Compositional Incoherence is a Parity

Local Audit Carries Zero Signal; Recovering Conventions from Observation is SQ- and LPN-Hard

John Komkov, Independent Researcher

June 2026

Contents

Abstract

A composition of agents or tools is semantically coherent when the conventions its components attach to shared quantities agree around every cycle. A companion result (Komkov, Local Validity Does Not Compose, 2026) shows that no bounded-radius local audit can certify coherence: in a rank-1 signed model the global obstruction is a cohomology class in $H^1(G; Z/2)$, and on high-girth graphs two compositions can be locally identical yet differ arbitrarily in this obstruction. That result is an information-theoretic statement about two adversarially chosen instances. This paper recasts the problem as learning and quantifies the barrier, starting from one elementary observation: the holonomy of a composition around a cycle is a parity of edge signs, so recovering its coherence structure is learning a parity. From this we obtain three results. (1) Information-theoretic — on a graph of girth $> 2r$, the radius-r local views are identical across all obstructions, so local interpretability obtains zero signal at any sample size and cannot even detect incoherence; this generalizes the companion theorem from a pair of instances to the whole obstruction family, and we close an operator-class gap by showing non-backtracking spectral statistics are also blind below the girth (their low-order moments vanish identically). (2) Unconditional — detecting whether a composition is incoherent is globally trivial (one query measuring average holonomy), but recovering which conventions disagree — the repair-relevant obstruction class, of size k — requires $2^{\Omega(k)}$ correlational statistical queries or tolerance $2^{-\Omega(k)}$. This is the class mechanistic interpretability lives in, and it explains the measured failure of Komkov, Interpretability Frontier, 2026 (99.8% edge-local probe accuracy, zero compositional signal). (3) Cryptographic — under realistic observation noise, recovering the obstruction reduces from Learning Parity with Noise (F_2 model), and — in the abelian/linearized Z_q regime — from Learning With Errors, so it is hard for every probabilistic polynomial-time algorithm. The results stack into a clean separation: the noiseless recovery problem is SQ-hard but PPT-easy (a linear solve, given the declared structure); the noisy problem is hard for everyone; and interpretability, being local and statistical, is defeated in every regime. The consequence is constructive: conventions cannot be recovered from observation, so they must be declared and typed — which is exactly the program's prescription, here shown to be a necessity rather than a convenience.

---

1 1. Introduction

1.1 1.1 The question

Assurance for composed AI systems is overwhelmingly local: each model or tool is benchmarked in isolation, each producer–consumer interface is checked pairwise, agent trajectories

are sampled within a depth budget, and mechanistic interpretability inspects features and circuits inside individual components. The companion paper [LV] proves that this evidence does not compose: there exist families of compositions identical to every bounded-radius local audit yet carrying arbitrarily many independent semantic obstructions. The obstruction — the coherence fee — is the rank difference `rank(δ_full) − rank(δ_obs)`, and in the distilled rank-1 model it is the number of fundamental cycles whose holonomy is nontrivial, an element of `H¹(G; Z/2)`.

[LV] is an impossibility theorem about distinguishing two fixed instances. But real interpretability and audit tools are not handed two adversarial twins; they are handed one system and asked to learn its coherence structure by computing statistics of observations. The right question for that practice is not "can two instances be told apart" but "what can be learned about the obstruction from statistical observations, and at what cost?" This paper answers that question, and the answer turns out to depend sharply on what is being asked (detect vs. recover) and who is asking (local vs. global, statistical vs. unrestricted).

1.2 1.2 The unifying observation

Fix a connected graph `G` with first Betti number `β₁(G) = k` and a spanning tree `T`; the `k` non-tree edges close fundamental cycles `γ₁,…,γ_k`. In the rank-1 model a connection assigns a sign `g_e ∈ {±1}` to each edge, and the holonomy around a cycle is `Hol(γ) = ∏_{e∈γ} g_e`. After gauge-fixing on `T` ([LV, Lemma 1]), the connection is determined by its holonomies on the fundamental cycles, i.e. by a vector `ω ∈ F₂^k` with `ω_i = [Hol(γ_i) = −1]`. For any cycle `z = ∑_i z_i γ_i` in the cycle space `F₂^k`,

$$\mathrm{Hol}(z) \;=\; (-1)^{\langle \omega, z\rangle} \;=\; \chi_\omega(z).$$

Observing the holonomy of a cycle is evaluating a parity; recovering the obstruction `ω` — which conventions disagree, hence what to repair — is learning the parity `χ_ω`. Every result below is an instance of this identification.

1.3 1.3 Contributions

We separate two tasks and two auditors.

Tasks: detection (decide `ω = 0` vs `ω ≠ 0`: is the composition coherent at all?) and recovery (output `ω`: which fundamental cycles are twisted — the witness a repair needs, e.g. the disclosure normal form of [CD, Thm 7.1]). Auditors: local-statistical (interpretability: statistics of radius-`r` neighborhoods), global-statistical (an idealized auditor that may correlationally query arbitrary cycles to bounded precision), and global-unrestricted (any PPT algorithm).

- Theorem 1 (§4.1–4.2): local audit sees nothing. On a graph of girth `> 2r`, the radius-`r` local views are identical across all `ω`. Hence local interpretability obtains zero signal — it cannot detect or recover, at any sample size, under any cross-root aggregation. This generalizes [LV, Thm A] from the pair `(C₀, C_ω)` to the whole family `{C_ω}`, and (Prop 4.1) we extend

[LV, Lemma 6] to non-backtracking spectral statistics, whose low-order moments vanish identically below the girth.

- Theorem 2 (§4.3): detection is globally trivial, recovery is globally SQ-hard. A single correlational query ($\varphi \equiv 1$, measuring average holonomy) detects incoherence. But recovering ω requires $2^{\Omega(k)}$ correlational statistical queries or tolerance $2^{-\Omega(k)}$ — an unconditional lower bound against the entire statistical-query class, which contains essentially every interpretability method. This explains [IF] mechanistically: edge-local probes saturate at their local task yet carry zero composition signal, with the gap widening as k grows.
- Theorem 3 (§5): recovery under noise is cryptographically hard. When holonomy is observed through a noisy channel, recovering ω reduces from search-LPN ($\mathbb{F}_2$ model) and search-LWE ($\mathbb{Q}$-valued Bulla model). Hence no PPT algorithm recovers the obstruction from noisy observation, under standard assumptions.

The results stack (§5.3): noiseless recovery is SQ-hard but PPT-easy (Gaussian elimination, given declared structure); noisy recovery is hard for every PPT algorithm. The throughline (§5.4): one cannot recover conventions from observation — so they must be declared and typed. The program's global fee computation (Bulla) is easy precisely because it reads declared conventions (noiseless, direct algebraic access to δ_{full}); the hardness bites only when one tries to infer undeclared conventions from behavior, which is exactly what interpretability attempts.

1.4 1.4 What this paper is about: the certification axis, not severity

The program measures two different quantities that must not be conflated. The coherence fee $\dim H^1$ (a rank/count) lives on the certification axis — "is there an obstruction, which ones, how many?" A separate magnitude, cycle frustration $\|\Pi R_e - I\|$, lives on the severity axis — "how large is the resulting error?" These are orthogonal: a companion scaling study finds $\dim H^1$ a poor severity predictor while frustration magnitude predicts well. This paper is entirely on the certification axis, where the rank/parity object is provably the right one; nothing here rests on the contested claim that the fee predicts failure magnitude. This is the cleaner half of the program and the half that survives any reviewer.

1.5 1.5 Why a hardness theorem, given the impossibility theorem

Three additions over [LV]. Model: [LV] distinguishes two instances; we learn over a distribution, the model interpretability inhabits, and Theorem 1 upgrades "two instances are indistinguishable" to "the obstruction is absent from the local-view distribution entirely." Quantification: [LV] is yes/no; Theorem 2 gives an exponential resource lower bound parameterized by the cycle count, explaining why scaling interpretability does not help. Computational hardness: Theorem 3 moves from "no statistical method works" to "no efficient method works at all, un-

der noise" — a cryptographic-grade barrier the impossibility theorem does not reach. The cost: Theorem 3 is conditional on LPN/LWE; Theorems 1–2 are unconditional.

1.6 1.6 Claim discipline

[Established]: the holonomy model and Lemmas 1–2, 6 of [LV] (recalled in §2); orthonormality of characters and the parity SQ lower bound (Kearns [Ke], Blum–Furst–Jackson–Kearns–Mansour–Rudich [BFJKMR]); the LPN/LWE hardness assumptions ([BFKL], [Re]). [New, proven here]: Theorem 1 (whole-family local blindness, in the support-radius formulation, + non-backtracking extension), Theorem 2 (the detection-easy/recovery-hard split with explicit `q·τ² ≥ 2^k − 1`), Theorem 3 (the LPN/LWE reductions for convention recovery, the LWE half confined to the abelian/linearized regime). [Mechanized, Lean — sorry-free, standard axioms]: Theorem 2's counting/Markov step (`few_large_coeffs`), the adversary pigeonhole (`two_indistinguishable`, `csq_identifies_forces_many_queries`), and the assembled bound `q ≥ (2^k − 1) τ²` (`csq_query_lower_bound_parity`) are formalized in `CompositionDoctrine/LocalCertificationSQ.lean` and independently Aristotle-verified (stamp `b4f87803`); the orthonormality/Parseval input (`∑_ω ⟨φ, χ_ω⟩² ≤ 1`) and the SQ-model facts (per-query functional, union bound, identification) enter there as cited hypotheses, not re-proved — so the mechanized claim is precisely "counting skeleton + bound assembly," not "all of Theorem 2." [Deferred]: the Lean formalization of Theorem 1 (stubbed as `eval_correctness_inseparability_abstract`; the genuine obstruction is the unproved `…_exists`) and the [IF]-probe LLM scaling experiment (§6) remain follow-on work — a finite arithmetic demonstration of Theorems 1–2 (no LLMs; a genuine measurement with a negative control) accompanies the paper as `qa_sq_scaling.py`. Falsification (§7.4): a radius-`r` statistical query correlated with the obstruction (refutes Thm 1); a `poly(k)`-query constant-tolerance learner recovering `ω` (refutes Thm 2); a PPT recoverer from noisy observation (refutes Thm 3 or breaks LPN/LWE).

---

2 2. Setup: holonomy as a parity

Definition 2.1 (composition graph, connection). A composition graph `G = (V, E)` is a connected finite simple graph (vertices = tools/agents, edges = data-flow interfaces). A rank-1 signed connection assigns `g_e ∈ {±1}` to each edge; a gauge transformation `h ∈ {±1}^V` acts by `g_e ↦ h_v g_e h_u`. Gauge models local relabeling (renaming a field, rescaling units at one tool) and is semantically inert; all observations below are gauge-invariant, per [LV, Def 2.6].

Definition 2.2 (holonomy, obstruction). For a cycle `γ`, `Hol(γ) = ∏_{e∈γ} g_e ∈ {±1}` is gauge-invariant. Fixing a spanning tree `T`, the `k = β₁(G)` fundamental cycles identify gauge classes of connections with `H¹(G; Z/2) ≅ F₂^k`; write `ω` for the class, `ω_i = [Hol(γ_i) = −1]`. The composition is coherent iff `ω = 0`; the coherence fee is `wt(ω)` (equivalently `rank(δ_full) −`

rank(δ_obs); see [LV, §7], [CD]).

Lemma 2.3 (parity form; [Established]). In the fundamental-cycle basis of $Z_1(G; F_2) \cong F_2^k$, for every $z \in F_2^k$ the holonomy of $\sum_i z_i \gamma_i$ is $\mathrm{Hol}(z) = (-1)^{\{\omega,z\}} = \chi_\omega(z)$.

Proof. Holonomy is multiplicative over edges and (± 1)-valued. Passing to additive F_2 notation, the edge-support of $\sum_i z_i \gamma_i$ is the symmetric difference of the supports of the chosen γ_i: an edge shared by two of them is traversed twice and contributes $g_e^2 = 1$, so it cancels. Hence the holonomy of $\sum_i z_i \gamma_i$ is the F_2-sum of the per-cycle holonomy bits, namely $\langle \omega, z \rangle$. ■

Lemma 2.4 (tree triviality; [Established] = [LV, Lemmas 1–2]). Every rank-1 connection restricted to a tree is gauge-trivial. Hence if girth(G) > 2r, every radius-r ball $B_r(v)$ is a tree, and the connection restricted to $B_r(v)$ is gauge-equivalent to the trivial connection — for every ω.

The characters $\{\chi_\omega : \omega \in F_2^k\}$ are orthonormal under the uniform distribution U on F_2^k: $E_{z\sim U}[\chi_\omega \chi_{\omega'}] = [\omega = \omega']$ (O'Donnell [OD]). This is the only analytic fact we use.

---

3 3. The statistical-query model

We use the correlational statistical query (CSQ) model, which is exactly the model interpretability inhabits: a measurement estimates the average, over inputs, of a feature-statistic (optionally multiplied by an outcome), to finite precision.

Definition 3.1 (CSQ oracle). Fix a target function $f: X \to \{\pm 1\}$ and a distribution D over X. A query is a function $\varphi: X \to [-1,1]$. The oracle STAT_D(f) answers (φ, τ) with any v such that $|v - E_{x\sim D}[\varphi(x) f(x)]| \le \tau$; τ is the tolerance and $1/\tau^2$ the effective sample size. A CSQ learner makes adaptive queries and outputs a hypothesis.

Definition 3.2 (cycle-query model — global statistical auditor). $X = F_2^k$ (the cycle space), $D = U$ uniform, target χ_ω (Lemma 2.3). A query $\varphi: F_2^k \to [-1,1]$ is a chosen weighting of cycles when estimating average holonomy. This grants global cycle access (any cycle, including long ones) but only statistical, bounded-precision answers.

Definition 3.3 (local-view model — interpretability auditor). The example is a uniformly random rooted radius-r ball $\mathrm{view}(B_r(v))$, $v \sim U(V)$, recording the ball's structure and the gauge class of the connection on it. An r-local query is any function of $\mathrm{view}(B_r(v))$. Every SAE-feature statistic, linear probe, attention statistic, and representational-similarity score is a bounded function of node/edge-local gauge-invariant measurements (the σ-algebra $\mathcal{F}_{\mathrm{edge}}$ of [IF, Prop 3.3]), hence an r-local query for the relevant receptive radius r.

Remark 3.4 (CSQ vs general SQ). A general SQ $\psi(x,y)$ decomposes into a target-independent part and a correlational part; lower bounds for CSQ transfer to general SQ up to constants

(Feldman [Fe]). We state CSQ results and write "SQ" loosely.

---

4 4. Unconditional results

4.1 4.1 Theorem 1: local audit carries zero signal

Define the support radius $\rho(\varphi)$ of an observation φ to be the smallest ρ such that φ depends only on the gauge class of the connection restricted to some ball $B_\rho(v)$ (single-ball; see Remark 4.2 for aggregates). An r-local query (Definition 3.3) has $\rho \le r$.

Theorem 1 (whole-family local blindness). Let G have girth $> 2\rho$. Then the gauge class of the connection restricted to any ball $B_\rho(v)$ is the trivial-tree class for every obstruction ω — the same data across all ω. Consequently every observation of support radius $\le \rho$ (in particular every r-local statistical query with $r \le \rho$, with the oracle averaging a single-ball function over a random root) has an ω-independent value; the channel of support-ρ observations carries zero mutual information about ω; and no learner restricted to such observations can detect incoherence, recover ω, or estimate $\mathrm{wt}(\omega)$ to additive error $< k/2$.

Proof. By Lemma 2.4, $\mathrm{girth}(G) > 2\rho$ forces each $B_\rho(v)$ to be a tree, on which the connection is gauge-trivial. The observation records the gauge class, which is therefore the trivial-tree class for every v and every ω — identical data across all ω, not merely identically distributed. Any function of ω-independent data is ω-independent; mutual information is 0; a channel independent of ω permits no inference about ω. For the fee estimate, an ω-independent output differs from the truth by $\mathrm{wt}(\omega)$, which is 0 at $\omega = 0$ and k at any weight-k ω, forcing additive error $\ge k/2$ on one of them. ■

This is the sharpest form of the program's central impossibility: the obstruction is absent from the support-ρ data entirely, so no statistic, however clever or data-rich, can detect it. It is strictly stronger than [LV, Thm A], which fixes one pair (C_0, C_ω); here all 2^k obstruction classes present identical local data.

Remark 4.2 (what "local" must mean — aggregation and support radius). The theorem is about an observation's support radius, not the radius of any single ball one happens to name. A gauge-invariant quantity of a union of balls is an observation of larger support radius, and it can be ω-dependent: on a girth-g graph, the holonomy of a cycle is a gauge-invariant function supported on a region of radius $\approx g/2$, so an auditor that re-glues observations across overlapping balls into support radius $\rho' \ge g/2$ is not covered — and should not be, since computing the global fee is exactly such a (global) observation, which the program advocates. The content is therefore precise: any audit of bounded locality ρ is blind whenever $\mathrm{girth} > 2\rho$; mechanistic interpretability, whose features and probes have bounded receptive field, is the $\rho = O(\text{depth})$ case, and high-girth (long-cycle) compositions push $g/2$ past any fixed ρ. Cross-correlational probes (e.g. CKA between two sites at distance d) have support radius $\approx d/2 + r$ and are blind whenever

girth exceeds twice that. The single SQ query never re-glues: it averages one bounded-support function over roots, so its support radius is that of the function, and Theorem 1 binds it.

4.2 4.2 Closing the operator-class gap: non-backtracking statistics are also blind

[LV, Lemma 6] shows any spectral statistic built from walk-sum moments tr(A^m) of order m < girth is ω-independent (a closed walk of length < girth must backtrack, and backtracking weights cancel). A skeptic asks about non-backtracking (Hashimoto) operators B, known to detect cycle structure that adjacency spectra miss. We close the gap.

Proposition 4.1 (non-backtracking blindness). Let B be the connection-weighted non-backtracking operator of G, (B)_{e,e'} = g_{e'} if e' follows e without backtracking, else 0. For every m < girth(G), tr(B^m) = 0 identically (hence ω-independent); more generally any statistic that is a function of non-backtracking closed walks of length < girth is ω-independent.

Proof. The diagonal of B^m sums products B_{e_1 e_2} ⋯ B_{e_m e_1}, and the closing factor B_{e_m e_1} imposes non-backtracking on the wrap-around, so tr(B^m) counts cyclically reduced (tailless) closed non-backtracking walks of length m, weighted by ∏ g_{e_i}. Such a walk's edge-image contains a cycle as a subgraph, so in a simple graph its length is at least girth(G). Thus for m < girth(G) there are no such walks and tr(B^m) = 0 identically — independent of ω. ■

So non-backtracking statistics are not merely blind below the girth — their low-order moments vanish identically, conveying even less than backtracking ones. Every spectral statistic of order below the girth — walk-sum or non-backtracking — is an r-local query in the sense of Theorem 1 and carries zero signal. Detecting the obstruction spectrally requires moments of order ≥ girth, i.e. global spectral access (consistent with [LV, §7]).

4.3 4.3 Theorem 2: detection is trivial, recovery is hard — globally and statistically

Theorem 1 forbids local learning outright. With global cycle access the two tasks separate sharply.

Theorem 2 (CSQ complexity of detection vs recovery). In the cycle-query model (Definition 3.2):

1. (Detection is one query.) The constant query φ ≡ 1 returns ⟨1, χ_ω⟩_U = E_{z~U}[χ_ω(z)] = [ω = 0]. Hence a single CSQ at tolerance τ < 1/2 decides coherence (ω = 0) vs incoherence.
2. (Recovery is exponentially hard.) Any CSQ learner that outputs ω using queries of tolerance τ makes a number of queries q with

$$q \;\geq\; (2^k - 1)\,\tau^2.$$

Thus recovery needs $q \geq 2^{\Omega(k)}$ (constant τ) or $\tau \leq 2^{-\Omega(k)}$ (effective sample size $2^{\Omega(k)}$).

Proof. (1) $E_z[(-1)^{\langle \omega, z \rangle}] = 1$ if $\omega = 0$ and 0 otherwise, by orthonormality; $\varphi \equiv 1$ is a legal CSQ. (2) Fix a query $\varphi: \mathbb{F}_2^k \to [-1,1]$; since $|\varphi| \leq 1$ pointwise, $E_U[\varphi^2] \leq 1$. By Parseval's identity over the complete orthonormal basis $\{\chi_\omega\}_{\omega \in \mathbb{F}_2^k}$, $\sum_\omega \langle \varphi, \chi_\omega \rangle_U^2 = \|\varphi\|_U^2 \leq 1$, so $\#\{\omega : |\langle \varphi, \chi_\omega \rangle| > \tau\} < 1/\tau^2$. Run the adversary that answers 0 to every query; 0 is a legal τ-approximation for target ω whenever $|\langle \varphi, \chi_\omega \rangle| \leq \tau$, i.e. for all but $< 1/\tau^2$ targets per query. After q queries, the set of targets ever forced to a nonzero answer has size $< q/\tau^2$; every other target produces the identical all-zero transcript, and — since the adversary's answers do not depend on the true target — the learner's full adaptive query sequence and final output are a fixed object across that set (adaptivity gives no leverage). If $q/\tau^2 < 2^k - 1$, at least two distinct targets share it, so a deterministic learner correct on all targets has $q \geq (2^k - 1)\tau^2$. A randomized learner succeeding with probability $\geq 2/3$ over a uniform target faces an identical output distribution across the indistinguishable set, capping success at $1/|\text{set}|$, giving the same bound up to a constant. ■

The proof is deliberately self-contained — one Parseval step plus an adversary — because the parity class is exactly orthonormal and the elementary bound is sharp and auditable; it is the parity SQ lower bound (Kearns [Ke]; Blum–Furst–Jackson–Kearns–Mansour–Rudich [BFJKMR]) in explicit form. The interpretation is the crux: detecting that something is wrong is cheap with global access, but recovering what is wrong — the information a repair requires — is exponentially hard for any statistical method, and mechanistic interpretability is a statistical (CSQ) method. This is exactly the measured behavior in [IF]: edge-local probes reach 99.8% accuracy at their local task yet carry zero composition-level signal, and the gap widens with scale — the empirical shadow of an exponential lower bound. (Locality makes it worse still: by Theorem 1 even detection is impossible for a local auditor, since $\varphi \equiv 1$ over global cycles is unavailable to it.)

---

5 5. The noisy regime: LPN- and LWE-hardness

Noiseless recovery is SQ-hard (Theorem 2) but not hard for all algorithms: given global noiseless samples $(z, \chi_\omega(z))$, Gaussian elimination over $\mathbb{F}_2$ recovers ω in $O(k^3)$. (This is the classical separation: parities are SQ-hard but PPT-easy.) Why, then, is recovery hard in practice for every method? Because real observation is noisy — conventions are inferred from finite, fallible evidence. Noise converts the PPT-easy problem into a canonical hard one.

5.1 5.1 The noisy observation model

Definition 5.1 (noisy holonomy samples). Fix G with $\beta_1 = k$ and obstruction ω. A noisy holonomy sample is (z, b) with $z \sim U(\mathbb{F}_2^k)$ and $b = \langle \omega, z \rangle \oplus e$, $e \sim \mathrm{Bernoulli}(\eta)$ iid, $\eta \in (0, 1/2)$ a fixed constant. The primitive that carries the noise is the holonomy measurement itself: each round-trip consistency test of a cycle z succeeds or fails with an independent error e, because the auditor infers each convention from finite, fallible behavioral evidence. This per-measurement model is what the reduction uses, and it is the honest one. (A per-edge $\mathrm{BSC}(\eta_0)$ story — read each edge transport noisily and XOR along the cycle — is a special case valid only for bounded cycle length: along a length-ℓ cycle the effective rate is $\eta = (1 - (1-2\eta_0)^\ell)/2$, which tends to $1/2$ as ℓ grows. Since LPN draws $z \sim U(\mathbb{F}_2^k)$ of typical weight $\approx k/2$, the per-edge story does not by itself realize a fixed η bounded away from $1/2$; the per-measurement model above does, and is the model we posit.)

5.2 5.2 Theorem 3: convention recovery is LPN-hard (and LWE-hard)

Theorem 3 (cryptographic hardness of recovery). ($\mathbb{F}_2$.) If a PPT algorithm, given $\mathrm{poly}(k)$ noisy holonomy samples (Definition 5.1) for unknown $\omega \in \mathbb{F}_2^k$, outputs ω with non-negligible probability, then there is a PPT algorithm solving search-$\mathrm{LPN}_{k,\eta}$. Under the LPN hardness assumption [BFKL] (the best known algorithm is the sub-exponential $2^{O(k/\log k)}$ of [BKW]), no such recoverer exists. (Q-valued / Bulla, abelian regime only.) Take the linearized witness model [CD]: a single $\mathbb{Z}_q$-valued convention per dimension, so the holonomy of a cycle z is the additive form $\langle \omega, z \rangle \bmod q$ (the abelian coboundary / rank object of [CD §3], not the non-commutative matrix product). Observed with discretized-Gaussian noise, $b = \langle \omega, z \rangle + e \bmod q$ is by definition an LWE sample, so an efficient recoverer of $\omega \in \mathbb{Z}_q^k$ yields a search-$\mathrm{LWE}_{k,q,\chi}$ solver; under LWE [Re], none exists. Scope. This covers the certification-axis object only. The severity-axis holonomy $\prod_e R_e$ (a product of matrices, §1.4) is non-abelian — its "holonomy" is a conjugacy class, not a linear form $\langle \omega, z \rangle$ — and the LWE reduction does not apply to it; non-abelian convention recovery is left open (and [CD] flags non-abelian cohomology as unformalized).

Proof. ($\mathbb{F}_2$.) The reduction is the identity on data. Given an $\mathrm{LPN}_{k,\eta}$ instance — secret $s \in \mathbb{F}_2^k$, samples (a_i, c_i), $c_i = \langle a_i, s \rangle \oplus e_i$ — take the bouquet graph with k fundamental cycles (any connected G with $\beta_1 = k$ works), set $\omega := s$, and read each $a_i \in \mathbb{F}_2^k$ as the cycle $\sum_j (a_i)_j \gamma_j$. By Lemma 2.3 the noiseless holonomy of a_i is $\langle \omega, a_i \rangle$, so (a_i, c_i) is exactly a noisy holonomy sample (Def 5.1) for the composition with obstruction ω. A recoverer outputting ω solves search-LPN. (Q.) Identical with additive characters of $\mathbb{Z}_q$: a noisy linear holonomy observation $\langle \omega, z \rangle + e$ over $\mathbb{Z}_q$ is by definition an LWE sample with secret ω and coefficient vector z; recovery is search-LWE. ■

The reduction is a relabeling — the strongest possible form, with no loss or blowup: a noisy holonomy observation is an LPN/LWE sample. The content is the identification, not the manipulation. (Coarse detection, by contrast, stays globally easy under noise: $\mathbb{E}[b] = \eta$ if $\omega = 0$

and 1/2 if ω ≠ 0, a bias test — the noisy analogue of Theorem 2(1). Detection easy, recovery LPN-hard.)

5.3 5.3 The separation

Recovering the k-cycle obstruction, indexed by auditor and observation regime:

Auditor regime	Local statistical (interpretability)	Global statistical (idealized SQ)	Global, unrestricted PPT
Noiseless	impossible — zero signal (Thm 1)	2^{Ω(k)} queries/precision (Thm 2)	easy — Gaussian elimination
Noisy	impossible — zero signal (Thm 1)	2^{Ω(k)} (Thm 2; noise only worsens)	LPN/LWE-hard (Thm 3)

The only efficient recoverer needs all three of: global access, no observation noise, and a non-statistical (algebraic) algorithm. Reality offers an auditor none of them: interpretability is local and statistical (top-left/bottom-left — zero signal), and any observational inference is noisy (bottom row — LPN/LWE-hard). Detection of some incoherence is the lone easy cell (one global query), and even that is unavailable locally.

5.4 5.4 Consequence: declare conventions, do not recover them

The program computes the fee globally and efficiently (Bulla) — no contradiction with Theorem 3, because Bulla reads declared conventions: it has direct, noiseless, algebraic access to δ_full (top-right "easy" cell). The hardness bites only when one tries to infer undeclared conventions from observed behavior — precisely what mechanistic interpretability attempts, and precisely the local-and-statistical regime that is impossible (Thm 1) or noisy-hard (Thm 3). The constructive reading: there is no bottom-up route from observation to a coherence certificate; conventions must be made explicit — declared and typed — at the interface. This is the negative half of the program's thesis (coherence is locally unlearnable); the positive half (therefore carry it top-down as a typing invariant) is the subject of the companion type-system development, for which this paper is the justification.

---

6 6. The empirical signature (summary; experiment deferred)

[IF] runs six families of edge-local interpretability diagnostics on 240+ compositions across two models (GPT-2 Small, Gemma 2 2B) and reports 99.8% edge-local probe accuracy with zero composition-level signal, decaying further with scale. Theorems 1–2 explain this exactly: the

probes are r-local CSQs (Def 3.3), so Theorem 1 forces zero correlation and Theorem 2 forces the gap to widen with cycle count. The 99.8%/zero-signal pair is not a tooling deficiency to engineer away; it is the predicted shadow of an unconditional lower bound.

Pre-registered confirmatory experiment (follow-on). Theorem 2 predicts, beyond [IF]'s qualitative null, a quantitative law: probe sample-complexity for a fixed-accuracy holonomy classifier should grow exponentially in cycle length/girth. The experiment fixes [IF]'s probe family and convention generator, sweeps girth at fixed k, and measures sample-complexity to a target AUC; the prediction is a 2^{Θ(g)} curve. A polynomial curve refutes the assumption that real probes are CSQs. Flagged as the next deliverable; not claimed here.

---

7 7. Discussion

7.1 7.1 Relation to locally-checkable labelings

Theorem 1 belongs to the LCL tradition (Naor–Stockmeyer [NS], Linial): some global predicates admit local certificates, others provably do not. Our contribution within it is the identification of the obstruction with a parity, which licenses importing the parity SQ and LPN/LWE lower bounds; the parity SQ bound itself is classical [Ke], and we contribute its application, the explicit $q \cdot \tau^2 \geq 2^k - 1$ form, the detection/recovery split, and the observation that interpretability probes are exactly the CSQs it binds.

7.2 7.2 Relation to mechanistic interpretability

Not a critique of interpretability's goals but a delineation of its reach. Interpretability recovers local mechanism; Theorems 1–3 say global compositional coherence is a parity of edge data that no local statistic can correlate with and no efficient statistical method can recover. The fee is therefore not a competitor to interpretability but a missing global layer above it, and that layer cannot be reconstructed bottom-up from the lower one.

7.3 7.3 What the results do not prove

Theorem 1 bounds r-local statistical queries; it does not bound algorithms with global graph access (component counting, full-spectrum methods), which the program advocates (the fee is computed globally). Theorem 2 bounds the SQ class; noiseless recovery is broken by Gaussian elimination — which is why Theorem 3 (noise) is needed for hardness against all PPT algorithms. Theorem 3 is conditional on LPN/LWE. The high-girth hypothesis of Theorem 1 is worst-case; short-cycle compositions admit local detection up to the cycle length, so practical hardness scales with dependency-chain length. None of the theorems addresses severity (§1.4).

7.4 7.4 Falsification

Refuted by any one of: (a) a radius-`r` statistical query with nonzero correlation to the obstruction on a girth-`>2r` family (Thm 1); (b) a CSQ learner recovering `w` with `poly(k)` queries at constant tolerance (Thm 2); (c) a PPT recoverer from noisy observation, which would break LPN/LWE (Thm 3).

---

8 References

[LV] J. Komkov, Local Validity Does Not Compose: A Theorem on the Limits of Bounded AI Audit, Res Agentica program, 2026.

[IF] J. Komkov, Edge-Local Interpretability Is Not Enough for Cyclic Composition, Res Agentica program, 2026.

[CD] J. Komkov, A Witness Logic for Semantic Composition, Res Agentica program, 2026.

[Ke] M. Kearns, "Efficient noise-tolerant learning from statistical queries," Journal of the ACM 45(6) (1998), 983–1006.

[Fe] V. Feldman, "A general characterization of the statistical query complexity," COLT 2017.

[BFJKMR] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, S. Rudich, "Weakly learning DNF and characterizing statistical query learning using Fourier analysis," STOC 1994, 253–262. (Statistical dimension; parity has SQ-dimension `2^k`.)

[BFKL] A. Blum, M. Furst, M. Kearns, R. Lipton, "Cryptographic primitives based on hard learning problems," CRYPTO 1993, LNCS 773, 278–291. (LPN hardness assumption.)

[BKW] A. Blum, A. Kalai, H. Wasserman, "Noise-tolerant learning, the parity problem, and the statistical query model," Journal of the ACM 50(4) (2003), 506–519. (Best known LPN algorithm: sub-exponential `2^{O(k/\log k)}`.)

[Re] O. Regev, "On lattices, learning with errors, random linear codes, and cryptography," Journal of the ACM 56(6) (2009), Article 34.

[NS] M. Naor and L. Stockmeyer, "What can be computed locally?," SIAM Journal on Computing 24(6) (1995), 1259–1277.

[OD] R. O'Donnell, Analysis of Boolean Functions, Cambridge University Press, 2014.