Sparse Spectral Signatures of Reasoning: Model-Agnostic Verification via Sentence-Level Graph Signals

Arjun Balaji
Columbia University
ab6136@columbia.edu

Abstract

Recent work has shown that spectral properties of internal attention graphs can distinguish valid from invalid mathematical reasoning in LLMs. However, attention-based methods require access to model weights, excluding closed-source models and production deployments. We investigate whether analogous spectral signatures exist in *external* sentence-level semantic graphs constructed solely from chain-of-thought text. We construct cosine-similarity threshold graphs over sentence embeddings and compute spectral metrics from the graph Laplacian—requiring only black-box text output. Across 2,400 traces spanning three reasoning domains (mathematical, first-order logic, deductive) and four model architectures—including the closed-source Claude Sonnet 4—we find that spectral metrics reliably discriminate correct from incorrect reasoning, with 9 of 12 domain-model conditions significant at $p<0.05$ (AUC up to 0.77). Spectral features add up to +14.9% AUC over text-level baselines, with the largest gains when baselines are weakest—demonstrating that spectral analysis captures structural reasoning properties orthogonal to surface text quality.

1 Introduction

Large language models produce fluent chain-of-thought (CoT) reasoning (Wei et al., 2022), yet their outputs are frequently incorrect—even when the surface text appears convincing. Verifying reasoning quality is a critical challenge: process reward models require expensive step-level annotations (Lightman et al., 2024), self-consistency relies on sampling agreement (Wang et al., 2023), and LLM-as-judge approaches inherit the biases of the evaluating model.

A promising recent direction treats reasoning as a *graph signal processing* problem. Noël (2026) demonstrated that spectral properties of token-level attention graphs—algebraic connectivity, high-frequency energy ratio, spectral entropy—reliably distinguish valid from invalid mathematical proofs, achieving up to 95.6% accuracy with effect sizes of d=3.30. However, their method requires access to internal attention weights, limiting applicability to open-weight models and excluding closed-source APIs (GPT-4, Claude, Gemini) and production deployments where attention extraction is costly.

In this work, we ask: **can spectral signatures of reasoning validity be recovered from external text alone?** We construct sentence-level semantic graphs from CoT traces using sentence embeddings and cosine similarity, compute spectral metrics from the graph Laplacian, and test whether these metrics discriminate correct from incorrect reasoning—without any access to model internals.

Our key finding is that this works, and we provide a theoretical account of *why*: valid reasoning forms a semantic dependency chain with a characteristic, non-uniform eigenvalue distribution, while invalid reasoning disrupts this structure, producing higher spectral entropy. This spectral signature generalizes across mathematical, logical, and deductive domains and across four architectures including the closed-source Claude Sonnet 4 (AUC up to 0.77, effect sizes $|d|$ up to 0.92).

Contributions.

- We demonstrate that spectral properties of external sentence-level graphs discriminate correct from incorrect reasoning across three domains and four models—including the closed-source Claude Sonnet 4—with 9/12 conditions at $p<0.05$, establishing **external spectral analysis** as a viable, model-agnostic reasoning verification method.
- We provide a **theoretical account**: valid reasoning forms a dependency chain with a structured, non-uniform eigenvalue distribution; invalid reasoning disrupts this chain, producing higher spectral entropy. This explains why spectral entropy is the most consistent metric across all conditions.
- We show that spectral features provide **complementary signal beyond text-level baselines** (up to +14.9% AUC), with the largest gains precisely when text-level features are least discriminative.

2 Related Work

Spectral analysis of LLM reasoning. Noël (2026) introduced spectral diagnostics on internal attention graphs, showing that the Fiedler value, high-frequency energy ratio (HFER), and spectral entropy of token-level attention matrices 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*. Our work differs fundamentally: we construct graphs from external text output alone, enabling analysis of any LLM including closed-source APIs.

External CoT structure analysis. Xiong et al. (2025) build directed reasoning graphs from CoT traces and extract structural features (branching, convergence, exploration density) that correlate with accuracy. Jiang et al. (2025) convert deliberative long CoTs—with explicit backtracking and exploration—into hierarchical trees and apply supervised GNNs to extract structural patterns (branching, verification steps) that predict correctness. Unlike our approach, their method requires deliberative long-CoT traces with exploratory structure and trains a supervised model; our spectral method is training-free and operates on any single CoT trace. More broadly, both approaches operate externally but use fundamentally different analyses: their node-level structural features capture local tree properties, while our spectral decomposition of the graph Laplacian captures global coherence.

CoT verification. Self-consistency (Wang et al., 2023) uses majority voting over sampled paths. Process reward models (Lightman et al., 2024) train step-level verifiers. Graph-of-Thoughts (Besta et al., 2024) uses graph structure for prompting. Our method is *training-free* and requires only a single trace, complementing these approaches.

Graph signal processing. GSP (Shuman et al., 2013) extends classical signal processing to irregular graph domains via the graph Laplacian. The Fiedler value (Fiedler, 1973) measures algebraic connectivity. Fettal et al. (2024) applied graph smoothing on sentence embedding graphs for document classification. We bring these tools to reasoning verification.

3 Method

Our pipeline operates entirely on black-box text output. Given a CoT trace, we: (1) split into sentences, (2) embed each sentence, (3) construct a semantic similarity graph, (4) compute the graph Laplacian, and (5) extract spectral metrics. Before defining the formal machinery, we describe the core intuition.

3.1 Theoretical Foundation: Reasoning as a Structured Graph

We ground our approach in a connection between the structure of valid reasoning and the spectral properties of graphs.

Valid reasoning as a path graph. Consider a correct n-step proof. Each step follows logically from the previous: step 2 builds on step 1, step 3 on step 2, and so on. When we embed these sentences and measure their pairwise similarity, this produces a graph that approximates a *path*—each sentence is most similar to its neighbors and progressively less similar to distant steps. The Laplacian of a path graph on n nodes has eigenvalues $\lambda_k = 2(1 - \cos(\pi k/n))$ for $k = 0, \ldots, n-1$ (Chung, 1997). This distribution is highly non-uniform: eigenvalues are concentrated near zero, with a long tail. The resulting spectral entropy $H_{\text{path}} = -\sum_i p_i \log p_i$ is low.

How invalid reasoning disrupts this structure. Two common failure modes push the spectrum toward uniformity:

- *Redundancy and circular reasoning* create cross-edges between distant sentences that are semantically similar despite being logically unrelated. In the extreme, if every sentence is equally similar to every other, the graph approaches a *complete graph* K_n, whose Laplacian has eigenvalues $\{0, n, n, \ldots, n\}$. The $n-1$ non-zero eigenvalues are all equal; when normalized as a probability distribution they are uniform, yielding maximum spectral entropy $\log(n-1)$—far higher than the path graph's concentrated spectrum.
- *Incoherent jumps and missing steps* break the sequential similarity pattern, producing random-looking connectivity. An Erdős–Rényi random graph $G(n, p)$ has eigenvalues that converge to a more symmetric, diffuse distribution as n grows—also yielding higher entropy than the structured path spectrum.

Both failure modes therefore produce more uniform eigenvalue distributions than the path-like structure of valid reasoning. Our central hypothesis is: **spectral entropy of sentence-level semantic graphs measures how far a reasoning trace deviates from the ordered dependency chain that characterizes valid deduction.** We test this hypothesis empirically in Section 4.3.

3.2 Sentence-Level Semantic Graphs

Given a CoT trace, we split it into n sentences $\{s_1, \ldots, s_n\}$ using regex-based boundary detection tuned for reasoning text (handling numbered steps, equations, etc.). Each sentence is embedded using Sentence-BERT (Reimers & Gurevych, 2019) to obtain vectors $\mathbf{e}_i \in \mathbb{R}^d$. We then construct an adjacency matrix $A \in \mathbb{R}^{n \times n}$ based on cosine similarity.

Our primary construction is the **ε-threshold graph**: $A_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j)$ if $\cos(\mathbf{e}_i, \mathbf{e}_j) > \varepsilon$, and 0 otherwise, with $A_{ii} = 0$. Low ε produces sparse graphs (only the strongest semantic connections), while high ε produces denser graphs. We set $\varepsilon = 0.3$ based on sensitivity analysis (Section 4.4).

3.3 Spectral Metrics from the Graph Laplacian

The combinatorial graph Laplacian is $L = D - A$, where $D = \text{diag}(\sum_j A_{ij})$. Let $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of L, and $U = [u_1, \ldots, u_n]$ the corresponding eigenvectors. Using the sentence embeddings $X \in \mathbb{R}^{n \times d}$ as a signal on the graph, we compute five metrics, each capturing a different aspect of graph structure:

Spectral Entropy $H = -\sum_i p_i \log p_i$ where $p_i = \lambda_i / \sum_j \lambda_j$. *What it captures:* How structured vs. uniform the eigenvalue distribution is. Structured graphs (like paths) have concentrated spectra and low entropy; disrupted graphs have diffuse spectra and high entropy. This is our primary metric based on the theoretical account above.

Spectral High-frequency Score (SHS) $\sum_{k>n/2} \|U_k^\top X\|^2 / \sum_k \|U_k^\top X\|^2$. *What it captures:* The unweighted fraction of signal energy in high-frequency modes. Unlike HFER (which weights by eigenvalue), SHS measures raw energy concentration regardless of mode magnitude.

Graph Rayleigh Smoothness (GRS) $\mathrm{Tr}(X^\top LX)/\mathrm{Tr}(X^\top X)$. *What it captures:* How smoothly the embedding signal varies across connected sentences. Low GRS means adjacent sentences in the graph have similar embeddings—consistent with a coherent reasoning flow.

Fiedler value λ_2: the second smallest eigenvalue of L (Fiedler, 1973). *What it captures:* Algebraic connectivity—how tightly connected the graph is overall. Invalid reasoning with redundant sentences produces higher connectivity.

High-Frequency Energy Ratio (HFER) $\sum_{k>n/2} \lambda_k \|U_k^\top X\|^2 / \sum_k \lambda_k \|U_k^\top X\|^2$. *What it captures:* The fraction of embedding signal energy in rapidly varying (high-frequency) graph modes. Logical inconsistencies create abrupt embedding changes that concentrate energy in high-frequency components.

3.4 Graph Construction Variants

Unlike internal attention graphs (which are fixed by the model architecture), external sentence graphs can be *designed*. We compare five constructions, each encoding a different theory of what makes reasoning coherent:

Construction	Edges	Structural Assumption
Chain	Adjacent sentences ($\|i-j\|=1$)	Sequential coherence
k-NN	$k=3$ most similar by cosine	Semantic clustering
ε-threshold	Cosine $> \varepsilon$	Semantic neighborhoods
Hybrid	Chain + threshold	Sequential + semantic
Causal-marker	Chain + "therefore"/"because"	Logical connectives

The causal-marker graph is particularly motivated for logical reasoning: when a sentence contains a discourse connective like "therefore" or "because," it is connected back to the most semantically similar preceding sentence, encoding explicit logical dependencies.

4 Experiments

4.1 Setup

Datasets. We evaluate on three reasoning domains with n=200 per model per dataset: **GSM8K** (Cobbe et al., 2021) (mathematical), **FOLIO** (first-order logic NLI), and **PrOntoQA** (ontological deduction). The latter two directly test logical reasoning capabilities.

Models. We test four architectures: **Mistral-7B-Instruct** (7B, sliding window attention), **Llama-3.1-8B** (8B), **GPT-OSS-120B** (120B), and **Claude Sonnet 4** (closed-source frontier model; `claude-sonnet-4-20250514`). All use identical system prompts and greedy decoding. Total: 2,400 traces.

Baselines. We compare spectral metrics against text-level features: sentence count, token count, cosine drift, embedding variance, and mean pairwise similarity.

Embedding. All-MiniLM-L6-v2 (384-dim).

4.2 Discriminability

Table 1 shows the discriminative power of each spectral metric on 200 GSM8K traces (Llama-3.1-8B). **Spectral entropy** is the strongest single metric (d=−0.800, p=1.7×10^{-7},

Table 1: Discriminability of spectral metrics on 200 GSM8K traces (Llama-3.1-8B, threshold ε=0.3).

Metric	Cohen's d	p-value	AUC-ROC
GRS	-0.590	1.9×10^{-6}	0.718
SHS	$+0.388$	2.1×10^{-2}	0.606
Spectral Entropy	$\mathbf{-0.800}$	$\mathbf{1.7 \times 10^{-7}}$	**0.740**
Fiedler	-0.628	1.3×10^{-5}	0.694
HFER	$+0.668$	3.9×10^{-5}	0.689

Table 2: Cross-domain and cross-model results (3 domains × 4 models, all n=200). Best spectral metric per condition. Bold: $p < 0.01$; italic: $p < 0.05$.

Domain	Model	Acc.	Best Metric	d	p	AUC
GSM8K	Llama-8B	72%	Sp. Entropy	$\mathbf{-0.80}$	$\mathbf{2 \times 10^{-7}}$	**0.74**
GSM8K	Mistral-7B	39%	GRS	$\mathbf{-0.67}$	$\mathbf{4 \times 10^{-6}}$	**0.69**
GSM8K	GPT-OSS-120B	48%	Fiedler	$\mathbf{-0.61}$	$\mathbf{2 \times 10^{-8}}$	**0.73**
GSM8K	Claude Sonnet 4	82%	Fiedler	$\mathbf{-0.55}$	$\mathbf{5 \times 10^{-7}}$	**0.77**
FOLIO	Llama-8B	42%	Sp. Entropy	$\mathbf{-0.47}$	$\mathbf{4 \times 10^{-4}}$	**0.65**
FOLIO	Mistral-7B	46%	Fiedler	$+0.19$	0.061	0.58
FOLIO	GPT-OSS-120B	62%	GRS	$+0.17$	0.21	0.56
FOLIO	Claude Sonnet 4	76%	GRS	*-0.33*	*0.035*	*0.60*
PrOntoQA	Llama-8B	39%	Sp. Entropy	$\mathbf{-0.92}$	$\mathbf{3 \times 10^{-9}}$	**0.75**
PrOntoQA	Mistral-7B	40%	HFER	$+0.23$	0.13	0.56
PrOntoQA	GPT-OSS-120B	12%	GRS	$\mathbf{+1.02}$	$\mathbf{2 \times 10^{-3}}$	**0.70**
PrOntoQA	Claude Sonnet 4	73%	HFER	*-0.39*	*0.023*	*0.61*

AUC=0.740). The negative effect size indicates that incorrect traces have *higher* spectral entropy: their eigenvalue distributions are more diffuse, reflecting less structured semantic organization.

4.3 Cross-Domain and Cross-Model Generalization

Table 2 presents results across all twelve domain-model combinations (uniform n=200). **7 of 12 conditions are significant at** p<0.01, with 9 of 12 at p<0.05. GSM8K is the strongest domain, with all four models significant at $p<10^{-5}$. Notably, **Claude Sonnet 4**—a frontier closed-source model with 82% accuracy—shows strong spectral discrimination (d=−0.55, AUC=0.77, $p=5 \times 10^{-7}$). The strongest individual result is PrOntoQA/Llama (d=−0.919, AUC=0.750).

An instructive anomaly: GPT-OSS on PrOntoQA. The GPT-OSS/PrOntoQA result (d=+1.02) shows a *positive* effect size—correct traces have *higher* GRS than incorrect ones, the opposite of the typical pattern. This model achieves only 12% accuracy on PrOntoQA; its default behavior is to produce short, low-effort incorrect traces (mean 5.6 sentences). The rare correct traces are longer (7.9 sentences) and more elaborate—the model "tried harder." This reversal does not contradict our theory; rather, it reveals a distinct failure regime. When a model's primary failure mode is *insufficient reasoning* rather than *flawed reasoning*, the spectral signal inverts: more structural complexity indicates effort, not confusion. The method still discriminates effectively (AUC=0.70, p=0.002), but the direction of the signal depends on the model's failure mode.

4.4 Graph Construction Ablation

Table 3 compares five graph constructions on 200 GSM8K traces (Llama-3.1-8B). All constructions achieve AUC $\geq$ 0.736, with spectral entropy as the best metric for all. Sparser

Table 3: Graph construction ablation on 200 GSM8K traces (Llama-3.1-8B). All constructions use spectral entropy as the best metric.

Graph Type	AUC-ROC
Chain	0.737
k-NN (k=3)	0.739
Threshold (ε=0.3)	0.740
Hybrid (ε=0.5)	0.736
Causal-Marker	0.741
Threshold (ε=0.2)	**0.744**

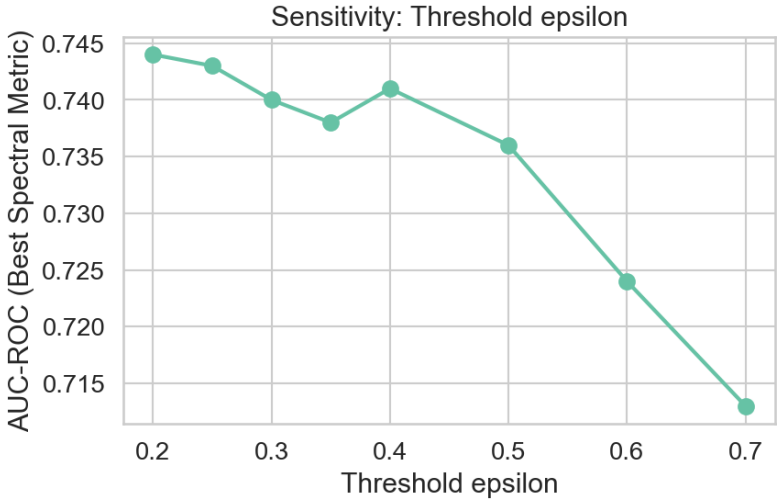

Figure 1: AUC-ROC vs. threshold ε on 200 GSM8K traces (Llama-3.1-8B). Sparser graphs (lower ε) produce modestly stronger discrimination (0.744 vs. 0.713).

threshold graphs (ε=0.2) achieve the highest AUC (0.744), though the differences are modest.

The sensitivity analysis (Figure 1) confirms a consistent trend: sparser graphs produce stronger discrimination (AUC 0.744 at ε=0.2 vs. 0.713 at ε=0.7), though the effect is gradual rather than dramatic. This suggests that valid reasoning has somewhat sparser semantic neighborhoods, but the spectral signal is robust to graph construction choices.

4.5 Spectral Metrics vs. Text-Level Baselines

Table 4 compares text-level baselines against spectral features via 5-fold cross-validated logistic regression on GSM8K. Spectral features provide the most complementary value when text-level baselines are weakest. For GPT-OSS-120B, where baselines achieve only 0.578 AUC, spectral features add +14.9%. Claude Sonnet 4 (baselines 0.638) gains +10.9%, and Llama-3.1-8B (0.696) gains +8.6%. For Mistral-7B, where baselines are already strong (0.761), spectral features provide no additional value (−1.8%). This complementarity pattern—confirmed across four models including the closed-source Claude Sonnet 4—suggests that spectral analysis captures structural reasoning properties orthogonal to surface text quality.

4.6 Eigenvalue Spectrum Analysis

Figure 2 shows the average Laplacian eigenvalue spectrum across 145 correct and 55 incorrect GSM8K traces (Llama-3.1-8B), directly illustrating the theoretical prediction from Section 3.1. Correct traces exhibit a more concentrated spectrum (mean spectral entropy

Table 4: Logistic regression ablation: text-level baselines vs. spectral features (5-fold CV AUC, GSM8K). Δ = Combined − Baselines. Combined can be lower than Spectral-only due to overfitting with more features at small n.

Model	Baselines	Spectral	Combined	Δ
Mistral-7B	0.761	0.688	0.743	−0.018
Llama-3.1-8B	0.696	0.772	0.782	+0.086
Claude Sonnet 4	0.638	0.706	0.747	+0.109
GPT-OSS-120B	0.578	0.732	0.728	+0.149

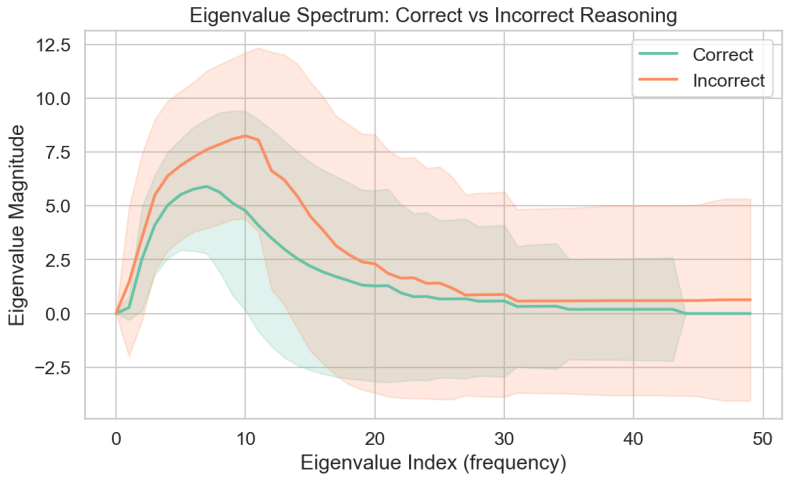

Figure 2: Average Laplacian eigenvalue spectrum for correct vs. incorrect reasoning traces (200 GSM8K, Llama-3.1-8B). Correct traces have more concentrated spectra.

2.18 ± 0.50) compared to incorrect traces (2.55 ± 0.36), consistent with the path-like vs. disrupted structure predicted by our theoretical account.

Controlling for trace length. Spectral entropy correlates strongly with trace length ($r = 0.92$), raising the concern that we are simply detecting that incorrect traces are longer. Indeed, sentence count alone achieves AUC=0.737, compared to spectral entropy's 0.740—a negligible single-metric difference. However, the full spectral feature set achieves AUC=0.772 (Table 4), substantially above sentence count's 0.737. This gap persists after controlling for length: comparing entropy between correct and incorrect traces *within the same length bin* yields $p < 10^{-4}$. Crucially, the full baseline set in Table 4 *already includes* sentence count, token count, and three embedding-level features—yet spectral features still add +8.6% AUC on top. The *combination* of spectral metrics captures structural information that no subset of length and embedding features provides.

4.7 Cross-Domain Spectral Entropy

Table 5 shows mean spectral entropy by domain and correctness for Llama-3.1-8B. The entropy gap between correct and incorrect traces is consistent across all three domains, confirming that the pattern is not GSM8K-specific. The absolute entropy values increase from GSM8K to PrOntoQA, reflecting longer traces for ontological reasoning. Note that absolute entropy is length-influenced (as discussed in Section 4.6), but the *within-domain gap* between correct and incorrect—which is the discriminative signal—is consistent across domains and persists after length-matching.

Table 5: Mean spectral entropy (± std) by domain and correctness (Llama-3.1-8B, threshold ε=0.3). Incorrect traces consistently have higher entropy across all domains.

Domain	Correct	Incorrect	Δ
GSM8K	2.18 ± 0.50	2.55 ± 0.36	+0.37
FOLIO	2.68 ± 0.36	2.87 ± 0.44	+0.19
PrOntoQA	2.96 ± 0.34	3.33 ± 0.45	+0.37

4.8 Qualitative Example

To make the spectral signal concrete, consider two GSM8K traces of similar length (6 and 10 sentences):

Correct trace (spectral entropy = 0.96). "Johnny took his allowance of \$20 and added \$10... This sum tripled in a year... so \$30 × 3 = \$90." Each sentence introduces one new fact and builds directly on the previous, producing a clean path-like graph with concentrated eigenvalues.

Incorrect trace (spectral entropy = 2.14). "To find how much money Greta has left, we need to calculate... 50% of \$2400 = \$1200... uses 20% of her remaining pay for her car..." Despite being well-written, the trace contains multiple sub-problems with overlapping vocabulary (repeated references to amounts, percentages, and "remaining pay"), creating spurious semantic cross-edges between sentences that discuss different calculation steps. This elevates spectral entropy by pushing the graph away from the clean path structure—even though the surface text reads coherently.

5 Discussion

Theoretical account: empirical support. The hypothesis from Section 3.1—that spectral entropy measures deviation from valid reasoning's dependency chain—is supported by three findings: spectral entropy is the best metric across all graph constructions (Table 3), the signal generalizes across reasoning domains (Table 2), and the predicted failure mode (structurally uniform traces) is confirmed for Mistral-7B on logic tasks, where correct and incorrect traces differ by only 0.2–0.8 sentences compared to 4.2–10.8 for Llama.

When do spectral features help most? The ablation reveals a clear complementarity pattern across all four models: spectral features add +14.9% AUC for GPT-OSS-120B (where baselines are weakest at 0.578), +10.9% for Claude Sonnet 4 (0.638), +8.6% for Llama-8B (0.696), but −1.8% for Mistral-7B (where baselines already reach 0.761). This suggests that spectral graph structure captures reasoning properties *orthogonal* to surface text quality. When simple features like length and coherence already discriminate well (Mistral), spectral analysis is redundant. When they do not (GPT-OSS, Claude), spectral topology becomes essential.

Graph construction robustness. Unlike prior work on internal attention graphs where graph structure is fixed by architecture, external graphs can be designed. We find that all five constructions produce significant discrimination (AUC 0.736–0.744), with sparser graphs modestly better. This robustness is encouraging: practitioners need not carefully tune graph construction to benefit from spectral analysis.

Limitations. Our AUC values (0.56–0.77) are substantially lower than the 85–96% accuracy achieved by Noël (2026) using internal attention graphs. External analysis necessarily loses information. The signal is strongest for mathematical reasoning (all models at $p<10^{-5}$) but weaker for first-order logic (FOLIO: 2/4 models significant) and deductive reasoning (PrOntoQA: 3/4), suggesting that spectral discrimination depends on the model producing structurally diverse correct and incorrect traces. Mistral-7B shows no significant signal on

FOLIO or PrOntoQA; inspection reveals that Mistral's correct and incorrect traces have similar lengths and embedding variance on these tasks, consistent with our theory's prediction that structurally uniform outputs will lack spectral signal. The graph ablation (Table 3) and baseline comparison (Table 4) are conducted only on GSM8K/Llama; whether these findings transfer to other domains remains to be verified. Finally, at 82% accuracy, Claude Sonnet 4's GSM8K condition has only 35 incorrect traces, limiting statistical power for that cell despite the strong p-value.

When to use external spectral analysis. This method is most valuable when: (1) model internals are unavailable—as demonstrated with Claude Sonnet 4, (2) surface text quality does not indicate reasoning quality, and (3) a training-free, single-trace quality signal is needed.

6 Conclusion

We have shown that spectral properties of sentence-level semantic graphs constructed from chain-of-thought text provide a reliable, model-agnostic signal for reasoning validity. Spectral entropy emerged as the most consistent discriminator across three reasoning domains and four model architectures—including the closed-source Claude Sonnet 4. Our finding that spectral analysis provides the most value precisely when text-level baselines are weakest suggests that external spectral methods will become increasingly important as models produce ever more polished text.

Future work could explore richer embedding models, integration with process reward models as a complementary verification signal, and extension to multi-turn dialogue reasoning.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Nyczyk, Ronny Muller, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of AAAI*, 2024.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Hallucination detection in LLMs using spectral features of attention maps. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24354–24385, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1239. URL https://aclanthology.org/2025.emnlp-main.1239/.

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

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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Chakib Fettal et al. More discriminative sentence embeddings via semantic graph smoothing. *arXiv preprint arXiv:2402.12890*, 2024.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23 (2):298–305, 1973.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Qi Wang, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2505.22148.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of ICLR*, 2024.

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

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

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, 2019.

David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of ICLR*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 2022.

Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. Mapping the minds of large language models: A graph-based analysis of reasoning llms. In *Proceedings of EMNLP*, 2025.