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

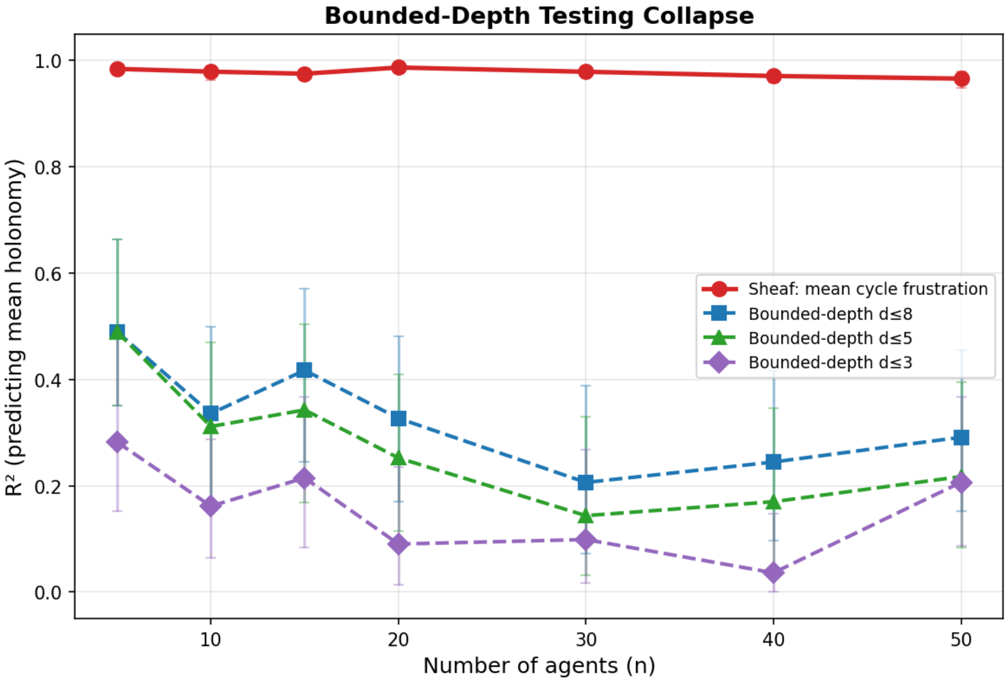

Figure 1: **Bounded-depth testing collapse (the title-earning figure).** R^2 for predicting mean holonomy vs. agent count. Bounded-depth baselines ($d \leq 3, 5, 8$) collapse with scale while mean cycle frustration (sheaf diagnostic) holds steady above $R^2 = 0.96$. Error bars: 95% bootstrap CI (1000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

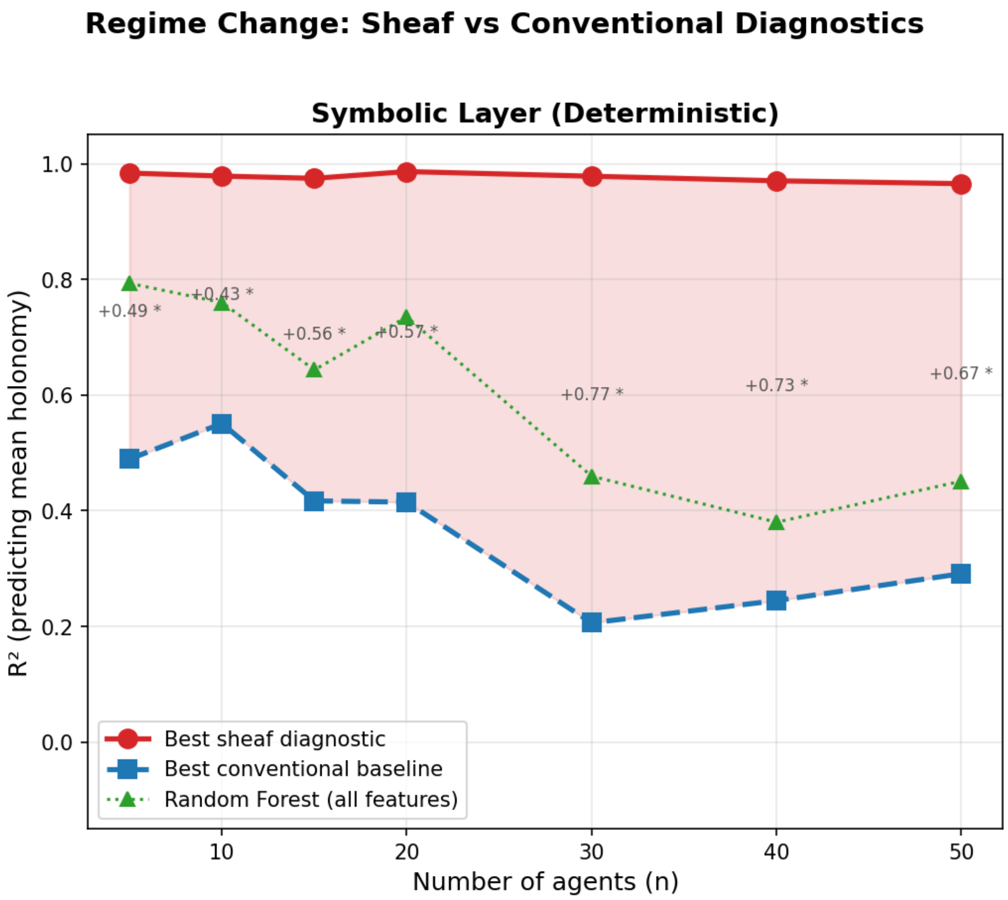

Figure 2: **Regime change: sheaf vs. conventional diagnostics.** R^2 for the best sheaf diagnostic, best conventional single-predictor baseline, and Random Forest across all scales. The shaded region shows the widening gap. Asterisks denote scales where the selection-safe bootstrap CI excludes zero.

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.

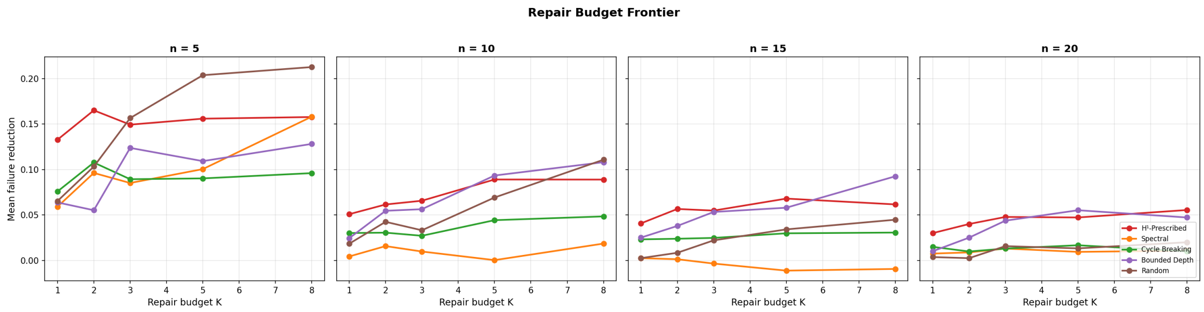

Figure 3: **Repair budget frontier.** Mean failure reduction vs. repair budget K (number of edge-level convention-harmonization operations), faceted by scale. H^1-prescribed repair dominates or matches the frontier at every budget level. Evaluated on all frustrated graphs (38 at $n = 5$, 75 at $n \geq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

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.