On the Geometry of Reasoning Traces: Spectral and Ollivier–Ricci Signatures of Chain-of-Thought

Anonymous ACL submission

Abstract

We study the geometry of chain-of-thought (CoT) reasoning traces, treating each as a semantic ε-proximity graph and analyzing its spectral and discrete-curvature structure. We find that the 9-dimensional geometric signature decomposes interpretably: spectral entropy detects errors disproportionately on arithmetic / number-dense reasoning, while Ollivier–Ricci negative curvature detects errors disproportionately on symbolic-logic reasoning. Across 6 LLMs and 3 benchmarks (18 within-corpus (model, domain) conditions, 3,493 traces), the only-spectral and only-Ricci error groups differ on per-sentence numeric-token rate (Cliff's $\delta = +0.29$, $p = 9.0 \times 10^{-10}$) and on domain composition (χ^2 $p = 1.3 \times 10^{-11}$), with model composition serving as an independent corroborating signal (χ^2 $p = 8.2 \times 10^{-10}$). The same signature carries a model-characteristic fingerprint that beats a trace-length baseline on 6-way model identity (GBT accuracy 0.439 vs. length-only 0.349; $B - A = +0.090$, 95% CI $[0.072, 0.108]$), with the advantage carried by 7 length-independent *shape* features alone ($B_{\text{shape}} - A = +0.070$, CI $[0.052, 0.088]$). Both findings hold under three sentence encoders (all-MiniLM-L6-v2, all-mpnet-base-v2, BGE-small) and survive a uniform model-blind truncation-exclusion control. These results establish geometric trace structure as a decomposable, interpretable representation of CoT reasoning rather than a monolithic quality signal.

1 Introduction

Chain-of-thought (CoT) prompting (Wei et al., 2022) elicits step-by-step reasoning from large language models, and a growing line of work treats the resulting trace as a graph and analyzes its geometric structure. Noël (2026) use spectral properties of *internal* token-level attention graphs (algebraic connectivity, high-frequency energy, spectral entropy) to discriminate valid from invalid mathematical proofs at up to 95.6% accuracy. This is encouraging but requires attention extraction, which excludes closed-source models and adds nontrivial cost in deployment.

We extend this line of work in two directions and ask a different question than prior work. First, we replace internal access with *external* semantic ε-proximity graphs and combine spectral analysis with discrete Ollivier–Ricci curvature (Ollivier, 2009; Topping et al., 2022), yielding a 9-dimensional geometric signature requiring no per-step labels, no resampling, and no model-internal access. Curvature is included because spectral features measure global coherence but are blind to local bottlenecks. Second, instead of asking only *whether* geometry predicts correctness, we ask whether the geometric signature is *monolithic* (one "reasoning quality" axis) or *decomposable* into interpretable, specialized components.

We find decomposition. Across 6 LLMs (Claude Sonnet 4, gpt-oss 120B, Llama 3.1 8B, Mistral 7B, GPT-4o-mini, Gemini 2.5 Flash) and 3 reasoning benchmarks (GSM8K (Cobbe et al., 2021), FOLIO (Han et al., 2024), PrOntoQA (Saparov and He, 2023)), the spectral and curvature halves identify non-overlapping subsets of incorrect traces on every (model, domain) condition we measured (range 30.5% to 74.5%). The two halves are specialized: only-spectral catches are concentrated on GSM8K (number-dense arithmetic); only-Ricci catches on PrOntoQA (symbolic-logic deduction). Beyond modality specialization, the same 9-dimensional signature predicts model identity above a trace-length baseline, with the advantage carried by 7 length-independent shape features. Both findings replicate across three sentence encoders chosen to be distinct in architecture and training objective.

Contributions.

1. We show that the 9-dimensional geometric signature of CoT traces decomposes by rea-

soning type: spectral entropy specializes to errors on arithmetic / number-dense reasoning, while Ollivier–Ricci negative curvature specializes to errors on symbolic-logic reasoning. The split is confirmed on per-sentence numeric-token rate (Cliff's $\delta = +0.29$, length-controlled $\delta = +0.34$) and on domain composition (χ^2 $p = 1.3 \times 10^{-11}$), with model composition as an independent corroborating signal (χ^2 $p = 8.2 \times 10^{-10}$).

2. We show that the geometric signature carries a model-characteristic fingerprint, beating a strong trace-length baseline on 6-way model identity (GBT 0.439 vs. 0.349, $B - A = +0.090$, CI $[0.072, 0.108]$), with the advantage carried by 7 length-independent shape features alone.

3. We verify both findings under three sentence encoders (MiniLM, MPNet, BGE) at 6 models. The robustness gate is the intersection of {model-pool expansion, truncation decontamination, encoder change}; both findings pass.

4. We release a 130-line python Ollivier–Ricci implementation, validated against canonical graphs (P_6, K_4, two-triangle bridge) to three decimal places of published analytical values, developed after the standard `GraphRicciCurvature` library produced bus errors on our compute environment.

2 Background and Related Work

Spectral analysis of LLM reasoning. Noël (2026) introduced spectral diagnostics on internal attention graphs, showing that Fiedler value, high-frequency energy ratio, and spectral entropy of token-level attention Laplacians distinguish valid from invalid mathematical proofs. Noël (2025) extended this to hallucination detection, and Binkowski et al. (2025) used top-k eigenvalues of attention Laplacians for similar purposes. All of these methods require internal model access. We adapt the spectral framework to external sentence-level graphs and combine it with discrete curvature, with the goal of decomposing the geometric signal into interpretable, modality-specific components.

Structural analysis of CoT traces. Xiong et al. (2025) build directed reasoning graphs from CoT traces and extract node-level structural features (branching, convergence, exploration density) that correlate with accuracy. Jiang et al. (2025) convert deliberative long CoTs into hierarchical trees and apply supervised GNNs to extract structural patterns. Closest to our geometric framing, Tan et al. (2025) apply topological data analysis to reasoning traces and report that persistent-homology features predict reasoning quality more strongly than standard graph metrics; notably, they also caution that such features are embedding-dependent signals correlated with quality rather than faithful maps of the underlying reasoning program, a caveat our encoder-robustness analysis (Section 7) addresses directly. Saha et al. (2025) extract causal CoT graphs and use graph-aligned interventions to argue that reasoning nodes are causal contributors to the answer; their causal apparatus is complementary to our descriptive decomposition, for which we make no causal claim. These methods operate externally but require either deliberative long-CoT traces with exploratory structure, supervised training, or causal-intervention machinery. Our method is training-free, operates on any single linear CoT, and decomposes the trace's semantic graph spectrally and curvature-wise rather than by node-level descriptors.

Process supervision and verifier training. The dominant approach to CoT verification trains an external model to score reasoning at the step level. Lightman et al. (2023) demonstrate that process reward models trained on dense per-step human labels substantially outperform outcome-only supervision. Subsequent work has reduced labeling cost via MCTS rollouts (Wang et al., 2024; Luo et al., 2024) or LLM-as-judge (Zheng et al., 2023). Our approach is complementary and orthogonal in goal: rather than score correctness with a learned verifier, we characterize and decompose the geometric structure of the trace itself.

Discrete Ricci curvature on graphs. Ollivier (2009) introduced the random-walk-based curvature notion we use. On graphs it has been most prominently applied by Topping et al. (2022), who show that negatively curved edges drive "over-squashing" bottlenecks in graph neural networks, with Ollivier-Ricci-specific refinements by Nguyen et al. (2023). This GNN-side literature is directly relevant to our modality-split finding (Section 5): the bottleneck phenomenon on message-passing graphs is, on our semantic ε-proximity graphs, the structural correlate of reasoning that loses internal coherence. To our knowledge, Ollivier–Ricci cur-

vature has not previously been applied to LLM CoT traces.

Reasoning benchmarks. Cobbe et al. (2021) introduce GSM8K, an arithmetic word-problem benchmark on which CoT is heavily number-dense. Han et al. (2024) introduce FOLIO, an expert-annotated first-order logic benchmark with True/False/Unknown labels. Saparov and He (2023) introduce PrOntoQA, a synthetic symbolic-logic deduction benchmark. Together, the three benchmarks span a numeric vs. symbolic reasoning spectrum that turns out to be the axis along which our modality decomposition cleanly resolves.

3 Method

3.1 Geometric Signature Construction

Pipeline. For each CoT trace, we produce a 9-dimensional geometric signature. The trace is split into sentences via a regex tokenizer handling punctuation and line-break list markers; traces with fewer than three sentences are discarded. Each sentence is encoded with `sentence-transformers/all-MiniLM-L6-v2` (Reimers and Gurevych, 2019) (384-dim, L_2-normalized). We build an undirected graph $G = (V, E)$ where V is the set of sentences and $(u, v) \in E$ iff $\langle e_u, e_v \rangle > 0.3$. If G is disconnected, sequential edges $(i, i+1)$ are added to restore connectivity without perturbing the thresholded structure elsewhere. Threshold $\varepsilon = 0.3$ was selected via sensitivity analysis (Appendix A). For alternative encoders we re-derive ε to match the reference mean edge density (see Section 7).

Theoretical motivation: paths and uniformity. Valid step-by-step reasoning forms a semantic dependency chain: step $k+1$ builds on step k, so the semantic ε-proximity graph approximates a path. The Laplacian of an n-node path has eigenvalues $\lambda_k = 2(1 - \cos(\pi k/n))$ (Chung, 1997), concentrated near zero with a long tail, and therefore low spectral entropy. Two failure modes push the spectrum toward uniformity: redundancy and circular reasoning create cross-edges between distant sentences (in the limit, K_n, which has maximum spectral entropy $\log(n-1)$); incoherent jumps produce random-looking connectivity. Curvature features capture a complementary failure mode: a chain that loses contact with itself contains negative-curvature edges where the random-walk neighborhoods on the two endpoints have minimal overlap. These two failure modes are the structural correlates of the spectral and Ricci specializations we report in Section 5.

Spectral features (5). From the graph Laplacian $L = D - A$ and its eigendecomposition $L = U \Lambda U^\top$, we extract: *spectral entropy* $H = -\sum_i p_i \log p_i$ with $p_i = \lambda_i / \sum_j \lambda_j$; *Fiedler value* λ_2 (algebraic connectivity); *graph Rayleigh smoothness* (GRS) $= \mathrm{Tr}(X^\top L X)/\mathrm{Tr}(X^\top X)$ measuring how smoothly the embedding signal varies across the graph; *spectral high-frequency score* (SHS), the unweighted fraction of eigenvalues with $\lambda > \lambda_{\max}/2$; and *high-frequency energy ratio* (HFER), the eigenvalue-weighted fraction.

Ollivier–Ricci features (4). For each edge $(u, v) \in E$ we compute

$$\kappa(u, v) = 1 - \frac{W_1(m_u, m_v)}{d(u, v)},$$

where m_u is the lazy random-walk measure (laziness $\alpha = 0.5$) and W_1 is computed by linear programming. From the edge-wise distribution we summarize: *mean* κ, *std* κ, *frac-negative* (fraction of edges with $\kappa < -0.05$), and $\min \kappa$. We implemented Ollivier–Ricci from scratch in 130 lines of python after `GraphRicciCurvature` produced bus errors and `POT` failed to import on our environment. Validation against analytical values: P_6 interior edges $\kappa = 0$, boundary $\kappa = 0.5$; K_4 all edges $\kappa = 2/3$; bridge in two-triangles-joined $\kappa \approx -0.33$. All within three decimal places.

3.2 Datasets

Models. 6 LLMs spanning proprietary, open-weight, and commercial-distilled training recipes: Claude Sonnet 4 (Anthropic), gpt-oss 120B (open-weight), Llama 3.1 8B Instruct (Meta), Mistral 7B Instruct v0.3, GPT-4o-mini (OpenAI), and Gemini 2.5 Flash (Google).

Benchmarks. 3 reasoning datasets, 200 problems each: GSM8K (Cobbe et al., 2021) (grade-school arithmetic); FOLIO (Han et al., 2024) (first-order logic with True / False / Unknown labels); PrOntoQA (Saparov and He, 2023) (synthetic symbolic deduction).

Generation and filtering. Each trace is generated at temperature 0.0 on the same 200 problems per benchmark, with per-model generation settings

Model	FOLIO	GSM8K	PrOntoQA
Claude Sonnet 4	**30%**	42%	49%
gpt-oss 120B	49%	**67%**	34%
Llama 3.1 8B	53%	47%	49%
Mistral 7B	58%	45%	54%
GPT-4o-mini	44%	39%	50%
Gemini 2.5 Flash	38%	44%	**74%**

Table 1: Per-condition disagreement rate (fraction of traces caught by exactly one of spectral or Ricci) across 18 (model, domain) cells. Range 30.5% to 74.5%, mean 48.1%, median 47.9%. Bold: minimum and maximum. Per-condition n ranges from 143 to 200 (Appendix C).

Test	Stat.	p-value
Cliff's δ, numeric rate	$+\mathbf{0.29}$	9.0×10^{-10}
Cliff's δ, length-resid.	$+0.34$	9.6×10^{-13}
Domain comp., $\chi^2(2)$	50.07	1.3×10^{-11}
Model comp., $\chi^2(5)$	51.12	8.2×10^{-10}

Table 2: Test statistics for the S vs. R comparison (incorrect-only, $|S| = 345$, $|R| = 265$, pooled over 18 conditions). The conservative *per-sentence rate* is the primary headline; the length-residualized δ corroborates under explicit length control.

and prompt templates listed in Appendix E. The raw $6 \times 3 \times 200 = 3{,}600$ trace corpus is filtered by (i) a uniform model-blind exclusion of generation-truncated traces (15 dropped; rule and trace IDs in Appendix D) and (ii) removal of traces with fewer than three sentences, yielding 3,493 traces for the analyses below. Because trace length varies with model and generation settings, the modality analyses (Section 5) are reported as per-sentence rates and under length residualization, and the model-identity analysis (Section 6) reports against a trace-length baseline.

Graph sizes. After the three-sentence filter, traces span $|V| = 3$ to 92 sentences (median 16, mean 18.5), with edge counts $|E|$ from 2 to 2,427 (median 80). Graph size varies by domain: median $|V|$ is 14 for GSM8K, 18 for FOLIO, and 21 for PrOntoQA, with median $|E|$ of 61, 82, and 129 respectively. PrOntoQA traces are roughly 50% longer and twice as edge-dense as GSM8K traces; this cross-domain length variation is precisely what the per-sentence rate and length-residualized controls in Section 5 absorb.

4 Spectral and Curvature Halves Are Orthogonal

For each (model, domain) condition we threshold spectral entropy at its per-condition median and frac-negative at its per-condition median, flipping direction if anti-correlated with correctness, and count traces where exactly one of the two predictors is correct.

Table 1 reports per-condition disagreement. Across the 18 within-corpus conditions, disagreement ranges from 30.5% (Claude $\times$ FOLIO) to 74.5% (Gemini Flash $\times$ PrOntoQA), with mean 48.1% and median 47.9%. Disagreement exceeds 50% on 5 of 18 conditions. The asymmetry is structured by model: Claude prefers spectral features on FOLIO and Ricci on PrOntoQA; the two commercial-distilled models concentrate their disagreement on PrOntoQA. The non-overlap holds consistently across models with very different architectures, sizes, and training recipes, which we take as evidence that the two feature families capture genuinely different aspects of reasoning structure rather than a model-family artifact.

5 Modality Specialization

The disagreement is structured. For each (model, domain) condition we identify the *only-spectral* error group S (incorrect traces caught by spectral but not Ricci) and the *only-Ricci* error group R (caught by Ricci but not spectral), pooled across all 18 conditions. Restricting to incorrect traces (the primary analysis), $|S| = 345$ and $|R| = 265$. The all-trace pool gives $|S| = 747$ and $|R| = 940$. We test whether S and R differ on non-geometric, non-circular axes: per-sentence text properties, domain composition, and model composition.

5.1 Spectral catches number-dense errors; Ricci catches sparse symbolic errors

Table 2 summarizes the key statistics. S traces contain substantially more numeric tokens per sentence than R traces: Cliff's $\delta = +0.29$, $p = 9.0 \times 10^{-10}$. We report the per-sentence rate as the primary headline because it makes no length-extrapolation assumption. Under explicit length control (residualizing numeric-token count on sentence count and comparing residuals), the effect strengthens slightly to $\delta = +0.34$ ($p = 9.6 \times 10^{-13}$). Both measures point in the same direction: spectral features preferentially catch errors on number-dense traces.

The same direction is confirmed by domain composition (Table 3). The fraction of S traces drawn from GSM8K is 0.42 vs. 0.19 for R; the fraction of R traces drawn from PrOntoQA is 0.49 vs. 0.24

Domain	S frac	R frac
GSM8K (arithmetic)	**0.42**	0.19
FOLIO (mixed)	0.34	0.32
PrOntoQA (symbolic logic)	0.24	**0.49**

Table 3: Domain composition of S (only-spectral) and R (only-Ricci) error groups, pooled across 18 conditions. S is GSM8K-heavy, R is PrOntoQA-heavy, FOLIO splits roughly evenly.

for S ($\chi^2(2)$ $p = 1.3 \times 10^{-11}$). FOLIO (mixed in numeric content) splits roughly evenly.

5.2 One axis, two correlated measures, one independent confirmation

The numeric-token-density result and the domain-composition result are not independent. GSM8K is constitutively the most number-dense benchmark, so the per-sentence rate effect and the GSM8K-vs-PrOntoQA fraction effect are two correlated projections of the same underlying axis: reasoning type. We are reporting *one reasoning-type effect confirmed by two correlated measures*, not two independent effects.

The genuinely independent corroborating signal is model composition. S and R differ significantly on which models contributed each group's traces ($\chi^2(5)$ $p = 8.2 \times 10^{-10}$). The direction is real but less cleanly interpretable than the modality split: it reflects that different models err in structurally different proportions on arithmetic vs. logic. We treat the model-composition effect as evidence that the S/R split is not a single model's quirk; the rate and domain effects together tell us what the split is *about*.

5.3 Held-out confirmation and exact tests

Since S and R are constructed by thresholding model outputs that may themselves correlate with domain and length, the tests above could in principle be optimistically biased by using the same data to construct and to evaluate the groups. We address this with a held-out confirmatory split: within each condition we pick the per-condition spectral-entropy and frac-negative thresholds (and the direction-flip) on one random half, freeze them, and assign the held-out half's incorrect traces to S or R using only the frozen thresholds. We then run the numeric-rate and domain tests on the held-out half alone. The modality split survives: on the held-out half the per-sentence numeric-rate effect is Cliff's $\delta = +0.30$ (single split, seed 0) with the domain composition retaining direction (S more GSM8K-heavy, R more PrOntoQA-heavy). Repeating the split over 50 random seeds, the held-out numeric-rate δ is positive in 50 of 50 splits (median $+0.27$, 5th–95th percentile $[+0.13, +0.39]$) and the domain direction holds in 49 of 50; the small shrinkage from the full-data $\delta = +0.29$ is the expected cost of held-out evaluation, not a collapse of the effect. The effect is therefore not an artifact of constructing and testing the groups on the same data.

We further replace the asymptotic χ^2 approximation with exact permutation tests, shuffling the S/R labels across the pooled 610 incorrect traces. Over 10^5 permutations, no relabeling produced a domain, model, or numeric-rate statistic as extreme as observed ($p < 10^{-5}$ for all three), consistent with the asymptotic values and confirming they are not anti-conservative at this sample size.

6 Model-Characteristic Geometric Signature

If geometric structure varies systematically by model, the 9-dimensional signature should predict model identity above a length-only baseline. We test this with three feature sets evaluated under a gradient-boosted classifier (150 trees, max depth 3, learning rate 0.05), with StratifiedGroupKFold cross-validation (5 folds, groups by `(domain, problem_idx)` to prevent problem-content leakage), seed 0, and 1,000-iteration group-resampled bootstrap confidence intervals: A length-only = $\{n_{\text{sent}}, n_{\text{edges}}\}$; B geometry-9 = 5 spectral + 4 Ricci; $C = B+$ length. We additionally split B into length-correlated and length-independent (*shape*) subsets using $|\text{corr}(f, n_{\text{edges}})| \geq 0.5$ as the boundary. The 7 shape features are Fiedler value, SHS, HFER, mean κ, std κ, frac-negative κ, and $\min \kappa$; the 2 size features are spectral entropy and GRS.

Table 4 reports the result. The 9-feature signature (B) reaches 0.439 accuracy at 6-way (chance 0.167); the length-only baseline (A) reaches 0.349. The paired difference $B - A = +0.090$ with 95% bootstrap CI $[0.072, 0.108]$ comfortably excludes zero. Adding length to geometry gives a small further lift ($C - A = +0.098$, CI $[0.080, 0.115]$), so the two information sources are partly complementary. Critically, the 7 length-independent shape features alone beat the length baseline: $B_{\text{shape}} - A = +0.070$, CI $[0.052, 0.088]$. The fingerprint is not

Feature set	Accuracy	95% CI
A length-only (2 feats)	0.349	$[0.332, 0.364]$
B geometry-9 (9 feats)	**0.439**	$[0.418, 0.456]$
$C = B+$ length	0.447	$[0.428, 0.467]$
B_{shape} (7 feats)	0.419	$[0.401, 0.436]$
B_{size} (2 feats)	0.354	—
$B - A$ (paired)	$+0.090$	$[0.072, 0.108]$
$B_{\text{shape}} - A$ (paired)	$+0.070$	$[0.052, 0.088]$
$C - A$ (paired)	$+0.098$	$[0.080, 0.115]$

Table 4: 6-way model identity, chance accuracy 0.167. The 9-dimensional geometric signature beats the length-only baseline by $+0.090$ accuracy with bootstrap CI excluding zero. The 7 length-independent shape features alone beat length by $+0.070$. CIs are paired group-resampled bootstrap, 1,000 iterations.

Model	F1	Support
Claude Sonnet 4	**0.607**	600
gpt-oss 120B	0.507	511
Mistral 7B	0.413	595
Llama 3.1 8B	0.398	596
Gemini 2.5 Flash	0.387	591
GPT-4o-mini	0.342	600

Table 5: Per-model F1 under the 9-feature classifier at 6-way (chance-level F1 ≈ 0.17 on a balanced 6-class problem). Two models carry the macro signal. The two commercial-distilled models have the lowest F1, suggesting they occupy geometric territory shared with other models.

just length in disguise.

Per-model F1 (Table 5) shows the texture: Claude (0.607) and gpt-oss (0.507) carry the macro signal; the other four models cluster between 0.34 and 0.41. The two commercial-distilled models (GPT-4o-mini, Gemini 2.5 Flash) have the lowest per-class F1. The three models sharing a common generation prompt (Claude Sonnet 4, gpt-oss 120B, Llama 3.1 8B) remain mutually distinguishable (F1 0.607, 0.507, 0.398), indicating the signature reflects model-level structure beyond prompt formatting.

Absolute accuracy is modest: 0.439 at 6-way is 2.6× chance, and identifiability is uneven across the model pool. The result supports the claim that geometry carries model-characteristic information beyond trace length, not that the signature is a high-accuracy model identifier. The shape-only comparison is the cleanest statement of this: the geometric signature carries information independent of length.

	MiniLM	MPNet	BGE
D1: modality split			
Rate δ	$+0.29$	$+0.28$	$+0.18$
Residual δ	$+0.34$	$+0.35$	$+0.25$
Domain χ^2 p	1.3e-11	1.3e-12	9.0e-4
D2: 6-way model identity (GBT)			
A length-only	0.349	0.361	0.331
B geometry-9	0.439	**0.483**	0.411
B_{shape}	0.419	0.449	0.382
$B - A$	$+0.090$	$\mathbf{+0.122}$	$+0.079$
$B - A$ CI	$[0.072, 0.108]$	$[0.105, 0.138]$	$[0.062, 0.094]$

Table 6: Robustness of both findings across 3 sentence encoders at 6 models (3,493 traces). MiniLM and MPNet at $\varepsilon = 0.3$; BGE at $\varepsilon = 0.60$ (re-derived to match MiniLM mean density). All three encoders show directional D1 (positive δ on numeric-token rate and length-residualized density; significant domain χ^2) and directional D2 ($B - A$ CI excludes zero). MPNet shows the strongest effect on both axes; BGE the weakest but still significant. Shape-only minus length-only is also positive across all three encoders (Appendix B).

7 Robustness

The two findings above are computed on `all-MiniLM-L6-v2` embeddings. We test whether they survive replacing the sentence encoder with two alternatives chosen to be distinct in architecture and training: `all-mpnet-base-v2` (Reimers and Gurevych, 2019) (768-dim, MPNet pre-training) and `BAAI/bge-small-en-v1.5` (384-dim, BERT/RetroMAE plus contrastive fine-tuning; we use the model name without an originating citation as no peer-reviewed publication matches our usage).

Density-matched ε for BGE. On the 3,493-trace 6-model corpus, MiniLM at $\varepsilon = 0.3$ produces graphs with mean edge density 0.635. MPNet at $\varepsilon = 0.3$ produces density 0.677 (ratio 1.07× MiniLM), within an acceptable band, so we keep $\varepsilon = 0.3$. BGE at $\varepsilon = 0.3$ produces density 0.999 (ratio 1.57×), which is degenerate: at near-complete graphs, spectral and curvature features lose discriminative content. We re-derive ε for BGE by grid search to match MiniLM's mean density, settling on $\varepsilon = 0.60$ (mean density 0.600, gap 0.035 from the MiniLM target). All BGE results below use this re-matched threshold.

Table 6 reports both findings under all three encoders. The modality split is directional in all three; the fingerprint is directional in all three. MPNet, the encoder most architecturally distinct from

MiniLM (different pre-training objective, $2\times$ embedding dimension), shows the strongest effect on both axes. BGE is the weakest but still significant, even after the ε re-derivation that the degenerate density at $\varepsilon = 0.3$ forces.

8 Discussion

This matters for NLP evaluation because most approaches to analyzing reasoning quality require model-specific training, step-level labels, repeated sampling, or access to internal activations. The geometric signature studied here is computed from a single generated trace, with no per-step labels, no resampling, and no model-internal access, so it applies uniformly to open and closed models. This makes it usable as a low-cost diagnostic layer for comparing reasoning behavior across heterogeneous LLMs, complementing rather than replacing learned verifiers.

The geometric structure of CoT traces is a decomposable representation. Spectral and Ollivier–Ricci features detect non-overlapping subsets of errors in every (model, domain) condition we measured, and the decomposition is interpretable: spectral specializes to arithmetic / number-dense reasoning, Ricci to symbolic-logic reasoning. The same 9-feature signature carries a model-characteristic fingerprint that beats a strong length baseline ($B - A = +0.090$, CI $[0.072, 0.108]$), with the advantage carried by length-independent shape features alone. Both findings hold across three distinct sentence encoders at 6 models.

Representation, not a drop-in verifier. Geometric structure should be used to *characterize* reasoning, not as a single-boundary correctness verifier. Geometric fusion can reach high verification AUC in clean cases (0.971 5-fold CV GBT AUC on Claude $\times$ PrOntoQA, a $+0.120$ lift over the best single feature, Ricci frac-negative at 0.851), but the value of geometric structure is better understood as decomposable, error-type-specific detection than as a single monolithic verification score. A naive monolithic binary classifier trained on a small calibration-error subset can appear to score at chance (e.g. Gemini 2.5 Pro $\times$ FOLIO with $n = 26$ incorrect traces, AUC 0.485 with CI spanning 0.5 in earlier exploratory work). Such results are more parsimoniously attributed to trace-length confounds and small-sample variance than to any structural property of the signature itself, and they do not arise when the analysis is framed around modality-specialized error decomposition and length-controlled fingerprinting at the corpus scale of this paper. The appropriate quantitative targets are decomposition and length-controlled signal carriage, not aggregate verification AUC.

Connection to GNN over-squashing. The Ricci-side specialization to symbolic-logic errors is consistent with the GNN literature on negative-curvature bottlenecks (Topping et al., 2022; Nguyen et al., 2023): symbolic deduction traces that lose internal coherence develop negative-curvature edges where the random-walk neighborhoods of the two endpoint sentences have minimal overlap. The same structural signature that drives over-squashing in message-passing networks appears in CoT traces where the model loses contact with its own deduction chain.

9 Conclusion

We have shown that the 9-dimensional geometric signature of CoT traces decomposes by reasoning type (spectral $\rightarrow$ arithmetic, Ricci $\rightarrow$ symbolic-logic) and carries a model-characteristic fingerprint that beats a length baseline. Both findings are verified across three sentence encoders at 6 models with a uniform truncation-exclusion control. The decomposition is interpretable, encoder-robust, and length-independent in the part of the signal that matters. We release the 130-line python implementation of Ollivier–Ricci on graphs and the full 6-model trace corpus to encourage further analysis of geometric structure in chain-of-thought reasoning.

Limitations

Modest absolute accuracy. Per-model identifiability at 6-way is 0.439 (chance 0.167). The claim is "geometry beats length"; we do not claim that the geometric signature alone is a high-accuracy model identifier. The 0.090 advantage over length-only baseline is statistically robust (CI excludes zero with margin) but the absolute headroom is limited at this scale.

Model coverage. 6 models from 5 organizations covering proprietary, open-weight, and distilled-commercial training recipes. We do not test models trained with reasoning-specific reinforcement learning (e.g. models in the o1 / R1 family), where error distributions may differ qualitatively.

Embedding-model dependence. Both findings are verified across 3 sentence encoders (MiniLM,

MPNet, BGE) at 6 models. BGE is the weakest of the three and required ε re-derivation to avoid near-complete graphs at the default $\varepsilon = 0.3$. Encoders more distant from sentence-transformer training (e.g. instruction-tuned text embedders, multilingual encoders) have not been tested.

Benchmark coverage. Three reasoning benchmarks (GSM8K, FOLIO, PrOntoQA) covering arithmetic, mixed-logic, and synthetic symbolic-logic. Open-ended reasoning, code-as-thought, multilingual reasoning, and multi-turn dialogue reasoning are not tested.

Single-trace assumption. The signature is computed from a single CoT trace per query. We do not investigate whether geometric properties of *sampled distributions* of traces (e.g. variance across self-consistency samples) carry additional signal.

No causal claim about modality specialization. We document that the only-spectral error group is more number-dense / arithmetic-domain and the only-Ricci group is more symbolic-logic / PrOntoQA-domain. We interpret this as "spectral catches arithmetic-style failures, Ricci catches logic-style failures" but do not establish a causal mechanism. The pattern could in principle correlate with an upstream factor (problem difficulty, tokenizer behavior across domains, prompt-format sensitivity) we have not isolated.

Threshold and graph-construction choices. We fix the cosine similarity threshold at $\varepsilon = 0.3$ for MiniLM and re-derive per encoder to match mean edge density. Sensitivity analysis (Appendix A) shows the spectral signal is robust to ε in $[0.2, 0.5]$ on MiniLM. We have not exhaustively tested alternative graph constructions (k-NN, hybrid, causal-marker) for the modality-split or fingerprint findings.

References

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. 2025. Hallucination detection in LLMs using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 24354–24385.

Fan R. K. Chung. 1997. *Spectral Graph Theory*. American Mathematical Society.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, and 7 others. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Qi Wang, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. 2025. What makes a good reasoning chain? Uncovering structural patterns in long chain-of-thought reasoning. *arXiv preprint arXiv:2505.22148*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

Khang Nguyen, Nong Minh Hieu, Vinh Duc Nguyen, Nhat Ho, Stanley Osher, and Tan Minh Nguyen. 2023. Revisiting over-smoothing and over-squashing using Ollivier–Ricci curvature. In *International Conference on Machine Learning (ICML)*.

Valentin Noël. 2025. A graph signal processing framework for hallucination detection in large language models. *arXiv preprint arXiv:2510.19117*.

Valentin Noël. 2026. Geometry of reason: Spectral signatures of valid mathematical reasoning. *arXiv preprint arXiv:2601.00791*.

Yann Ollivier. 2009. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Soumadeep Saha, Akshay Chaturvedi, Saptarshi Saha, Utpal Garain, and Nicholas Asher. 2025. KisMATH: Do LLMs have knowledge of implicit structures in mathematical reasoning? *arXiv preprint arXiv:2507.11408*.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations (ICLR)*.

Xue Wen Tan, Nathaniel Tan, Galen Lee, and Stanley Kok. 2025. The shape of reasoning: Topological analysis of reasoning traces in large language models. *arXiv preprint arXiv:2510.20665*.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. 2022. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations (ICLR)*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. ArXiv preprint.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. 2025. Mapping the minds of large language models: A graph-based analysis of reasoning LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Sensitivity to graph construction

Across cosine similarity thresholds $\varepsilon \in [0.2, 0.7]$ and across five graph construction variants (chain, k-NN, cosine threshold, hybrid, causal-marker), the spectral entropy single-feature AUC on the Llama-8B GSM8K validation set is stable in the range 0.727 to 0.738. The method is not sensitive to graph construction within this regime. We use cosine threshold $\varepsilon = 0.3$ on MiniLM in the main analysis; for MPNet and BGE we use density-matching as described in Section 7.

B Robustness: supplementary detail

Shape-only-minus-length-only Δ accuracy (paired group-resampled bootstrap, 1,000 iterations, 95% CI), per encoder at 6 models: MiniLM $+0.070$ $[0.051, 0.089]$; MPNet $+0.087$ $[0.069, 0.106]$; BGE $+0.051$ $[0.032, 0.069]$. All three CIs exclude zero.

Per-model F1 under MPNet and BGE follows the same ordering as MiniLM (Claude $\rightarrow$ gpt-oss $\rightarrow$ remainder), with absolute F1 slightly higher under MPNet (reflecting MPNet's stronger B accuracy) and slightly lower under BGE.

C Full 18-condition orthogonality matrix

Per-condition counts under the median-threshold reconstruction on the 6m clean corpus (3,493 traces). "both right" is the count of traces correctly classified by both spectral and Ricci medians; "only sp" is only-spectral-right; "only ri" is only-Ricci-right; "both wrong" is neither. Per-condition n is below 200 for cells where gpt-oss 120B has empty traces (72 / 600 total, gpt-oss only; see Section 3) or where the model has truncated traces excluded under the rule in Appendix D. Compact listing only:

Model names abbreviated: Claude = Claude Sonnet 4; gpt-oss = gpt-oss 120B; Llama = Llama 3.1 8B; Mistral = Mistral 7B; Gemini Flash = Gemini 2.5 Flash.

Model	**Domain**	n	**br**	**sp**	**ri**	**bw**	**dis**
Claude	folio	200	76	34	27	63	0.31
Claude	gsm8k	200	81	20	65	34	0.42
Claude	prontoqa	200	75	29	69	27	0.49
gpt-oss	folio	168	51	40	42	35	0.49
gpt-oss	gsm8k	200	39	66	68	27	0.67
gpt-oss	prontoqa	143	73	11	37	22	0.34
Llama	folio	200	58	67	39	36	0.53
Llama	gsm8k	200	86	39	55	20	0.47
Llama	prontoqa	196	72	64	32	28	0.49
Mistral	folio	195	51	53	60	31	0.58
Mistral	gsm8k	200	64	64	26	46	0.45
Mistral	prontoqa	200	67	40	67	26	0.54
GPT-4o-mini	folio	200	68	59	29	44	0.44
GPT-4o-mini	gsm8k	200	87	13	64	36	0.39
GPT-4o-mini	prontoqa	200	67	46	54	33	0.50
Gemini Flash	folio	195	84	41	33	37	0.38
Gemini Flash	gsm8k	200	96	4	84	16	0.44
Gemini Flash	prontoqa	196	40	57	89	10	0.74

D Truncation exclusion

We adopt a uniform model-blind rule: a trace is excluded iff it is generation-truncated, detected by a truncation finish-reason (`MAX_TOKENS`, `length`, or equivalent) where the generation pipeline logged one, or by spot-confirmed mid-reasoning truncation where it did not (the four within-paper Stage-1 models were generated without a finish-reason field). A strict finish-reason-only rule was rejected as logging-blind rather than model-blind: it would exclude Gemini Flash's truncated traces but retain

equally-truncated gpt-oss and Llama traces purely because of a logging difference.

The rule excludes 15 traces from the 6-model corpus: 9 from Gemini 2.5 Flash (5 FOLIO problem indices {134, 142, 158, 173, 187} and 4 PrOntoQA {15, 60, 110, 120}, all `MAX_TOKENS`); 2 from gpt-oss 120B (FOLIO 99, PrOntoQA 168, both spot-confirmed mid-step truncations); and 4 from Llama 3.1 8B (PrOntoQA {59, 78, 101, 198}, all spot-confirmed; three are length-capped repetition loops). All 15 are logic-domain, all are marked `correct = False`, and all carry garbage final-answer fragments.

Dual report. Both D1 and D2 are run with and without the exclusion. The headline numbers are within bootstrap CI of each other: D1 length-residualized $\delta = +0.34$ (excluded) vs. $+0.31$ (included); D2 $B - A = +0.090$ (excluded) vs. $+0.097$ (included). The exclusion is hygiene: it removes mislabeled length-correlated noise from the analysis without changing the qualitative result.

Empty traces (disclosure). 72 traces (12%) from gpt-oss 120B and 5 (1%) from Mistral 7B have blank `cot_trace` and produce no graph. These were already outside the geometric analysis (no graph $\rightarrow$ no 9 features) and are not affected by the truncation rule.

E Implementation Details

Data and code availability. All data, code, and configuration are available at https://anonymous.4open.science/r/geometry-cot-release-61F5, including the trace corpus, the feature extraction pipeline, and scripts that reproduce every number reported in this paper.

E.1 Generation settings

All traces were generated at temperature 0.0 on the same 200 problems per benchmark. Generation budgets were 1,024 tokens for Claude Sonnet 4, gpt-oss 120B, Llama 3.1 8B, and Mistral 7B, and 8,192 tokens for GPT-4o-mini and Gemini 2.5 Flash. Each model was prompted for step-by-step reasoning through its provider interface. Claude Sonnet 4, gpt-oss 120B, and Llama 3.1 8B used a system prompt requesting numbered reasoning with the final answer on a line after `####`, and a separate logic phrasing directing a True/False/Unknown answer for FOLIO and PrOntoQA. Mistral 7B used an instruction-style template (`[INST] ... [/INST]`) instructing the model to reason step by step and give its final answer after `####`. GPT-4o-mini and Gemini 2.5 Flash used per-benchmark instructions requesting a final answer in "The answer is *N*" form (numeric for GSM8K, True/False/Unknown for FOLIO, True/False for PrOntoQA). Verbatim prompt strings for every model are in the released repository.

E.2 Answer extraction and grading

Final answers were extracted by regular expression, prioritizing an explicit answer delimiter and falling back to a search over the trace tail. Grading used exact match after normalization of case, whitespace, and thousands separators; numeric answers (GSM8K) were compared with a tolerance of 10^{-6}, and FOLIO and PrOntoQA answers by case-insensitive string match. The exact extraction and grading code is in the released repository.

E.3 Sentence segmentation

Each trace is split into sentences with the regular expression `(?<=[.!?])\s+(?=[A-Z])|\n+`, which splits on sentence-final punctuation followed by whitespace and a capitalized token, or on line breaks. Fragments of two or fewer characters are discarded, and traces with fewer than three sentences are removed from the corpus.

E.4 Geometric features

Each sentence is embedded with `all-MiniLM-L6-v2` (384-dimensional, L_2-normalized), giving a signal matrix $X \in \mathbb{R}^{n \times 384}$ for an n-sentence trace. The combinatorial Laplacian $L = D - A$ is formed from the semantic ε-proximity graph, and its eigenvalues $\{\lambda_k\}$ are taken in ascending order with $p_k = \lambda_k / \sum_j \lambda_j$. The five spectral features are: spectral entropy $H = -\sum_k p_k \log p_k$; the Fiedler value λ_2; graph Rayleigh smoothness $\mathrm{GRS} = \mathrm{Tr}(X^\top L X)/\mathrm{Tr}(X^\top X)$; the spectral high-frequency score (SHS), the unweighted fraction of eigenvalues with $\lambda_k > \lambda_{\max}/2$; and the high-frequency energy ratio (HFER), the eigenvalue-weighted fraction $\sum_{\lambda_k > \lambda_{\max}/2} \lambda_k / \sum_k \lambda_k$.

The four Ollivier–Ricci features summarize the edge-curvature distribution. For an edge (u, v), $\kappa(u, v) = 1 - W_1(m_u, m_v)/d(u, v)$, where m_u is the lazy random-walk measure with laziness

$\alpha = 0.5$, d is the graph hop distance, and the Wasserstein-1 distance W_1 is computed by linear programming (`scipy.optimize.linprog`, method `highs`). From the per-edge curvatures we take the mean, standard deviation, fraction below -0.05 (frac-negative), and minimum. The implementation was validated against analytical curvature values on P_6 (interior edges $\kappa = 0$, boundary 0.5), K_4 (all edges $2/3$), and a two-triangle bridge ($\kappa \approx -0.33$), matching to three decimal places.

E.5 Modality groups and statistical tests

Within each (model, domain) condition, spectral entropy and frac-negative are each thresholded at their per-condition median, with a feature's direction reversed if its AUC against correctness falls below 0.5. An incorrect trace is assigned to the only-spectral group S if the spectral predictor is correct and the Ricci predictor is not, and to the only-Ricci group R in the reverse case. Cliff's δ is computed from the Mann–Whitney U statistic (two-sided). Length residualization regresses the per-trace count on sentence count by ordinary least squares, pooled over $S \cup R$, with Cliff's δ recomputed on the residuals. Domain and model composition are compared by Pearson χ^2 on the 2×3 and 2×6 contingency tables.

The held-out confirmation splits each condition into two stratified halves, selects the per-condition thresholds and directions on one half, and assigns the other half's traces using those frozen thresholds; the S/R tests are then run on the held-out half. This is repeated over 50 random splits, with the two halves exchanged symmetrically. The permutation tests shuffle the S/R label across the pooled incorrect traces over 10^5 draws, reporting the fraction of shuffles with a statistic at least as extreme as observed.

E.6 Model-identity classifier

The model-identity classifier is a gradient-boosted decision tree (150 trees, maximum depth 3, learning rate 0.05) evaluated under StratifiedGroupKFold cross-validation (5 folds, grouped by (domain, problem) to prevent problem-content leakage), with out-of-fold predictions pooled across folds. Confidence intervals are paired group-resampled bootstrap (1,000 iterations, resampling groups). A feature is classified as size-correlated if $|\mathrm{corr}(f, n_{\text{edges}})| \geq 0.5$ and as shape otherwise, yielding the 7 shape features (Fiedler value, SHS, HFER, mean κ, std κ, frac-negative, min κ) and 2 size features (spectral entropy, GRS) used in Section 6.

E.7 Models

In terms of the models, Claude Sonnet 4 (`claude-sonnet-4-20250514`) was queried through the Anthropic API. gpt-oss 120B and Llama 3.1 8B Instruct were queried through the Cerebras inference API. Mistral 7B Instruct v0.3 (`mistralai/Mistral-7B-Instruct-v0.3`) was served locally with 4-bit NF4 quantization. GPT-4o-mini was queried through the OpenAI API and Gemini 2.5 Flash through Google AI Studio.