∴

Technical Spine

Omnibus

∴

Technical Spine

Omnibus

Formal Hinge, Empirical Witness, Protocol Consequence,
Frontier Extension, Benchmark Evidence, and Interpretability Boundary

With Formal, Empirical, and Implementation Appendices

John Komkov

Technical Spine Omnibus

Collected technical companion

This volume gathers the technical center of the *Res Agentica* program: the formal hinge (*SCPI*), the empirical witness (*Bridge*), the protocol consequence (*Seam*), the frontier extension (*SHEAF*, *Communication Bottleneck*, *Coherence Cliff*), the benchmark evidence (*BABEL*, *Bronze+*, *Silver*), the interpretability boundary (*Edge-Local Interpretability Is Not Enough for Cyclic Composition*), and curated appendices on proofs, experiments, protocol artifacts, and frontier notes.

First edition: 2026-03-17

Contents

Part I: Formal Hinge

Part II: Empirical Witness

Part III: Protocol Consequence

Part IV: Frontier Extension

Part V: Benchmark and Real-Protocol Evidence

Part VI: Interpretability Boundary

Part VII: Formal and Implementation Appendices

Editorial Note

Title	*Res Agentica: Technical Spine Omnibus*
Edition	March 2026 (post-Interpretability Frontier)
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	2026-03-16
Canonical citation	*Res Agentica: Technical Spine Omnibus*, March 2026 ed.

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

The eight papers collected here develop a single argument across seven registers: 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 resolution has a computable cost (economics); that the problem grows worse, not better, at scale (a regime change); that all of this can be measured by a public benchmark anyone can run (an instrument); 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).

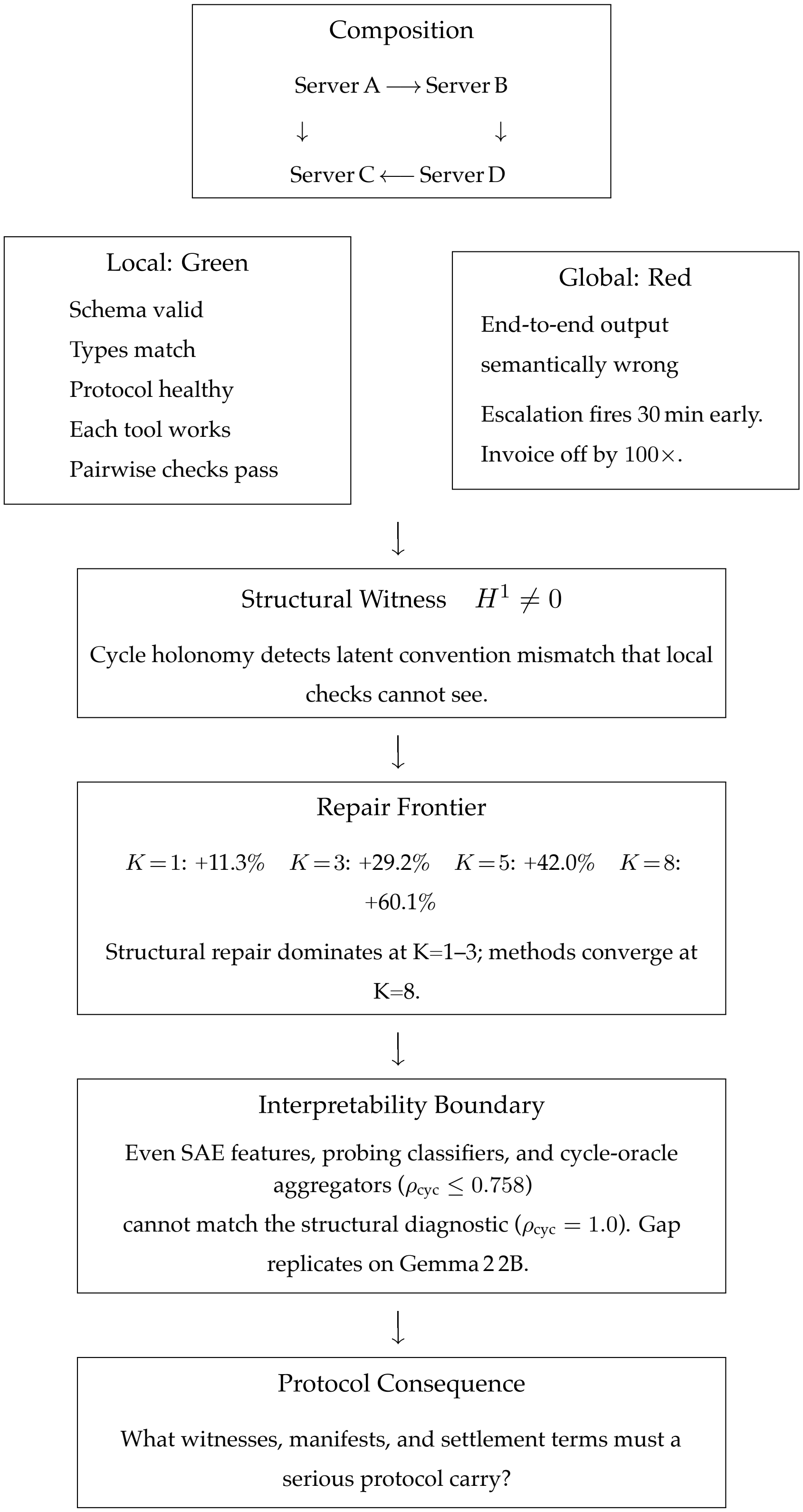
The escalation—from impossibility to detectability to accountability to cost to necessity to measurement to representational boundary—is the intellectual spine. The constitutional question that motivates *Res Agentica* is what happens when this same failure mode operates between autonomous agent runtimes 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. `SCPI` is the formal hinge. `Bridge` is the empirical witness. `Seam` is the protocol consequence. `SHEAF` is the frontier extension that becomes visible once the settled center is granted. `BABEL` (formerly COHERENCE-GYM) is the benchmark operationalization that turns the central claim into a public instrument. `Bronze+` and `Silver` are the real-protocol demonstrations that test the claim against actual MCP servers of mixed provenance across two domains—calendar/escalation and invoice/settlement—with the Silver track including two non-house servers. The *Interpretability Frontier* paper is the capstone: it tests whether mechanistic interpretability tools—SAE features, probing classifiers, attention diagnostics, CKA, and learned graph-level aggregators—can substitute for structural diagnosis, and shows they cannot, across two model architectures (GPT-2 Small and Gemma 2 2B), because the missing signal is representational rather than aggregational.

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 settled, what is empirical, what is operational, what is benchmarked, and what remains open.

Canonical Names

Name	Role	Part
SCPI	Formal hinge: the impossibility theorem	I
Bridge / *The Coherence Fee*	Empirical witness: the obstruction appears in concrete systems	II
Seam	Protocol consequence: what a serious response requires	III
SHEAF	Frontier extension: enriched cohomology and scaling theory	IV
BABEL	Public benchmark (v0.1): 932 instances, 7 families, 3 tracks, oracle CoT	V
Bronze+	Real-MCP calendar note: 3 custom + 1 official server	V
Silver	Real-MCP invoice note: 2 custom + 2 non-house servers	V
Interpretability Frontier	Representational boundary: edge-local interpretability is not enough	VI



Composition

Server A ⟶ Server B

↓ ↓

Server C ⟵ Server D

Local: Green

Schema valid

Types match

Protocol healthy

Each tool works

Pairwise checks pass

Global: Red

End-to-end output semantically wrong

Escalation fires 30 min early.

Invoice off by 100×.

Structural Witness $H^1 \neq 0$

Cycle holonomy detects latent convention mismatch that local checks cannot see.

Repair Frontier

$K = 1$: +11.3% $K = 3$: +29.2% $K = 5$: +42.0% $K = 8$: +60.1%

Structural repair dominates at K=1–3; methods converge at K=8.

Interpretability Boundary

Even SAE features, probing classifiers, and cycle-oracle aggregators ($\rho_{\text{cyc}} \leq 0.758$)

cannot match the structural diagnostic ($\rho_{\text{cyc}} = 1.0$). Gap replicates on Gemma 2 2B.

Protocol Consequence

What witnesses, manifests, and settlement terms must a serious protocol carry?

Local validity does not guarantee global coherence. 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.

Claim	Status	Evidence Level	Where
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, V
MCP-shaped workflows exhibit protocol-green / semantics-red	Demonstrated	Bronze+ (calendar, 1 official server, 44/44 checks pass); Silver (invoice, 2 non-house servers, all checks pass)	Part V
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, IV, 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: what witnesses, manifests, and settlement consequences a serious protocol response requires	Part III
Benchmark released	Complete	BABEL v0.1: 932 instances, 7 families, 3 tracks, 10 baselines, frozen evaluation protocol, replication pack	Part V
LLMs fail on compositional diagnosis	Demonstrated	5 models, 3 providers; oracle CoT decomposition isolates arithmetic bottleneck ($\varrho = 0.80$ ranking but $R^2 = -4.0$ magnitude)	Part V
Holonomy predicts real dollar error	Demonstrated	Live-pipeline validation: 27 runs across 9 convention-pair configs, $\varrho = 0.795$ ($p < 7.4 \times 10^{-7}$); non-circular gold standard	Part V
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 $\varrho = 1.0$ in every condition	Part VI

Claim	Status	Evidence Level	Where
External replication	Pending	Replication pack published; no outside confirmation yet; interpretability paper adds an independent empirical surface	—
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 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 ($\varrho = 0.80$) 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 (e.g., graph neural networks over interpretability features, or novel graph-level representation classes) 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 argues that composition protocols should carry semantic witnesses and settlement terms. 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 ≈ 1.0×), 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 **architectural**: the logical structure of the argument.

1. `SCPI` for the formal hinge: bilateral checks cannot see cycles.
2. `Bridge` for the empirical witness: LLMs cannot see them either.
3. `Seam` for the protocol consequence: and you can prove they didn't.
4. `SHEAF` for the frontier extension: and here is what resolution costs.
5. The *Coherence Cliff* and *Communication Bottleneck*: and the problem grows worse at scale.
6. `BABEL`, `Bronze+`, and `Silver`: and here is how to measure all of it.
7. The *Interpretability Frontier*: and even the best per-component tools cannot substitute for structural diagnosis.

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

1. `BABEL` (Part V) for the benchmark: 932 instances, 7 families, 3 tracks. Five frontier LLMs fail. The oracle CoT experiment decomposes why.
2. The *Coherence Cliff* (Part IV) for the scaling evidence: the regime change where bounded-depth testing collapses.
3. The *Interpretability Frontier* (Part VI) for the representational boundary: mechanistic interpretability at its best still cannot diagnose cyclic compositional failure.
4. `SCPI` (Part I) for the formal foundation: the obstruction is topological and the proof is machine-checked.
5. `Seam` (Part III) for the protocol response: what to build once the obstruction is granted.

The third is **evidentiary**: inspect the primary objects directly.

1. Read the editorial and part notes first.
2. Read the included papers as facsimiles of the primary technical objects.
3. Use the appendices to inspect proof status, experimental design, protocol artifacts, and frontier notes without losing the hierarchy of the main volume.

Two distinctions apply throughout:

- Settled core vs `frontier`.
- Paper material vs appendix burden.

The first three parts constitute the hard center of the present program. Part IV is included because it is the visible perimeter of the same argument, not because it is already closed to the same degree. Within Part IV, the *Linear Communication Bottleneck Theorem* appears as a separate facsimile after the SHEAF paper—the first result harvested from the frontier that has passed an autonomy test and can be read independently of the surrounding program. *The Coherence Cliff* follows as a scaling experiment that provides the program's current large-scale empirical evidence for the necessity of sheaf-cohomological diagnostics: a regime change where the predictive gap between the best sheaf diagnostic and the best conventional baseline nearly triples from 5 to 50 agents.

Part V 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. `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 — Claude Sonnet 4 achieves $\varrho = 0.80$ with oracle matrices but $R^2 = -4.0$, and Opus 4 shows the largest information-extraction delta ($\Delta\varrho = +0.70$). The live-pipeline validation (BABEL Section 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 VI 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 $\varrho = 0.758$. The decisive result: 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 formal theory predicts local blindness, the benchmark measures it, and now the interpretability paper shows that even the richest local tools inherit it — 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 interfaces 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.

Part I

Formal Hinge

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 $M: C^{\mathrm{op}} \to \mathbf{Grpd}$, 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 ∼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

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

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 $M\colon C^{\mathrm{op}} \to$ **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 $(C^{\mathrm{op}}, J_{\mathrm{at}})$ 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., $\mathbf{Set}^{C^{\mathrm{op}}}$)	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*

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

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.

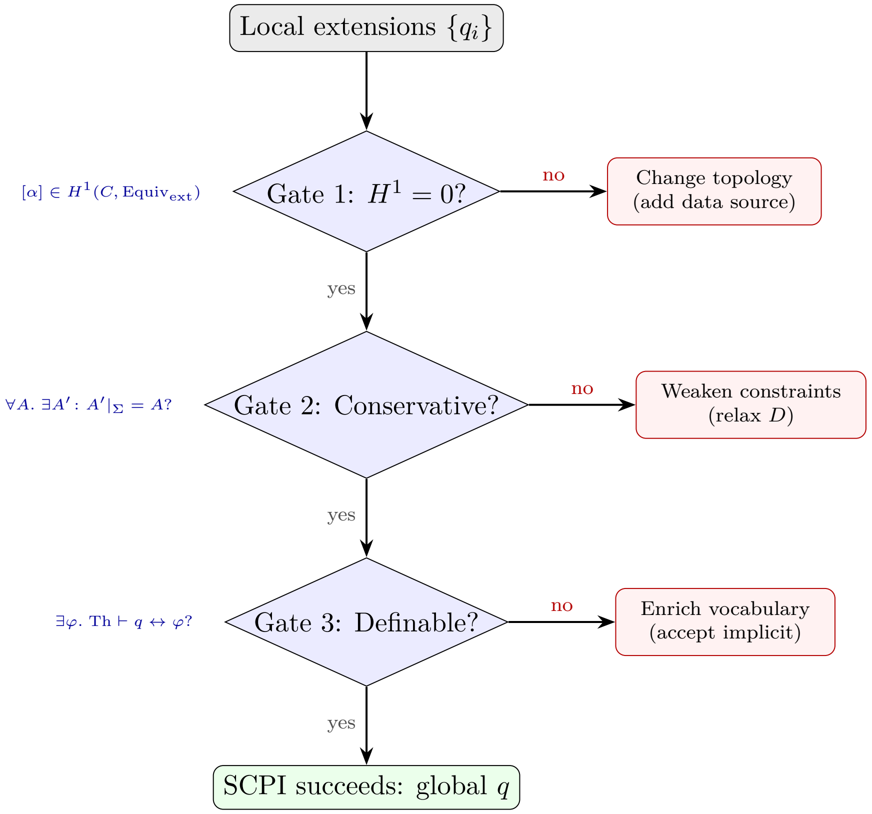

Local extensions $\{q_i\}$
$[\alpha] \in H^1(C, \mathrm{Equiv}_{\mathrm{ext}})$
Gate 1: $H^1 = 0$?
no
Change topology (add data source)
yes
$\forall A.\ \exists A' : A'|_\Sigma = A$?
Gate 2: Conservative?
no
Weaken constraints (relax D)
yes
$\exists \varphi.\ \mathrm{Th} \vdash q \leftrightarrow \varphi$?
Gate 3: Definable?
no
Enrich vocabulary (accept implicit)
yes
SCPI succeeds: global q

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 $(\{1, \dots, 2m+1\}, \leq)$ 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: $\exists$ 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. *$\mathrm{Equiv}_{\mathrm{ext1}} \cong \mathrm{Equiv}_{\mathrm{ext2}}$ as sheaves of groups on the respective nerves (compatible with the isomorphism in (1)).*

Then $H^1(\mathcal{U}_1, \mathrm{Equiv}_{\mathrm{ext1}}) \cong H^1(\mathcal{U}_2, \mathrm{Equiv}_{\mathrm{ext2}})$ 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 $\mathrm{Equiv}_{\mathrm{ext}} \cong \mathbb{Z}/2\mathbb{Z}$ (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 ∀-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 $U = \{u_1, \dots, u_n\}$ and sets $S_1, \dots, S_m$ 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:*

$$\begin{tikzcd}
\mathrm{Inv}(C,J,M) \arrow[r, "\mathrm{Inv}(F)"] \arrow[d, "\simeq"'] & \mathrm{Inv}(D,J',M') \arrow[d, "\simeq"] \\
C \arrow[r, "F"'] & D
\end{tikzcd}$$

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.

Part II

Empirical Witness

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).

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.

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.

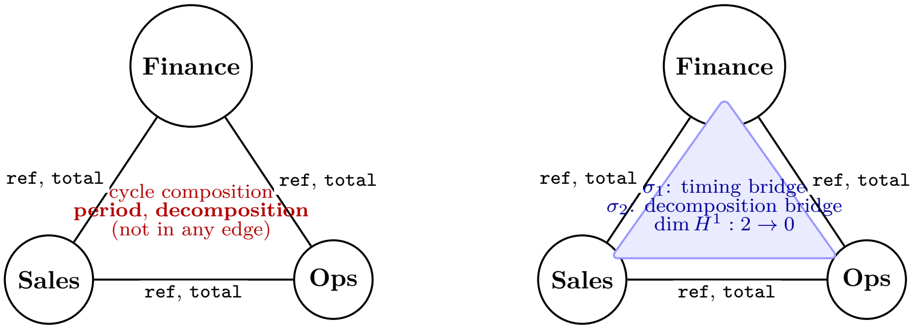


Figure 1: **Left:** the coordination graph with edge-local bilateral checks (shared fields) and the cycle-composition dimensions (private fields) that no edge observes. $\dim H^1 = 2$. **Right:** two bridge concepts (2-cells σ_1, σ_2) fill the triangle, killing both H^1 generators. $\dim H^1 = 0$.

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 $\delta_\sigma^1 : \mathbb{R}^6 \to \mathbb{R}^1$ 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: $Q1 \neq Q2$ 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: $1 \text{ item} \neq 2 \text{ entries}$.

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 handcrafted 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 ($\mathrm{rank}\,\mathcal{F}(\sigma) > 1$*),* $\dim_R H^1$ *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 $\dim_R H^1 = \sum_{e \in E} \mathrm{rank}_R\, \mathcal{F}(e) - \mathrm{rank}(\delta^0)$.

(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

$\mathrm{im}(\delta^1)$ 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 $\mathrm{im}(\delta^0)$ 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 $\mathrm{im}(\delta^0)$ (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.

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].

Part III

Protocol Consequence

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).

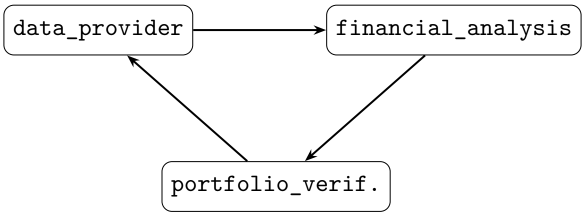
2 The Model

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. $\text{rank}(\delta^0_{\text{obs}}) = 6$, $\dim H^1(\mathcal{F}_{\text{obs}}) = 8 - 6 = 2$.
$\text{rank}(\delta^0_{\text{full}}) = 8$, $\dim H^1(\mathcal{F}_{\text{full}}) = 8 - 8 = 0$.

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 $H^1(\mathcal{F}_{\mathrm{obs}})$, 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" ≠ "trading" at `dp`→`fa`, and `risk_metric` = "sortino" ≠ "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.

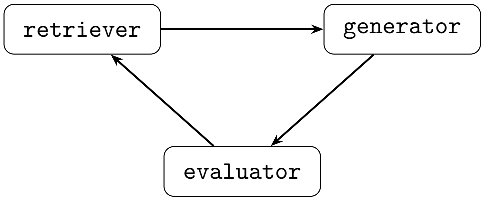
On a cycle, attribution fails. The inconsistency circulates: `dp` assumes calendar days; `fa` switches to trading days; `pv` interprets the result as if it were Sharpe when `fa` computed Sortino; `pv`'s feedback reaches `dp`, closing the loop. No single edge is "wrong." The composition is wrong.

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 ↔ 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 → Filesystem → 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 → Filesystem → 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 ↔ 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`.

Part IV
Frontier Extension

Frontier Extension

The final part of this volume should be read under a different sign from what precedes it.

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 the 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.

Two 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.

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.

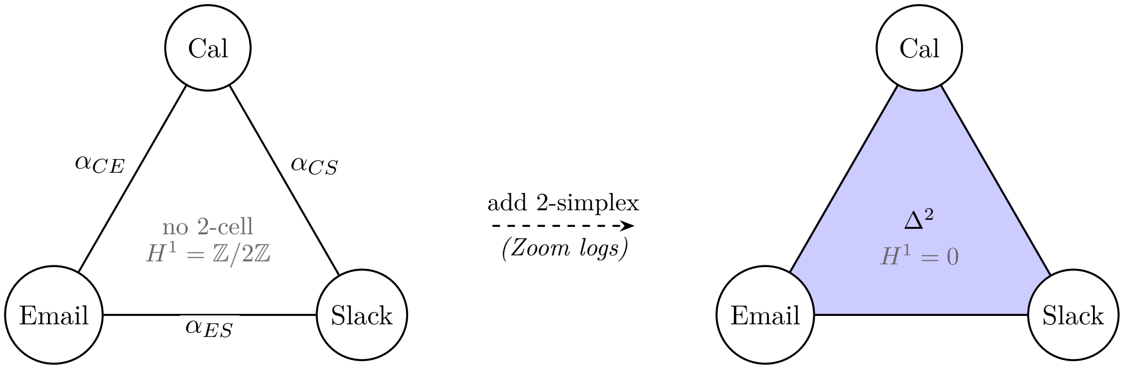


Figure 1: **Left:** three agents with pairwise overlaps but no triple overlap. The nerve is $\partial\Delta^2 \cong S^1$ (a 1-cycle with no filling 2-cell). **Right:** adding a shared data source (Zoom logs) creates a triple overlap, filling the triangle with a 2-simplex Δ^2. The nerve becomes contractible 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 $H^1(N, \mathrm{Equiv})$, there exist local adjustments $\{\beta_i\}$ making all agents' definitions globally consistent. A global consensus is structurally achievable.*
2. *If $[\alpha] \neq 0$ in $H^1(N, \mathrm{Equiv})$,* 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:

$$\begin{aligned}
&\alpha_{CE} \in \mathbb{Z}/2\mathbb{Z} : \text{do C and E agree on “is\_meeting” for their shared records?}\\
&\alpha_{ES} \in \mathbb{Z}/2\mathbb{Z} : \text{do E and S agree?}\\
&\alpha_{CS} \in \mathbb{Z}/2\mathbb{Z} : \text{do C and S agree?}
\end{aligned}$$

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 $H^1(N, \mathrm{Equiv})$—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 $H^1(N, \mathrm{Equiv})$. 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:

$$\begin{aligned} x_A &\leftarrow \min(x_B + w_{AB},\ x_C - w_{CA}) \\ x_B &\leftarrow \min(x_A - w_{AB},\ x_C + w_{BC}) \\ x_C &\leftarrow \min(x_B - w_{BC},\ x_A + w_{CA}) \end{aligned}$$

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."

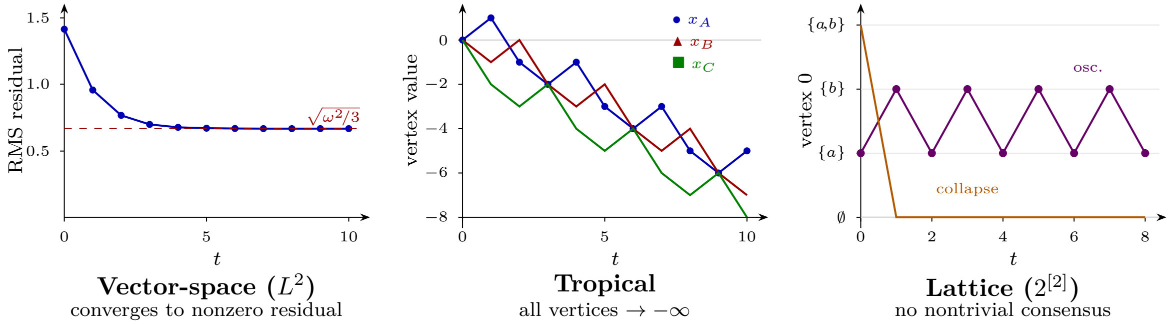
Consequence for the protocol: in the tropical regime, the Laplacian does not provide *graceful degradation* when $H^1 \neq 0$—there is no finite "best approximation" to return. The diagnostic detects the obstruction, but the resolution phase has no approximate output. Approximate consensus when $H^1 \neq 0$ 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.



Vector-space (L^2)
converges to nonzero residual

Tropical
all vertices → $-\infty$

Lattice ($2^{[2]}$)
no nontrivial consensus

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 $(\mathrm{swap}, \mathrm{swap}, \mathrm{swap})$; 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.

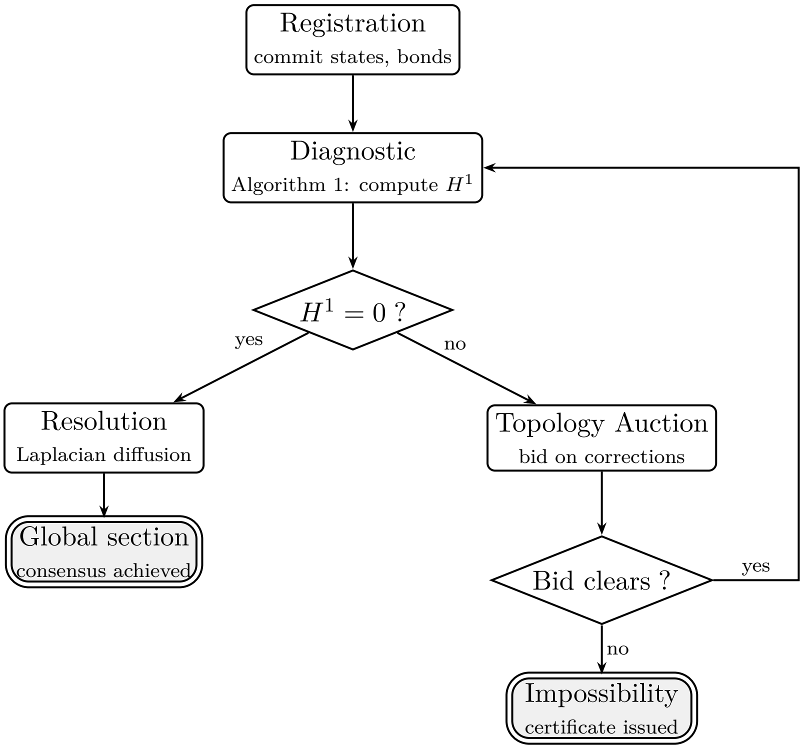


Figure 3: The SHEAF protocol loop. After registration, the H^1 diagnostic determines the protocol path. If $H^1 = 0$, the Laplacian resolves to a global section. If $H^1 \neq 0$, the topology auction prices corrections; a successful correction re-triggers the diagnostic. If no economically rational correction 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 $T \subseteq E$
Set $\beta_{\text{root}} = 0$
for each vertex $i$ in BFS order from root along $T$ do
    $\beta_i \leftarrow \beta_{\text{parent}(i)} + \alpha_{\text{parent}(i),i}$        ▷ Propagate along tree
end for
for each non-tree edge $(i, j) \in E \setminus T$ do
    $r_{ij} \leftarrow -\beta_i + \alpha_{ij} + \beta_j$        ▷ Residual on back-edge
    if $r_{ij} \neq 0$ then
        return NONTRIVIAL with certificate: cycle through $T$ from $i$ to $j$ plus edge $(i, j)$
    end if
end for
return TRIVIAL with witness $\{\beta_i\}$
```

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: $H^1(G, \mathrm{Equiv}) \cong \mathrm{Equiv}^{\beta_1}$ 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 → fuzzy-ϵ → 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 $f_0(S) = \operatorname{rank}(\delta^0_{G+S}) - \operatorname{rank}(\delta^0_G)$ 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 $\alpha_{ij} = \rho^{-1}_{j\to ij} \circ \rho_{i\to ij}$ 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 → lock → 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 Algorithm 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, $\star$ = 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	$\star$

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 $H^1(N, \mathrm{Equiv})$ 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

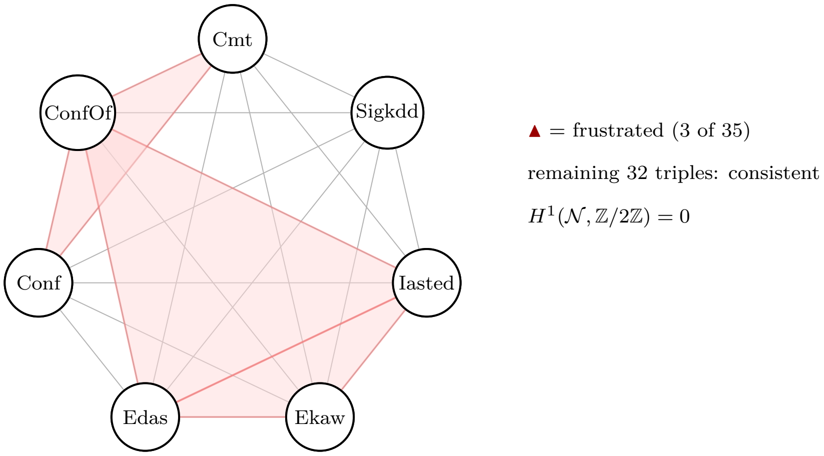

Cmt
ConfOf
Sigkdd
Conf
Iasted
Edas
Ekaw
▲ = frustrated (3 of 35)
remaining 32 triples: consistent
$H^1(\mathcal{N}, \mathbb{Z}/2\mathbb{Z}) = 0$

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 → Diagnostic → Auction → Verification → 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* $\perp$ *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, $\operatorname{meet}(x, \sigma(x)) = \perp$ 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: $R_{ij} = \arg\min_{R\in O(k)} \|X_iR - X_j\|_F$ 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 communication 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 → 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 integration 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 $\eta_G = \min_{g_1,\dots,g_n\in O(B)} \frac{1}{2|E|}\sum_{(i,j)\in E} w_{ij}\|g_iR_{ij} - g_j\|_F^2$. It measures how far the cocycle is from being a coboundary.*

By the Cheeger inequality for connection Laplacians [1]:

$$\frac{\lambda_1(L_1)}{2} \leq \eta_G \leq \sqrt{2\lambda_1(L_1)}.$$

3 Main Result

Theorem 5 (Linear Communication Bottleneck)**.** *Let $n \geq 3$ agents have d-dimensional representations with gauge transforms of angular magnitude θ, communicating through B-dimensional linear channels with $d \geq 2B$.*

(a) ***Aligned encoders, small gauge.*** *If the encoders share a common B-dimensional subspace up to gauge transforms of magnitude θ, then $\lambda_{B-1}(L_{\mathrm{comm}}) \leq C(B/d)\cdot\theta^2$.*

(b) ***Misaligned encoders.*** *If $E_1,\dots,E_n$ 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 ($0 \leq k \leq B$), 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],

$$\Pr\big[|f_x(E) - B/d| > t\big] \leq 2\exp\big(-c_0 \min(dt^2, d^{1/2}t)\big).$$

Setting $t = C_0\sqrt{(\log B)/d}$ and taking a union bound over a (1/4)-net of S^{B-1} with $|\mathcal{N}| \leq 9^B$, then lifting to operator norm via $\|M\| \leq 2\sup_{x\in\mathcal{N}} |x^\top Mx|$, 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 $d \geq 2B$. The Procrustes holonomy $H_{ijk} = R_{ij}R_{jk}R_{ki}$ satisfies*

$$c_1 \cdot \frac{B}{d} \leq \mathbb{E}\left[\frac{\|H_{ijk} - I_B\|_F^2}{2B}\right] \leq c_2 \cdot \frac{B}{d}$$

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 $\mathbb{E}[\|H_{ijk} - I_B\|_F^2/(2B)] \leq c_2B/d$.

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 $B \geq 2$. 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: $B - \mathbb{E}[\mathrm{tr}(H_{ijk})] \geq c_1'B^2/d$, giving $\mathbb{E}[\|H_{ijk} - I_B\|_F^2/(2B)] \geq c_1B/d$.

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 $c_1B/d \leq \eta \leq c_2B/d$.

By the Cheeger inequality for connection Laplacians [1], $\lambda_1(L_1)/2 \leq \eta \leq \sqrt{2\lambda_1(L_1)}$. 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

$$|\mathbb{E}[\lambda_{B-1}(L_{\mathrm{Proc}})] - \mathrm{Haar}(B, n)| \leq C' \cdot B/d.$$ □

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. 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.

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.

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.

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

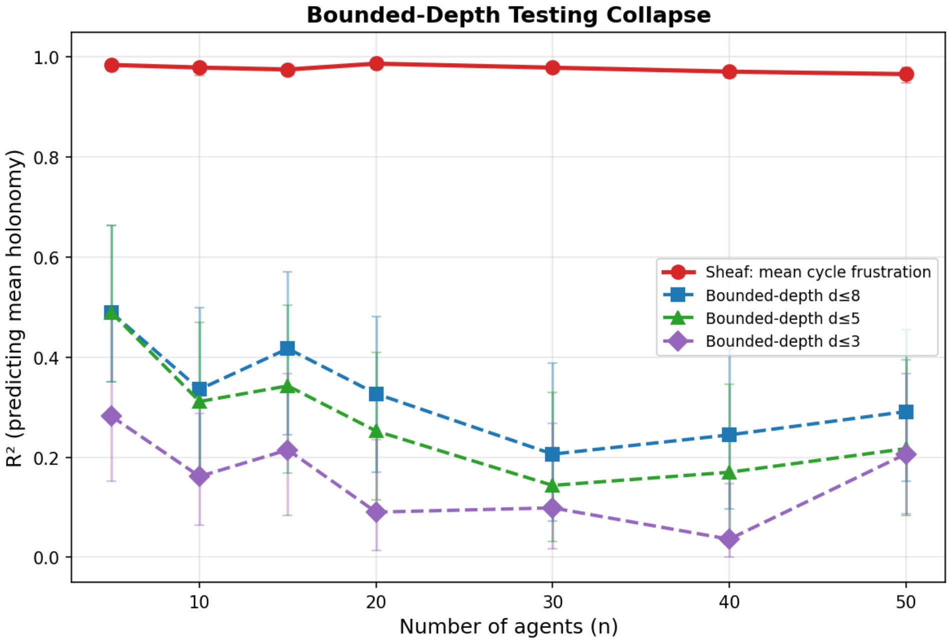


Figure 1: **Bounded-depth testing collapse (the title-earning figure).** R^2 for predicting mean holonomy vs. agent count. Bounded-depth baselines ($d \leq 3, 5, 8$) collapse with scale while mean cycle frustration (sheaf diagnostic) holds steady above $R^2 = 0.96$. Error bars: 95% bootstrap CI (1000 graph-level resamples; $n = 50$–75 graphs per scale).

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 the larger scales. This matters because it means the sheaf diagnostic does not merely predict failure better—it identifies *which edges to fix*.

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).

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` $\leftrightarrow$ `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	$[0.01, 0.25]$
$\dim H^{1\dagger}$	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

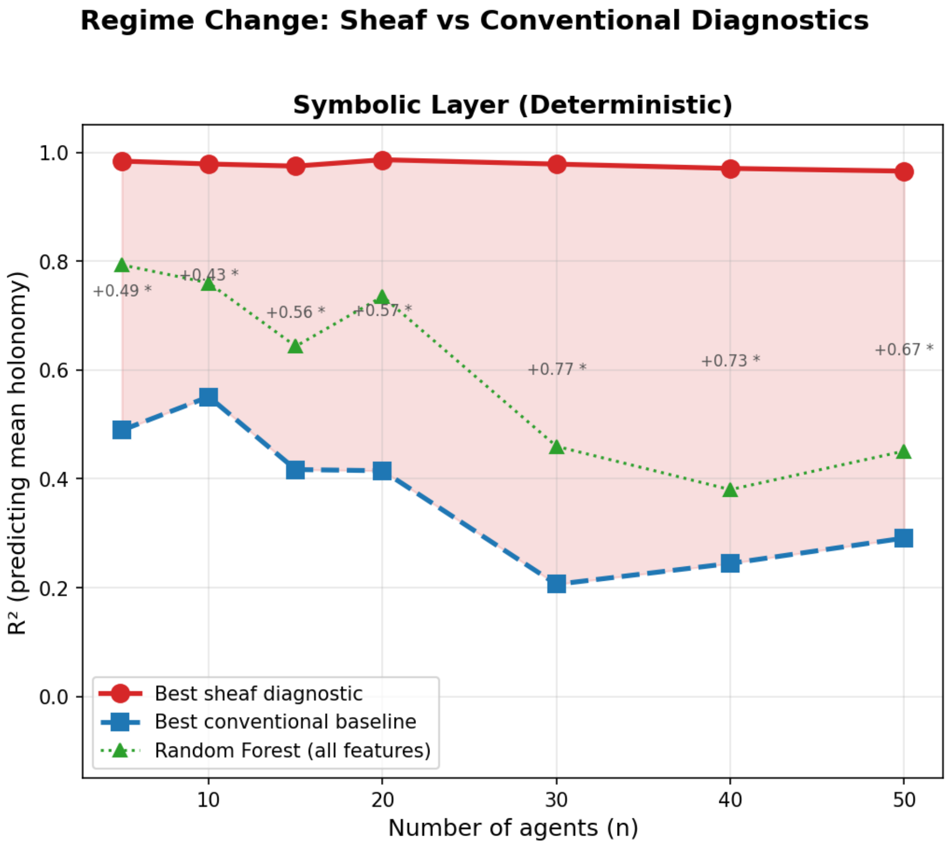


Figure 2: **Regime change: sheaf vs. conventional diagnostics.** R^2 for the best sheaf diagnostic, best conventional single-predictor baseline, and Random Forest across all scales. The shaded region shows the widening gap. Asterisks denote scales where the selection-safe bootstrap CI excludes zero.

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.

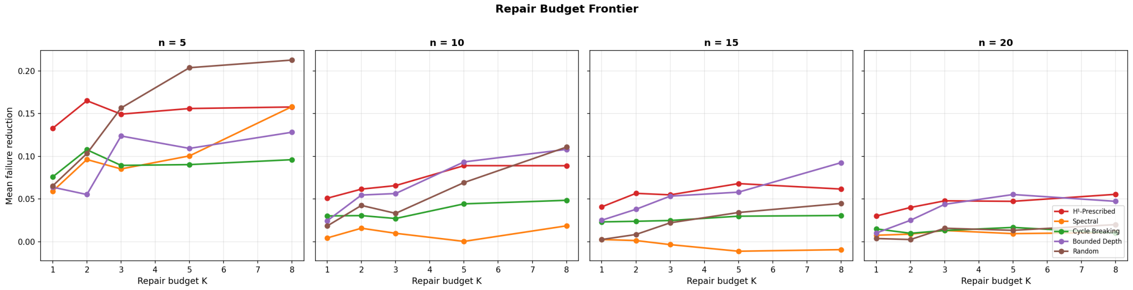


Figure 3: **Repair budget frontier.** Mean failure reduction vs. repair budget K (number of edge-level convention-harmonization operations), faceted by scale. H^1-prescribed repair dominates or matches the frontier at every budget level. Evaluated on all frustrated graphs (38 at $n = 5$, 75 at $n \geq 10$).

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.

Part V

Benchmark and Real-Protocol Evidence

Benchmark and Real-Protocol Evidence

Parts I–IV established the obstruction (SCPI), showed it appears in practice (Bridge), specified the protocol response (Seam), and demonstrated that the problem scales (SHEAF, Communication Bottleneck, Coherence Cliff). Part V 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**.

Three 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.

What follows are the Bronze+ and Silver technical notes in full, followed by the condensed BABEL benchmark summary, as the most compact exhibits of the program's current empirical reach.

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

```
┌──────────────────────────────────────────────────────────────────┐
│                    Workflow Composition                           │
│                                                                  │
│  ┌──────────────┐   ┌──────────────┐   ┌──────────────────┐     │
│  │ CalendarCore │──►│  TeamRouter  │──►│ EscalationEngine │     │
│  │  (custom)    │   │ (custom)     │   │   (custom)       │     │
│  │              │   │              │   │                  │     │
│  │ parse_invite │   │ resolve_tz   │   │ classify_priority│     │
│  │ normalize_time│    │ route_notif   │    │ compute_window   │  │
│  └──────┬────────┘    └───────────────┘    └────────┬─────────┘ │
│         │                                           │            │
│         │             ┌───────────────┐             │            │
│         └────────────►│   Memory      │◄────────────┘            │
│                       │  (official)   │                          │
│                       │               │                          │
│                       │ create_entities │                        │
│                       │ open_nodes      │                        │
│                       │ search_nodes    │                        │
│                       └───────────────┘                          │
└──────────────────────────────────────────────────────────────────┘
```

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.

Step	Server	Tool	Key output	Valid?	Latent issue
1	Calendar	`parse_invite`	event_time = 14.0 (UTC)	✓	Anchored to UTC
2	Calendar	`normalize_time`	normalized = 14.0	✓	—
3	Router	`resolve_timezones`	utc_event_time = 9.0	✓	Treats 14.0 as local, subtracts 5h
4	Escalation	`classify_priority`	priority = 1.0, escalation	✓	P-scale maps to high urgency
5	Router	`route_notification`	delivery windows computed	✓	—
6	Escalation	`compute_escalation`	**start = 13.5**	✓	Send-time anchor: 30 min before
7	Memory	`create_entities`	audit entity 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	R² 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.

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

Server	Provenance	Role	Convention Cluster
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
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):

Step	Server	Tool Call	Local Result	Valid?
1	MarkItDown	`convert_to_markdown`	Markdown text with amounts, dates	Yes
2	InvoiceParser	`parse_invoice_markdown`	total_amount: 18026.77 (dollars)	Yes
3	InvoiceParser	`extract_line_items`	4 line items extracted	Yes
4	SettlementEngine	`compute_settlement`	converted: 1,948,891.16 (cents!)	Yes (locally)
5	SettlementEngine	`compute_fees`	fee: 270.40 (on "net" of cents value)	Yes (locally)
6	SettlementEngine	`prepare_ledger_entry`	gross: 18026.77, fee: 270.40	Yes (locally)
7	Memory	`create_entities`	Ledger entity created	Yes
8	Memory	`open_nodes`	Ledger entity retrieved	Yes

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	R²	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
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.

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)	$\Delta\varrho$ 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 $\varrho = 0.80$ with oracle matrices — strong ranking — but $R^2 = -4.0$, meaning it cannot compute correct magnitudes. Opus 4 shows the largest information-extraction delta ($\Delta\varrho = +0.70$): 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]"


python -m coherence_gym evaluate-method --method structural_sheaf --split dev
python -m coherence_gym evaluate-method --method cycle_plain --split dev


python -m coherence_gym evaluate --split dev --output-dir results_check


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.

Part VI

Interpretability Boundary

Interpretability Boundary

Part V showed the obstruction is measurable and benchmarkable. Part VI 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 × 20 compositions, invoice domain. Scale sweep.
- **Phase 2c** (120): 3 domains × 2 scales × 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 × domain × 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× larger. SAE: `gemma-scope-2b-pt-res-canonical`, layer 20, 16k features.

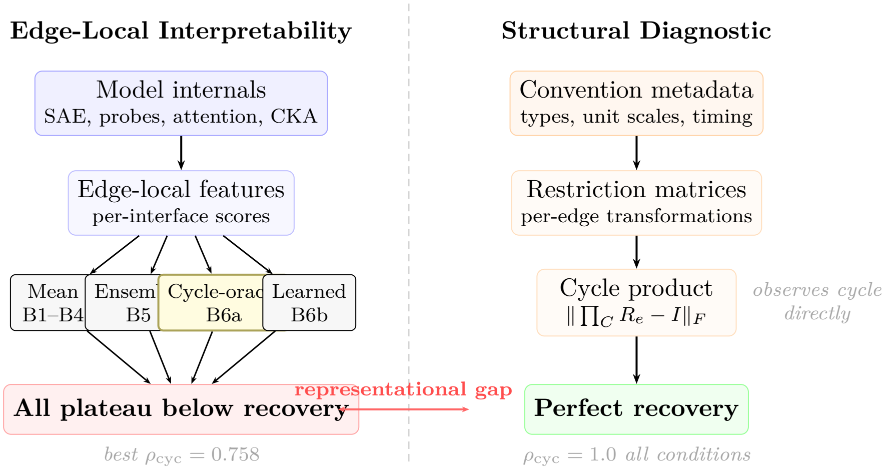


Figure 1: Two information paths to the same diagnostic target. **Left:** every tested aggregation of edge-local interpretability features—including cycle-oracle aggregation with oracle knowledge of graph topology (B6a, highlighted)—plateaus well below full recovery of compositional failure. **Right:** 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 ($0.685 \rightarrow 0.515$).

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 ($\beta_1 = 0$).

Part VII

Formal and Implementation Appendices

Appendix A: Proof Status Ledger

This appendix records the proof-status surface most directly relevant to the formal center of the omnibus.

SCPI Lean 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.

SCPI File Ledger

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.

File	Status	Main burden carried
`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.

What This Means

The formal center is not all at the same epistemic level.

- The core definitions and several decisive obstruction / conservativity results are machine-checked.
- Some higher-level schema-discovery statements are already typed precisely but are not yet fully closed at the proof layer.
- The paper is therefore strongest where it names the obstruction sequence and weaker where it reaches into the more ambitious functorial discovery perimeter.

That asymmetry is a feature of the present state, not a defect of the omnibus. The point of the ledger is to prevent machine-checked, paper-proved, and still-open material from being mistaken for one another.

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	Status
Tree control	0	n/a	n/a	5/5 contractible	n/a	topology-changing control
Single cycle (v1)	1	5/5	0/5	4/5	yes (A05)	structural-stable; repair mixed (behavioral tail)
Shared-edge double (v2)	2	5/5	0/5	5/5	yes (MC04)	repair-stable; behavior-sensitive on decomposition
Figure-eight (v3)	2	5/5	0/5	5/5	yes (F04)	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 → 0.515 → 0.321 → −0.067 from n = 5 to n = 20. The structural diagnostic remains at ϱ = 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 (≥ 0.995). The structural diagnostic achieves ϱ = 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.

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 H1(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`

File	Domain	Tools	Fee	Blind spots
`code_review_pipeline.yaml`	DevOps	4	3	`diff_format`, `severity_threshold`, `style_convention`
`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_tournam`
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.

Policy	Candidate	Fee	Review h	Rework h	Fail sign-off	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.py`
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 `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 thansettled center.′′ That asymmetry is a quality signal, and the omnibus should preserve it.