	Signed-Incidence + IncidenceLaplacian.lean
TU row structure (each row has ≤ one +1 and ≤ one −1)	1596/1596	Signed-Incidence + SignedIncidence.lean
Repair entropy basis count β = ∏|C_{d,j}|	240/240 (DFD+CHP regime)	Witness Geometry Beyond Fee + DFDBlockDiagonal.lean
Sharp bound ϕ+1 ≤ β ≤ 2^ϕ	240/240	Witness Geometry Beyond Fee §3
Endpoint coupling rank 2	240/240	Signed-Incidence Corollary

The corpus is drawn from a curated registry of MCP servers, not synthetic data. Each composition consists of a real schema pair; the structural claims hold uniformly across the corpus.

BABEL Completeness Surface

BABEL v0.1 contains 932 instances. The split structure is:

Split	Instances	Purpose
Dev	514	Public, baseline tuning
Test (visible)	209	Public, leaderboard evaluation
Test (hidden)	209	Maintainer-only, integrity check

Family	Provenance	Instances
`synthetic_scaling`	Synthetic	168
`invoice`	MCP-shaped facsimile	96
`calendar`	MCP-shaped facsimile	96
`policy`	MCP-shaped facsimile	96
`real_mcp_calendar`	Real MCP servers	192
`real_mcp_invoice`	Real MCP servers	192
`external_apis`	API-derived (Stripe/Twilio/SendGrid/GitHub/Shopify/etc.)	92

Five frontier LLMs (GPT-4o, Claude 3.5 Sonnet, Claude Sonnet 4, Claude Opus 4, Gemini 1.5 Pro) tested on the dev split: ϱ near zero on the failure-prediction track. Oracle CoT decomposition (matrix-aware prompts on all five): Claude Sonnet 4 reaches $\varrho = 0.80$ with oracle restriction matrices but R^2 = -4.0 (rank-correct, magnitude-wrong); Opus 4 shows the largest information-extraction delta ($\Delta\varrho$ = +0.70 from base prompt to oracle-CoT). The bottleneck is arithmetic reasoning rather than structural extraction.

Compositional Hiddenness Empirical Surface

The hiddenness study (Part VI) provides a complementary empirical surface separate from BABEL: a synthetic ecology with known ground truth on hidden conventions, and a context-ablation testing representational sensitivity.

Condition	Accuracy on hiddenness recovery
Structured relational + convention-specific task language	88%
Structured relational + neutral language	55% ($\Delta = -33$pp)
Flat representation + convention-specific language	55% ($\Delta = -33$pp)
Flat + neutral	46% (chance $\approx$ 50%)

Cross-model: Claude Sonnet 4 and GPT-4o both exhibit the relational-framing effect; failure modes differ. Bulla recovers the non-local predicate exactly across all conditions.

Appendix D: Protocol Artifacts

This appendix gathers the most useful implementation-facing surfaces from the protocol layer.

seam-lint

`seam-lint` is the current diagnostic tool for agent tool compositions.

Its governing distinction is:

- `Observable (F_obs)`: only what the tools declare in their schemas
- `Full (S)`: the full internal state the tools rely on

The coherence fee is computed as:

```
dim H^1(F_obs) − dim H^1(F_full)
```

The value of the tool is not merely that it reports a number. It identifies blind-spot dimensions and recommends bridge annotations that reduce the fee to zero.

Included Composition Surface

The current composition set includes:

File	Domain	Tools	Fee	Blind spots
`auth_pipeline.yaml`	Auth / audit	3	0	none
`financial_pipeline.yaml`	Finance	3	2	`day_convention`, `risk_metric`
`rag_pipeline.yaml`	RAG	3	3	`chunk_size`, `citation_mode`, `relevance_threshold`
`code_review_pipeline.yaml`	DevOps	4	3	`diff_format`, `severity_threshold`, `style_convention`

File	Domain	Tools	Fee	Blind spots
`data_etl_pipeline.yaml`	Data engineering	4	4	`timezone`, `null_convention`, `decimal_precision`

Additional composition files currently present:

- `web_research_pipeline.yaml`
- `mcp_filesystem_git.yaml`
- `mcp_fetch_memory.yaml`
- `mcp_fetch_filesystem_git.yaml`

Artifact Inventory

The implementation-facing surfaces presently visible in the repository are:

- `papers/seam/paper/seam.tex`
- `papers/seam/example/seam_example.py`
- `papers/seam/seam-lint/seam_lint.py`
- `papers/seam/seam-lint/compositions/*.yaml`
- `papers/seam/study/procurement_tournament/`
- `papers/seam/study/procurement_tournament/finance_rfp/`

The point of this inventory is not to print raw code. It is to make clear that the protocol layer already has executable and inspectable artifact surfaces, rather than existing only as whitepaper prose.

Procurement Consequence Surface

The protocol layer now also has a bounded procurement consequence study rather than only a diagnostic inventory.

The strongest current artifact is the finance-family procurement brief in `papers/seam/study/procurement_tournament` This is now an externally anchored consequence artifact. Its burden is deliberately narrow:

- all candidates are functionally matched
- all candidates pass local validation
- cost, latency, and fee are visible before selection
- holdout seam outcomes are revealed only after selection

The holdout regime includes six externally anchored cases across four consequence types:

- three committee-correction bulletins (internal governance, severity high to critical), anchored to ISDA EMU Memo (1998), Basel RCAP (2013), Brattle/ARRC (2021), and Basel FRTB (2016/2019)
- one regulatory-restatement notice (external regulatory, severity critical), anchored to ISDA 2021 Definitions overhaul
- one enforcement action (external enforcement, severity critical), anchored to the JPMorgan London Whale VaR model switch (US Senate PSI, 2013)
- one convention-insensitive null case (severity none, zero burden)

All non-null cases are grounded in independently verifiable standards-body, regulatory, or enforcement findings.

Hours are reported as `review / rework / fail-sign-off`.

Policy	Candidate	Fee	Hours	Burden	Regret
local_baseline	vendor_a	2	23 / 30 / 7	57.60	57.23
fee_threshold	vendor_b	1	14 / 18 / 4	34.20	33.96
naive_checklist	vendor_d	2	23 / 30 / 7	57.60	57.30
post_reveal_oracle	vendor_c	0	0 / 0 / 0	0.00	0.00

Among functionally matched, locally valid candidates in this finance-family brief, fee-aware qualification reduces downstream regret from 57.23 to 33.96 under hidden post-selection cases driven by `day_convention` and `risk_metric`. It outperforms both the local baseline and a naive convention-checklist rival (regret 57.30) under an externally anchored holdout regime spanning committee corrections, regulatory restatement, enforcement action, and a convention-insensitive null.

The artifact is externally anchored but still synthetic in construction: candidates and policies are designed, not observed from real procurement decisions. No broader procurement claim is warranted from this artifact alone.

Appendix E: Frontier Notes

This appendix records the most important frontier-facing surfaces behind `SHEAF`.

Current Frontier Status

From the current repository state:

Component	Status	Note
Paper outline	Draft	Major structure exists; later sections remain more open than the settled core papers.
Nerve computation	Implemented	`simulations/nerve.py`
`H^1` computation	Implemented	`simulations/cohomology.py`
Worked example	Implemented	`simulations/examples/calendar_email_slack.`
Sheaf Laplacian	Placeholder	explicit open perimeter
Topology auction	Placeholder	explicit open perimeter
Formal proofs	Not started at full program scale	later phase

Proof Note Surface

The current primary harvest route is the communication bottleneck theorem program.

Its paper-shaped boundary is now defined in:

- `papers/sheaf/COMMUNICATION-BOTTLENECK-HARVEST.md`

The proof note presently most legible as a stand-alone frontier appendix inside that route is:

- `papers/sheaf/proofs/haar-universality-lemma.tex`

Its status line:

- computationally verified
- Part (a) (Haar marginals via bi-invariance) proved

- Part (b) (pairwise independence of adjacent edges) proved — exact, TV = 0
- Part (c) (spectral gap via triangle holonomy) assembled from two derived lemmas (Grassmannian concentration citing Franke-Kabluchko-Prochno 2023, and triangle holonomy bound) plus direct citation of Bandeira-Singer-Spielman (2013) Theorem 2.6
- remaining risk ~3%, in the holonomy lower-bound constant

A self-contained paper draft now exists at `papers/sheaf/paper/communication-bottleneck-paper.tex`. The autonomy test is passed: the paper motivates the result from communication complexity and spectral graph theory without reference to the broader program. This paper is now included in the omnibus as a facsimile within Part IV, immediately following the SHEAF paper.

The current frontier boundary should therefore be read in six layers:

- the communication bottleneck theorem is now a paper-shaped second technical object with a self-contained draft
- the Haar-universality lemma is fully proved (all three parts) with ~3% residual risk in one constant
- The Coherence Cliff is a standalone scaling experiment (500 graphs, 7 scales, 5–50 agents) providing quantitative evidence that sheaf diagnostics are necessary at scale; its code, data, and figures live at `papers/coherence-cliff/` and the paper is included as a facsimile in Part IV
- BABEL (formerly COHERENCE-GYM) is the benchmark operationalization with 7 families, 932 instances, 3 tracks, and a frozen evaluation protocol
- The Interpretability Frontier is the representational boundary paper (Part VI): 240+ compositions, 6 baseline families, 3 domains, 4 scales, 2 model architectures. The decisive B6a result closes the aggregation objection. Cross-model replication on Gemma 2 2B confirms the gap widens at 20× scale. Replication on additional architectures (Llama 3, 70B-class models) is the priority next step
- the full enriched `H^1` and Laplacian-Cohomology Bridge agenda remains deferred frontier

Simulation Example Inventory

The current example-facing simulation set includes:

- `calendar_email_slack.py`
- `failure_benchmark.py`
- `multi_agent_llm.py`
- `protocol_trace.py`
- `uuv_coordination.py`

These examples show the breadth of the frontier surface: diagnostic carryover from the settled examples, protocol traces, irreducible failure benchmarks, and heterogeneous-agent coordination scenarios.

Frontier Usage Rule

The role of this appendix is to keep the frontier visible without collapsing it into either of two errors:

- treating it as already closed to the same degree as the Composition Doctrine, `SCPI`, `Bridge`, and `Seam`
- hiding it so completely that the visible perimeter of the program disappears

The correct reading posture is stronger than "mere speculation" but weaker than "settled center." That asymmetry is a quality signal, and the omnibus should preserve it.

Open Conjectures From the Composition Doctrine

The doctrine paper (Part 0) names five open conjectures that bound the frontier of the verified result. Each is recorded with the program's standard claim discipline: explicit falsification condition, expected difficulty, and program adjacency.

Conjecture 9.1 — Surrogate-regime characterization. A parallel uniqueness theorem on `SemComp_sur` (the class of complexes whose latent cochain complex is not rationally split) with an explicit correction term `κ(G) ≥ 0`. Status: open; multi-quarter program; partial results in `witness-geometry-beyond-fee` (leverage-conservation) and `hierarchical-fee` (additive decomposition).

Conjecture 9.2 — Verifier universality. Any sound verifier of compositional coherence factors through an equivalent of the witness functor (Yoneda-flavored). Status: open; substantially harder than the characterization theorem; would require representation-theoretic argument on the verification category.

Conjecture 9.3 — Semantic FLP-style impossibility. A distributed protocol exchanging messages between heterogeneous components cannot certify semantic agreement under asynchrony and bounded fault tolerance, with impossibility parametrized by witness rank. Status: open; the Communication Lower Bound (Theorem 8.1, doctrine paper) is a clean precursor but does not engage termination/liveness.

Conjecture 9.4 — Eval-correctness inseparability. Every locally-computable evaluation metric is asymptotically uncorrelated with the witness rank on a parameterized family of compositions. Status: open with empirical support (Coherence Cliff $R^2 > 0.96$ vs baselines degrading $0.83 \to 0.52$ across scales 5–50). Most empirically grounded of the four, theoretically tractable; closest to a clean theorem.

Conjecture 9.5 — Coherence-preserving update. An interface update preserving the chain-homotopy class of the seam complex preserves witness rank, obstruction class, and disclosure-normal-form receipt validity. Status: open; most product-adjacent of the conjectures; a proof would license incremental recertification under continuous interface evolution.

The doctrine paper does not require any of these to be true. Its main result — Theorem 5.1 — stands on its own in the exact regime. The conjectures describe the natural directions in which the doctrine generalizes.