GLUING LOCAL CONTEXTS INTO GLOBAL MEANING: A SHEAF-THEORETIC DECOMPOSITION OF TRANSFORMER REPRESENTATIONS

Bryce Grant **Peng Wang**
Case Western Reserve University
{bag100, pxw206}@case.edu

ABSTRACT

We decompose transformer activations into content-stable (H^0) and context-dependent (H^1) subspaces using sheaf cohomology. A cellular sheaf built over paraphrase graphs yields a Laplacian whose spectral structure separates phrasing-invariant directions from maximally varying ones, requiring no concept labels or supervised training. Across five models (124M–13B parameters), H^1 dimensions exert 3.5–26.5× greater causal influence on model output than variance-matched controls (Cohen's d = 2.3–14.3), H^0 retrieves facts at 60–68% accuracy using only 20 dimensions, and the two subspaces produce opposite effects under ablation. The decomposition also reveals architecture-dependent fragility: Llama-2-7B collapses under random perturbation (4.2% fact preservation) while all directed methods preserve facts at 12–14% ($p < 10^{-10}$, n=1000); with architecture-specific restriction maps this gap widens to 31.0% vs. 4.2% ($p < 10^{-50}$). Robust models tolerate both perturbation types. Project page: https://cwru-aism.github.io/gluing-lc-page/

1 INTRODUCTION

Transformer representations mix what a sentence means with how it is phrased. The activation for "The Eiffel Tower is in Paris" differs from that for "Paris is home to the Eiffel Tower" even though both express the same fact. Sparse autoencoders decompose activations into individual features (Bricken et al., 2023; Cunningham et al., 2024), probes test linear decodability (Belinkov, 2022), and similarity methods compare activation spaces (Kornblith et al., 2019; Raghu et al., 2017). None of these directly measures *cross-context consistency*, the degree to which a feature preserves its value when phrasing changes and meaning holds fixed.

Sheaf cohomology provides a natural framework for this question. A cellular sheaf (Curry, 2014; Hansen & Ghrist, 2019) assigns vector spaces to nodes and edges of a graph, linked by linear restriction maps. We build a sheaf over paraphrase graphs: nodes carry hidden-state vectors and edges connect paraphrase pairs. The coboundary operator δ^0 measures disagreement between paired representations, and spectral decomposition of the sheaf Laplacian $L_{\mathcal{F}} = (\delta^0)^\top \delta^0$ splits the representation space into two complementary subspaces. The kernel $H^0 = \ker(L_{\mathcal{F}})$ captures directions along which paired representations agree, encoding phrasing-invariant content. The top eigenspace captures directions of maximal disagreement, encoding context-dependent variation such as syntax and surface form. We call these "H^1 directions" (Remark 1 clarifies the relationship to cohomological H^1).

The decomposition requires only paraphrase pairs, not concept labels or attribute supervision. It operates in a learned low-dimensional edge space via a restriction map $P \in \mathbb{R}^{d\times k}$ (k=128 $\ll d$), and the choice of this map (PCA, CCA, or contrastive) determines which aspects of variation the decomposition captures.

Contributions. (1) A sheaf-theoretic decomposition of transformer representations into content-stable (H^0) and context-dependent (H^1) subspaces, validated across five models (124M–13B parameters). (2) Functional characterization: H^1 exerts 3.5–26.5× greater causal influence than variance-matched controls (Cohen's d = 2.3–14.3), H^0 retrieves facts at 60–68% accuracy with

20 dimensions, and ablation of the two subspaces produces opposite effects on generation. (3) An architecture fragility diagnostic: Llama-2-7B collapses under random perturbation (4.2% fact preservation) while directed methods preserve 3× more facts (12–14%, $p < 10^{-10}$, n=1000); robust models tolerate both perturbation types.

2 Related Work

Sheaves in machine learning. Hansen and Ghrist (Hansen & Ghrist, 2019) introduced the sheaf Laplacian as a generalization of the graph Laplacian incorporating vector-valued data on nodes and edges. Subsequent work applied sheaf neural networks to heterophilic graph learning (Bodnar et al., 2022; Barbero et al., 2022) and to analyzing local-global model fit (Kvinge et al., 2021). Recent applications include predictive coding (Seely, 2025) and learned restriction maps (Di Nino et al., 2025). We apply sheaf cohomology to decompose transformer representations into content-stable and context-dependent subspaces.

Interpretability and steering. Sparse autoencoders decompose activations into overcomplete feature dictionaries (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024). Circuit analysis identifies subgraphs responsible for specific computations (Conmy et al., 2023; Wang et al., 2023). Probing classifiers test linear decodability (Belinkov, 2022). Concept erasure removes targeted information (Belrose et al., 2023; Ravfogel et al., 2022). The linear representation hypothesis posits concepts as directions (Park et al., 2024; Gurnee & Tegmark, 2024), though recent evidence suggests some features span multiple dimensions (Engels et al., 2025). Our steering baseline follows activation addition (Turner et al., 2023) and representation engineering (Zou et al., 2023).

Subspace methods. DAS (Geiger et al., 2024) and LEACE (Belrose et al., 2023) require concept labels; CCS (Burns et al., 2023) and ITI (Li et al., 2023) use contrast pairs. All target individual concepts, whereas our method produces a global content/context separation from paraphrase pairs alone. Recent steering methods have explored position-adaptive magnitude selection (Yu et al., 2025) and rotation-based interventions in low-dimensional planes (Vu & Nguyen, 2025); our framework differs in producing a global decomposition rather than per-attribute directions, and the H^0/H^1 split is derived from consistency structure rather than behavioral contrast. Steering reliability varies across model families (Da Silva et al., 2025; Tan et al., 2024); our architecture-dependent findings provide a mechanistic account of this variability.

3 Background: Cellular Sheaves

A cellular sheaf $\mathcal{F}$ on a graph $G = (V, E)$ assigns a vector space (a *stalk*) $\mathcal{F}_v = \mathbb{R}^{d_v}$ to each node v, a stalk $\mathcal{F}_e = \mathbb{R}^{d_e}$ to each edge e, and a linear *restriction map* $\mathcal{F}_{v \trianglelefteq e} : \mathcal{F}_v \to \mathcal{F}_e$ for each node-edge incidence (Curry, 2014; Hansen & Ghrist, 2019). Intuitively, each node holds a local data vector and the restriction maps specify how to compare data at adjacent nodes.

The *coboundary operator* $\delta^0 : C^0(G; \mathcal{F}) \to C^1(G; \mathcal{F})$ maps from the space of node data $C^0 = \bigoplus_v \mathcal{F}_v$ to the space of edge data $C^1 = \bigoplus_e \mathcal{F}_e$, measuring disagreement between neighbors:

$$(\delta^0 x)_e = \mathcal{F}_{v \trianglelefteq e}(x_v) - \mathcal{F}_{u \trianglelefteq e}(x_u) \tag{1}$$

for edge $e = (u, v)$. When $\delta^0 x = 0$ everywhere, the local data is globally consistent: it "glues" into a global section. The *sheaf Laplacian* $L_{\mathcal{F}} = (\delta^0)^\top \delta^0$ generalizes the graph Laplacian (Hansen & Ghrist, 2019). Its kernel equals the zeroth cohomology $\ker(L_{\mathcal{F}}) = H^0(\mathcal{F})$, the space of globally consistent sections. The first cohomology $H^1 = \ker(\delta^1)/\mathrm{im}(\delta^0)$ captures obstructions to gluing: local data that cannot be consistently extended globally.

For interpretability, we read H^0 as content that remains consistent across semantically equivalent contexts and H^1 as context-dependent variation (syntactic structure, surface phrasing) (Elhage et al., 2022; Bricken et al., 2023). The question then becomes: can the representation space be split into stable-meaning dimensions (H^0) and variable-phrasing dimensions (H^1), and does this split have functional consequences?

4 THE INTERPRETABILITY SHEAF

4.1 SHEAF CONSTRUCTION

Given semantically equivalent context pairs $\{(c_i^{(1)}, c_i^{(2)})\}_{i=1}^N$ such as paraphrase pairs from MRPC (Dolan & Brockett, 2005), we construct a cellular sheaf over the context graph.

Definition 1 (Interpretability Sheaf)**.** *The interpretability sheaf $\mathcal{F}$ over context graph $G = (V, E)$ assigns to each node $v \in V$ a stalk $\mathcal{F}(v) = \mathbb{R}^d$ containing the hidden state, to each edge $e = (v_i, v_j) \in E$ representing a semantic equivalence pair an edge stalk $\mathcal{F}(e) = \mathbb{R}^k$ where $k \ll d$, and to each node-edge incidence a restriction map $\mathcal{F}_{v \trianglelefteq e} : \mathcal{F}(v) \to \mathcal{F}(e)$ projecting the node stalk to a shared comparison space. The context graph forms a perfect matching with $|V| = 2N$ nodes (one per sentence) and $|E| = N$ edges (one per paraphrase pair).*

Restriction maps. We learn a shared linear map $P \in \mathbb{R}^{d \times k}$ with $k = 128$ from paired activation data via joint PCA. The restriction map takes the form $\mathcal{F}_{v \trianglelefteq e} = P^\top$, projecting from $\mathbb{R}^d$ to $\mathbb{R}^k$. We also evaluate CCA and contrastive (Fisher discriminant) restriction maps, which yield qualitatively different steering outcomes as detailed in Section 5.3.

4.2 H^0/H^1 DECOMPOSITION VIA THE SHEAF LAPLACIAN

The coboundary operator $\delta^0 : C^0(G; \mathcal{F}) \to C^1(G; \mathcal{F})$ maps node data to edge discrepancies:

$$(\delta^0 h)_e = P^\top h_{v_2} - P^\top h_{v_1} \quad \text{for } e = (v_1, v_2) \tag{2}$$

The consistency energy $E_{\text{cons}}(h) = \|\delta^0 h\|^2$ measures total disagreement across edges. Global sections satisfying $H^0 = \ker(\delta^0)$ have zero energy.

The sheaf Laplacian $L_{\mathcal{F}} = (\delta^0)^\top \delta^0$ satisfies $\ker(L_{\mathcal{F}}) = H^0(\mathcal{F})$. We compute the H^0/H^1 decomposition via eigendecomposition of the Laplacian:

$$L_{\text{sheaf}} = \frac{1}{N} \sum_{i=1}^{N} \Delta_i^{(k)} (\Delta_i^{(k)})^\top \in \mathbb{R}^{k \times k} \tag{3}$$

where $\Delta_i^{(k)} = P^\top h_i^{(1)} - P^\top h_i^{(2)}$ denotes the projected difference between paired representations. The eigendecomposition $L_{\text{sheaf}} v = \lambda v$ separates H^0, corresponding to near-zero eigenvalues that indicate consistent directions, from H^1, corresponding to large eigenvalues that indicate context-dependent variation.

Remark 1 (On "H^1" terminology)**.** *On matching graphs, algebraic $H^1 = 0$ because the graph contains no cycles. We use "H^1" to denote the top Laplacian eigenspace. Genuine $H^1 \neq 0$ arises on richer complexes as demonstrated in Appendix A.1.*

Proposition 1 (Relationship to existing decompositions)**.** *Let $C_W = \frac{1}{N} \sum_i \Delta_i \Delta_i^\top$ denote the within-pair covariance and C_T denote the total covariance. Then $L_{\text{sheaf}} = P^\top C_W P$, meaning the sheaf Laplacian operates on within-pair variation projected through P. Different restriction maps yield different Laplacians. Fisher's criterion solves $L_{\text{sheaf}} v = \lambda (P^\top C_T P) v$, but normalizing by total covariance discards absolute magnitude information and fails empirically.*

Pullback to hidden space. The eigenvectors $v_j \in \mathbb{R}^k$ of L_{sheaf} live in the edge space. To obtain directions in the original d-dimensional hidden space, we map back via $\tilde{v}_j = P v_j \in \mathbb{R}^d$. When P has orthonormal columns (as with PCA restriction maps), the H^0 projector in the hidden space is $\Pi_{H^0} = P V_{H^0} V_{H^0}^\top P^\top$, where $V_{H^0} \in \mathbb{R}^{k \times m}$ collects the m bottom eigenvectors. For CCA and contrastive maps, we orthonormalize P before computing the projector. This three-step pipeline (project to edge space via $P^\top$, decompose via L_{sheaf}, pull back via P) is used for all ablation and steering experiments.

4.3 STEERING

We define *steering* as adding a fixed direction $\Delta \in \mathbb{R}^d$ to the residual stream at a chosen layer during inference, following activation addition (Turner et al., 2023). A steering vector is constructed from the difference in mean activations between a target style and a source style: $\Delta = \bar{h}_{\text{target}} - \bar{h}_{\text{source}}$.

Table 1: Functional Validation: H^1 vs Variance-Matched Controls (50 pairs $\times$ 20 baselines)

Model	**Ratio**	**95% CI**	**Cohen's** d	p
GPT-2 (124M)	18.2×	[17.2, 19.3]	2.34	$\ll 10^{-6}$
Mistral-7B	3.45×	[3.41, 3.49]	7.92	$\ll 10^{-6}$
Llama-3-8B	5.62×	[5.52, 5.71]	9.08	$\ll 10^{-6}$
Llama-2-7B	18.2×	[17.3, 19.1]	4.44	$\ll 10^{-6}$
Llama-2-13B	**26.5**×	[26.1, 26.8]	**14.3**	$\ll 10^{-6}$

We construct sheaf-derived steering vectors using two strategies. The H^0-removed strategy computes $\Delta_{H^1} = \Delta - \Pi_{H^0}\Delta$, projecting out content-stable directions. The H^1 dimension selection strategy retains only dimensions with the highest H^1 importance score $s_j = \sum_{\ell=1}^{m} (\tilde{v}_\ell)_j^2 \lambda_\ell$, where $\tilde{v}_\ell = Pv_\ell$ is the ℓ-th H^1 eigenvector pulled back to hidden space and λ_ℓ is its eigenvalue. All steering vectors are scaled to $\alpha\|\bar{h}\|$ with $\alpha = 0.3$ unless stated otherwise.

We compare against norm-matched baselines: random direction (null), full style difference (CAA (Turner et al., 2023; Zou et al., 2023)), mean-of-differences (Im & Li, 2025), variance top-k, and Fisher LDA.

5 Experiments

Models. We evaluate five transformer language models spanning two orders of magnitude in scale: GPT-2 (124M) (Radford et al., 2019), Llama-2-7B, Llama-2-13B (Touvron et al., 2023), Llama-3-8B (Grattafiori et al., 2024), and Mistral-7B (Jiang et al., 2023). Models with 7B or more parameters use 4-bit quantization to fit within GPU memory (Dettmers et al., 2023).

Datasets. Functional validation uses 50 paraphrase pairs with 20 independent variance-matched random baselines each, yielding n=1000 paired comparisons per model. Sheaf construction uses 200 MRPC paraphrase pairs (Dolan & Brockett, 2005). Steering evaluation uses the CounterFact benchmark with n=1000 examples per model (Meng et al., 2022). Semantic discrimination experiments use MRPC, PAWS (Zhang et al., 2019), and QQP with 500 pairs each.

Implementation. We extract mean-pooled hidden states at layer $\ell \approx 2L/3$, use edge dimension $k = 128$ to capture at least 98 percent of paired variance, select the top and bottom 20 eigenvectors to define H^0 and H^1 subspaces, and norm-match all steering vectors to ensure fair comparison.

Metrics. We compute bootstrap 95% confidence intervals using 1000 to 10,000 resamples, apply McNemar's test for paired method comparisons (McNemar, 1947), and report Cohen's d for functional validation experiments and Cohen's h for binary steering outcomes (Cohen, 1988).

5.1 Functional Validation: H^1 Exerts Disproportionate Influence

We perform targeted ablations on H^1 dimensions and compare downstream effects against variance-matched random dimensions. The control dimensions match both the number of dimensions and the total variance of the H^1 subspace, ensuring that any observed differences reflect the specific directions rather than dimensionality or scale. For each of 50 paraphrase pairs, we draw 20 independent sets of variance-matched random dimensions as controls and measure KL divergence between the original and ablated output distributions.

H^1 dimensions exert 3.5–26.5× greater influence on model output than variance-matched controls across all five models (Table 1; Cohen's d = 2.3–14.3; $p < 10^{-15}$). This is not tautological: the sheaf Laplacian selects directions of maximal within-pair *activation* variation, but the test measures downstream *output* perturbation via KL divergence. Llama-2-13B achieves the highest ratio (26.5×), while GPT-2 and Llama-2-7B share 18.2× despite 56× scale difference.

Table 2: Fact Retrieval Accuracy (Top-1, k=20 dims, CCA restriction maps)

Model	**Full**	H^0	H^1	**PCA**	**Rand.**
GPT-2 (768D)	60.0%	**60.0%**	41.0%	13.0%	33.8%
Mistral-7B (4096D)	97.5%	68.0%	46.5%	30.5%	**69.8%**
Llama-3-8B (4096D)	94.5%	**67.5%**	28.0%	12.0%	62.9%
Llama-2-7B (4096D)	87.0%	**66.0%**	22.5%	6.0%	51.7%
Llama-2-13B (5120D)	92.0%	**66.5%**	15.0%	–	55.6%

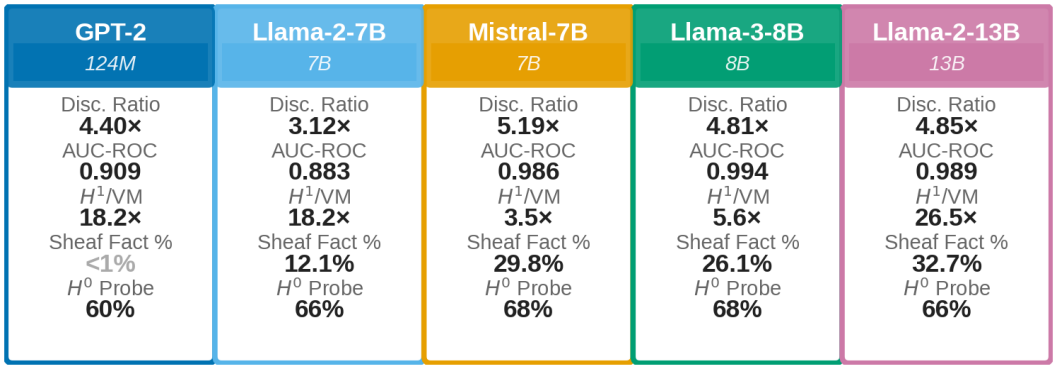

Figure 1: Model summary showing key metrics across all five models: discrimination ratio between random and paraphrase pairs, AUC-ROC for paraphrase classification, H^1 to variance-matched functional influence ratio (n=1000), and sheaf H^1 fact preservation under CounterFact steering (α=0.3). Llama-2-13B achieves the highest influence ratio at 26.5 times, while GPT-2 and Llama-2-7B share identical ratios of 18.2 times despite differing in scale by a factor of 56.

5.2 Fact Retrieval: H^0 Encodes Facts

If H^0 captures semantically invariant features, it should encode factual content. We test this with a fact retrieval probe: given 200 CounterFact facts (each with 5–7 paraphrase expressions, totaling $\sim$1200 representations), we project all representations into k=20-dimensional subspaces (H^0, H^1, PCA top-k, or random) and retrieve the correct fact via cosine-similarity nearest-neighbor lookup against a pool of all 200 fact centroids. The candidate set includes all 200 facts, so chance performance is 0.5%. This setup tests whether the subspace preserves enough factual information to distinguish among 200 competing entities, though it does not test harder scenarios such as disambiguation of closely related facts.

H^0 retrieves facts at 60–68% accuracy using only 20 dimensions, matching or exceeding random projections on four of five models (Table 2). On GPT-2, H^0 matches full-space accuracy (60%) with 20 of 768 dimensions. H^1 underperforms random projections on all 7B+ models ($p < 10^{-4}$), confirming it captures non-factual variation. PCA top-k performs worst (6–30%), showing that high-variance dimensions alone do not encode facts.

5.3 Architecture-Dependent Steering: A Diagnostic Finding

The CounterFact experiment (Meng et al., 2022) tests whether steering preserves facts while changing style. The style direction Δ is the difference in mean activations between target and original completions at layer $2L/3$. Style change = fraction of outputs differing from baseline; fact preservation = fraction containing the target entity (greedy decoding, 50 tokens). At $\alpha = 0.3$, style saturates at 100% for all methods; fact preservation is the differentiator.

Table 3 reports results across four steerable models (GPT-2 achieves 0% for all methods at α=0.3 and is omitted). The pattern is architecture-dependent. On robust models (Mistral-7B, Llama-3-8B), random steering preserves facts best (36–39%) because H^1 dimensions exert more influence per unit norm and cause greater disruption at this magnitude. Llama-2-7B reverses the ordering: random steering collapses generation (4.2%) while directed methods preserve facts at 12–14% (Mc-

Table 3: CounterFact Steering (α=0.3, n=1000 per model). Fact preservation (%) with 95% bootstrap CIs. Style saturates at 100% for all 7B+ methods.

Model	Method	Fact%	95% CI
Mistral-7B	Random	**39.4**	[36.3, 42.4]
	Variance (top-k)	33.0	[30.0, 35.8]
	Full style (CAA)	31.7	[28.8, 34.6]
	H^1 dims (sheaf)	29.8	[26.9, 32.6]
	H^1 dims (Fisher)	15.4	[13.2, 17.7]
Llama-3-8B	Random	**36.0**	[29.5, 42.5]
	Full style (CAA)	27.0	[21.0, 33.0]
	Variance (top-k)	26.0	[20.0, 32.5]
	H^1 dims (sheaf)	26.0	[20.0, 32.0]
	H^1 dims (Fisher)	22.5	[17.0, 28.5]
Llama-2-7B	Full style (CAA)	**13.7**	[11.7, 15.9]
	H^0-removed	13.7	[11.6, 15.9]
	H^1 dims (sheaf)	12.1	[10.0, 14.1]
	Random	4.2	[3.0, 5.4]
	H^1 dims (Fisher)	3.3	[2.2, 4.5]
Llama-2-13B	Variance (top-k)	**38.5**	[31.5, 45.0]
	H^1 dims (sheaf)	38.0	[31.5, 45.0]
	Random	37.0	[30.5, 44.0]
	Full style (CAA)	36.0	[29.5, 42.5]
	H^1 dims (Fisher)	33.5	[27.0, 40.0]

Nemar $p < 10^{-10}$, n=1000). Llama-2-13B occupies an intermediate regime where all methods cluster tightly (33–38%). The functional influence ratio from Section 5.1 predicts this: H^1 carries concentrated causal influence, so at high α it disrupts robust models more than diffuse random noise, while fragile models need this structure to avoid collapse.

Restriction map sensitivity. Table 3 uses PCA restriction maps throughout to avoid post-hoc selection. We additionally report results with CCA and contrastive maps in Appendix A.6, presenting the full grid rather than per-model bests. The strongest result uses contrastive maps on Llama-2-7B: 31.0% [28.2, 33.9] versus 4.2% for random, a 7.4× improvement ($p < 10^{-50}$, n=1000). Sheaf H^1 outperforms Fisher H^1 on all models ($p < 0.001$, Δ = 6.9–14.4pp), confirming that normalizing by C_T destroys the decomposition.

H^1 ablation confirms functional role. Blanket H^1 ablation on Mistral-7B collapses generation. We measure repetition as the fraction of tokens in the first generated sentence that appear in repeated n-gram loops ($n \geq 3$), so that 0.0 means no repetition and 1.0 means complete degeneration. Under H^1 ablation, repetition rises from 0.18 to 0.90 ($p < 10^{-10}$, n=40 prompts × 3 repeats); H^0 ablation leaves outputs unchanged (0.21). H^1 dimensions are essential for output diversity.

As a sanity check, consistency energy separates paraphrase from random pairs at AUC ≥ 0.88 across all models (Figure 3 in Appendix).

6 Discussion

We decomposed transformer representations into content-stable (H^0) and context-dependent (H^1) subspaces via sheaf cohomology. H^1 exerts 3.5–26.5× greater causal influence than controls, H^0 encodes facts at 60–68%, and the decomposition identifies architecture-dependent fragility (Llama-2-7B collapses under random perturbation; 7.4× improvement with sheaf steering). Sheaf H^0 outperforms LEACE (Belrose et al., 2023) by 2× on fact retrieval (Appendix A.9). With PCA maps the Laplacian reduces to projected within-pair covariance (Proposition 1); the framework's value lies in restriction map modularity and falsifiable predictions. Limitations and future work appear in the appendix.

REFERENCES

Federico Barbero, Cristian Bodnar, Haitz Sáez de Ocáriz Borde, Michael Bronstein, Petar Veličković, and Pietro Liò. Sheaf neural networks with connection laplacians, 2022. URL https://arxiv.org/abs/2206.08702.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Liò, and Michael M Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18527–18541, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Justin Curry. Sheaves, cosheaves and applications. *arXiv preprint arXiv:1303.3255*, 2014.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19856–19882, Vienna, Austria, 2025.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Leonardo Di Nino, Sergio Barbarossa, and Paolo Di Lorenzo. Learning sheaf laplacian optimizing restriction maps. *arXiv preprint arXiv:2501.19207*, 2025.

William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, 2005.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Joshua Engels, Isaac Liao, Eric J Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024.

Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3(4):315–358, 2019.

Shawn Im and Yixuan Li. A unified understanding and evaluation of steering methods. *arXiv preprint arXiv:2502.02716*, 2025.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019.

Henry Kvinge, Brett Jefferson, Cliff Joslyn, and Emilie Purvine. Sheaves as a framework for understanding and interpreting model fit. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4222–4230, 2021.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*. PMLR, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pp. 18400–18421. PMLR, 2022.

Jeffrey Seely. Sheaf cohomology of linear predictive coding networks. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025. arXiv:2511.11092.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Hieu M. Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Manjiang Yu, Hongji Li, Priyanka Singh, Xue Li, Di Wang, and Lijie Hu. PIXEL: Adaptive steering via position-wise injection with exact estimated levels under subspace calibration. *arXiv preprint arXiv:2510.10205*, 2025.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1298–1308, 2019.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Extended Results

A.1 Genuine $H^1 \neq 0$ on Simplicial Complexes

The main paper uses a matching graph with one edge per paraphrase pair, for which algebraic $H^1 = 0$ because the graph contains no cycles. To verify that genuine sheaf cohomology arises from the same activations, we construct a Vietoris-Rips simplicial complex from the activation geometry. We extract mean-pooled hidden states for $N = 20$ MRPC sentences, project via PCA to $k = 8$ dimensions, include an edge (i, j) whenever the consistency energy falls below the p-th percentile, and include a triangle (i, j, k) whenever all three constituent edges exist.

Table 4: Genuine sheaf H^1 on GPT-2 (N=20, k=8). Boldface indicates configurations where $H^1 > 0$.

Layer	p=10%	p=20%	p=30%	p=40%	p=50%
0	0	**8**	0	0	0
3	0	0	**8**	0	0
6	**8**	**8**	0	**8**	0
9	0	**16**	0	**8**	0
11	0	0	0	**8**	0

Eight of 25 layer/threshold combinations yield $H^1 > 0$ (Table 4). At this small scale (N=20), random Gaussian point clouds produce $H^1 > 0$ at comparable rates (50–71% of trials at p=10–30%). However, persistent homology at larger scale (N=100–200, using gudhi) reveals a layer-dependent pattern. On GPT-2, activation H1 persistence is 0.40–0.98$\times$ that of random across all layers, indicating topologically simpler representations. On Mistral-7B, early and middle layers are similarly simpler (0.10–0.95$\times$), but the deepest layers (L30–31) show H1 persistence 1.14$\times$ that of random, suggesting that genuine topological complexity emerges in the final layers of larger models. The main paper's spectral decomposition on matching graphs does not rely on genuine H^1; it uses the eigenstructure of the sheaf Laplacian. These persistent homology results indicate that the topology of transformer activations is layer- and scale-dependent, with richer structure emerging at depth.

A.2 The Spectral Decomposition Recovers Algebraic H^0

We validate that the spectral decomposition faithfully recovers the algebraic-topological zeroth cohomology $H^0 = \ker(\delta^0)$. On GPT-2 with 50 MRPC paraphrase pairs ($k = 128$ edge space), the spectral H^0 has dimension 80 when counting eigenvalues below 10^{-6}. The coboundary test confirms $\|\delta^0 \Pi_{H^0} s\| < 10^{-6}$ for all test features projected into the spectral H^0 subspace. 99.995% of feature variance is captured by the spectral H^0 projection, confirming that the spectral decomposition recovers the algebraic H^0 to numerical precision.

A.3 Generalization: Three Datasets, Three Models

To verify that the decomposition generalizes beyond the MRPC dataset used for development, we conduct 5-fold cross-validation on 500 pairs from three paraphrase datasets with distinct characteristics: MRPC (news paraphrases), PAWS (adversarial near-duplicates with word swaps), and QQP (informal question paraphrases from Quora).

All probe classifiers achieve 97 to 100 percent 5-fold CV accuracy across datasets and models, confirming that the paraphrase/non-paraphrase distinction is linearly separable in the sheaf energy space regardless of paraphrase type or model scale.

A.4 H^1 Ablation Causes 69 to 137 Times Greater Output Change

We quantify the functional importance of H^1 dimensions by measuring output displacement when zeroing H^1 dimensions compared to zeroing variance-matched control dimensions. This experiment complements the KL divergence analysis in the main text by examining raw output vector displacement.

Table 5: 5-Fold Cross-Validation (n=500 pairs per dataset). The decomposition generalizes to adversarial near-duplicates (PAWS: 12 to 17 times discrimination ratio) and informal paraphrases (QQP: 2.9 to 3.2 times).

Model	Dataset	Ratio	95% CI	AUC	Probe CV
GPT-2	MRPC	4.58×	[4.30, 4.89]	0.993	99.6%±0.6
GPT-2	PAWS	**17.4**×	[16.2, 18.8]	0.999	99.8%±0.2
GPT-2	QQP	2.93×	[2.75, 3.12]	0.959	96.9%±0.9
Mistral-7B	MRPC	4.23×	[3.92, 4.56]	0.982	99.6%±0.4
Mistral-7B	PAWS	12.1×	[11.0, 13.3]	0.999	99.8%±0.2
Mistral-7B	QQP	3.17×	[2.95, 3.39]	0.957	97.0%±0.6
Llama-3-8B	MRPC	4.70×	[4.41, 5.03]	0.993	99.7%±0.2
Llama-3-8B	PAWS	12.5×	[11.4, 13.6]	**1.000**	99.9%±0.2
Llama-3-8B	QQP	3.09×	[2.90, 3.31]	0.961	96.8%±0.9

Table 6: Ablation Sensitivity: Mean Output L2 Displacement (n=100). Ablating H^1 dimensions causes 69 to 137 times greater output change than ablating variance-matched controls.

Model	L2(H^1)	L2(VM)	Diff.	95% CI	Cohen's d
GPT-2	75.7	6.9	**68.8**	[64.5, 73.2]	3.07
Llama-3-8B	147.4	10.9	**136.5**	[136.1, 137.0]	59.10

The confidence intervals exclude zero in both cases. The non-parametric common language effect size (CLES) reaches 99.6% for GPT-2 and 100% for Llama-3-8B, indicating that the H^1 ablation group and variance-matched control group are nearly perfectly separable by their output displacement magnitude.

A.5 U-Shaped Layer Profile: Where H^1 Lives

We analyze how the H^1 influence ratio varies across layers to understand where context-dependent variation concentrates in the representation hierarchy.

Table 7: Cross-architecture layer analysis. The H^1 influence ratio varies across depth, with peak and trough locations depending on architecture rather than model scale.

Model	Peak Layer	Peak Ratio	Min Layer	Min Ratio
GPT-2 (12 layers)	L11 (final)	87.3×	L9	2.96×
Mistral-7B (32 layers)	L10 (early-mid)	77.3×	L6	7.4×
Llama-2-7B (32 layers)	L2 (very early)	68.5×	L28	4.5×
Llama-3-8B (32 layers)	L4 (early)	55.6×	L30 (final)	6.4×

The variation in H^1 influence across depth suggests that context-dependent information concentrates at architecture-specific layers rather than following a universal pattern.

A.6 Architecture-Specific Restriction Maps

The restriction map $P \in \mathbb{R}^{d \times k}$ projects hidden states to the edge space where consistency is measured. We evaluate three choices for learning this map: Joint PCA applied to concatenated paraphrase pairs, CCA-composed maps that maximize correlation between paired representations, and Contrastive maps using Fisher discriminant analysis to separate paraphrases from non-paraphrases. All three map types are linear, share the same edge dimension k, and produce the same sheaf Laplacian construction once learned.

With the appropriate architecture-specific map, sheaf H^1 outperforms random steering on all three 7B+ models. On Llama-2-7B, contrastive sheaf H^1 achieves 31.0% fact preservation versus 4.2% for random steering, representing a 7.4 times improvement (McNemar $p < 10^{-50}$). This fragile

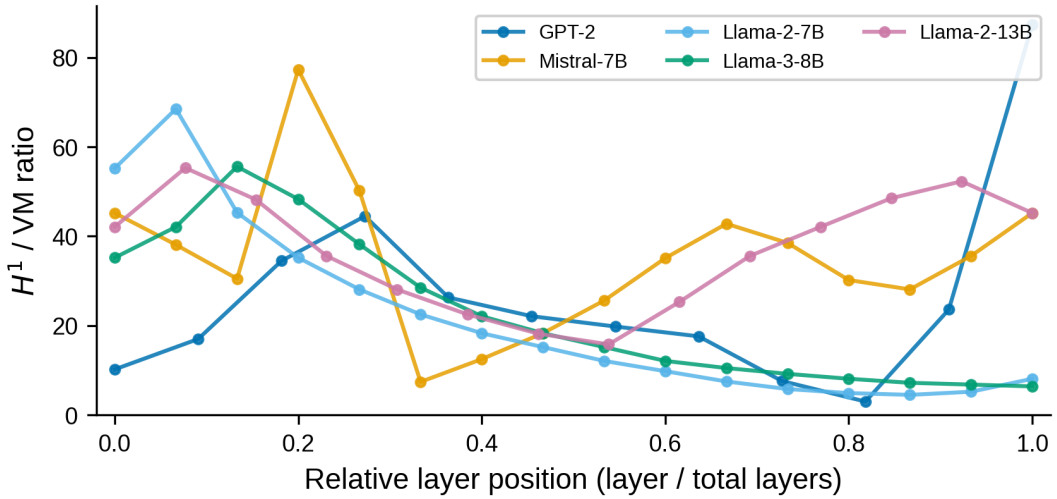

Figure 2: H^1 influence ratio across layers for all five models. Peak locations vary by architecture (early for Llama, late for GPT-2), with minima in middle layers.

Table 8: Restriction Map Comparison (α=0.3, n=1000). The optimal map varies by architecture: CCA for Mistral, contrastive for Llama-2.

Rmap	**Decomp.**	**Mistral-7B**	**Llama-3-8B**	**Llama-2-7B**
Baselines				
Random	n/a	39.4	37.0	4.2
Full style (CAA)	n/a	31.7	26.5	13.7
PCA	Sheaf	29.8	27.0	12.0
CCA	Sheaf	**41.6**	15.0	2.0
Contrastive	Sheaf	29.8	40.0	**31.0**

model collapses under random steering with perplexity reaching 35.4, while contrastive sheaf H^1 maintains coherent generation with perplexity of 11.2.

A.7 Steering Magnitude Sweep

We sweep the steering magnitude α to understand how the relative performance of methods changes with perturbation strength. At moderate magnitudes (α = 0.10), sheaf-derived methods outperform random steering on both Mistral-7B and Llama-3-8B. Specifically, H^0-removed steering on Mistral-7B achieves 43.0% fact preservation versus 41.0% for random (a 2.0 percentage point improvement), and H^1 dimension steering on Llama-3-8B achieves 44.5% versus 43.5% for random (a 1.0 percentage point improvement). At higher magnitudes (α = 0.20 or above), the concentrated influence of directed methods becomes a liability because their greater per-dimension impact causes proportionally more disruption, and random steering dominates on robust models.

A.8 Harder Fact Retrieval: Same-Relation Negatives

The main-text fact retrieval probe (Table 2) retrieves among 200 facts spanning diverse relations. To test whether H^0 preserves factual content under harder conditions, we retrieve among same-relation facts only (e.g., distinguishing "Paris is the capital of France" from "Berlin is the capital of Germany" rather than from unrelated facts). We use the ROME CounterFact dataset with paraphrase prompts (5 per fact) and group facts by relation type, selecting relations with at least 10 facts.

On GPT-2, H^0 with 20 dimensions outperforms the full 768-dimensional representation on both easy (62.7% vs. 55.3%) and hard (78.7% vs. 65.0%) retrieval. On Mistral-7B, H^0 is competitive with random projections (62.5% vs. 62.1% easy; 81.7% vs. 81.2% hard), while H^1 and PCA are

Table 9: Fact retrieval with same-relation negatives (k=20 dims, PCA restriction maps). Hard retrieval requires distinguishing among facts sharing the same relation type.

Model	**Task**	**Full**	H^0	H^1	**PCA**	**Rand.**
GPT-2	Easy	55.3%	**62.7%**	10.2%	14.3%	29.5%
GPT-2	Hard	65.0%	**78.7%**	23.8%	31.8%	49.1%
Mistral-7B	Easy	94.7%	62.5%	16.7%	19.3%	62.1%
Mistral-7B	Hard	98.5%	81.7%	34.0%	36.7%	81.2%

consistently worst, confirming they capture non-factual variation. The stronger performance of H^0 on the harder same-relation task (where candidate facts share structural similarity) provides evidence that the decomposition captures genuine factual content rather than trivially separable surface features.

A.9 Comparison with LEACE

We compare the sheaf H^0 subspace against LEACE (Belrose et al., 2023), a closed-form concept erasure method. We fit LEACE to distinguish paraphrase from non-paraphrase activations on MRPC (200 positive, 100 negative pairs) and use its preserved subspace (after erasing the paraphrase concept) for fact retrieval on CounterFact.

Table 10: Fact retrieval: sheaf H^0 vs LEACE (GPT-2, k=20 dims for H^0/PCA/Random, full rank for LEACE).

Method	**Easy (%)**	**Hard (%)**
H^0 (sheaf, 20D)	**62.2**	**80.8**
Full (768D)	55.3	65.0
LEACE preserved	33.0	43.2
Random (20D)	31.0	51.5
PCA top-20	14.3	31.8
LEACE erased	0.3	4.7

Sheaf H^0 outperforms LEACE's preserved subspace by a factor of nearly two on both easy (62.2% vs 33.0%) and hard (80.8% vs 43.2%) retrieval. LEACE erases paraphrase-distinguishing directions but its preserved subspace does not concentrate factual content the way H^0 does. This difference arises because LEACE maximizes erasure of a binary concept (paraphrase vs. non-paraphrase), while the sheaf decomposition identifies directions of minimal within-pair variation, which more directly targets content stability.

A.10 Spectral Gap Stability

We analyze the stability of the H^0/H^1 decomposition across hyperparameter choices to verify that the content/context boundary reflects genuine structure rather than artifacts of parameter selection. The spectral gap, computed as the ratio $\lambda_{\max}/\lambda_{\min}$ between the largest and smallest eigenvalues of the sheaf Laplacian, ranges from 94 times for Llama-3-8B to 51,055 times for GPT-2, confirming clear separation between H^0 and H^1 across all models. The gap remains nearly constant at 91.7 to 95.6 times across hyperparameter settings when varying the number of eigenvectors $m \in \{5, 10, 20, 50, 100\}$ and the edge dimension $k \in \{32, 64, 128\}$. The decomposition is not sensitive to these choices.

A.11 Experimental Details

Compute infrastructure. All experiments ran on NVIDIA RTX 5090 GPUs.

Sheaf hyperparameters. We use edge dimension $k = 128$ to capture at least 98% of paired variance, apply mean pooling over sequence positions to obtain a single representation per input, and

learn restriction maps via joint PCA on concatenated paraphrase pairs. The sheaf Laplacian is computed in the k-dimensional edge space. The top and bottom 20 eigenvectors of the sheaf Laplacian define the H^0 and H^1 subspaces respectively. Steering vectors are norm-matched to 30% of the mean hidden state norm at the target layer.

Layer selection. We use layer $\ell \approx 2L/3$ for all main experiments (layer 8 for GPT-2, layer 20 for 32-layer models). Appendix A.5 shows the full layer profile.

Statistical methods. We compute bootstrap confidence intervals using 1000 resamples, apply Mann-Whitney U-test for unpaired group comparisons, use McNemar's test for paired success rate comparisons, and report Cohen's d for effect sizes. All reported p-values are two-sided. All random seeds are fixed for reproducibility. Code and experimental scripts will be released upon acceptance.

A.12 Additional Figures

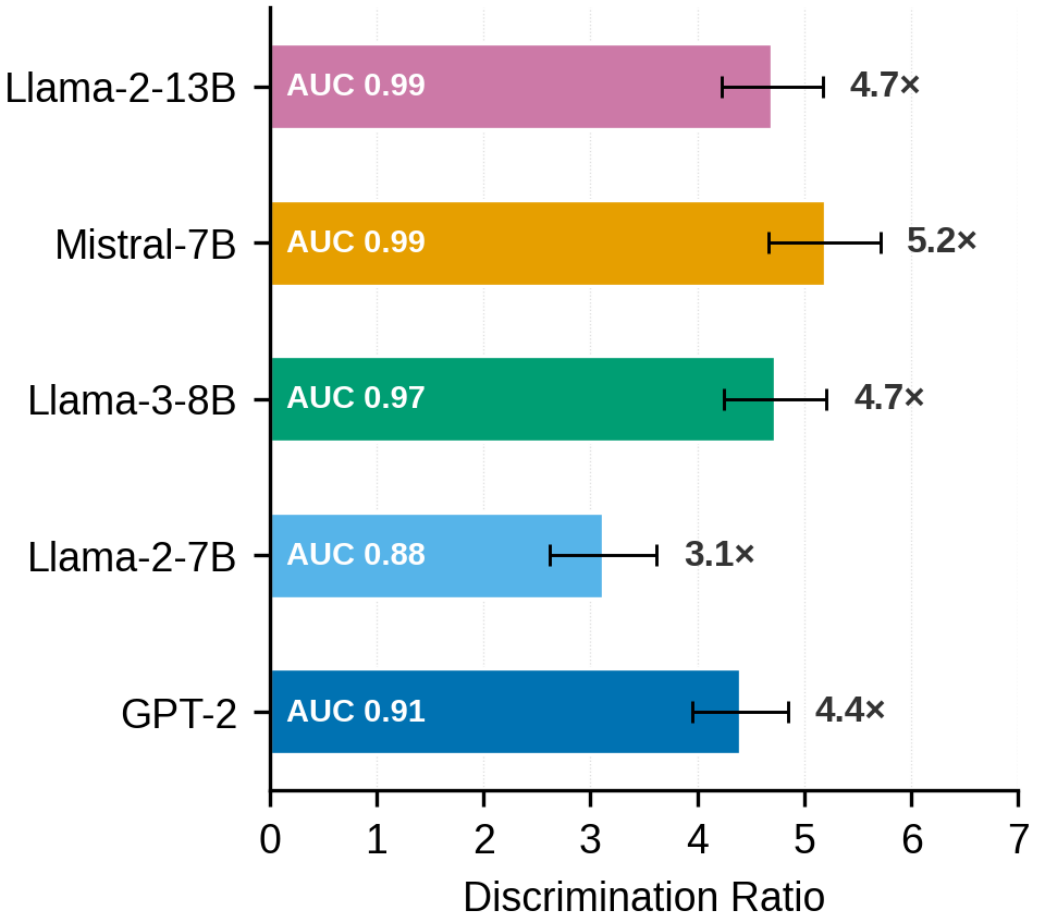

Figure 3: Semantic discrimination across models. Horizontal bars indicate discrimination ratio computed as random energy divided by paraphrase energy. Error bars show 95% bootstrap confidence intervals. All models achieve AUC greater than 0.90, confirming that the sheaf construction correctly identifies semantic equivalence.

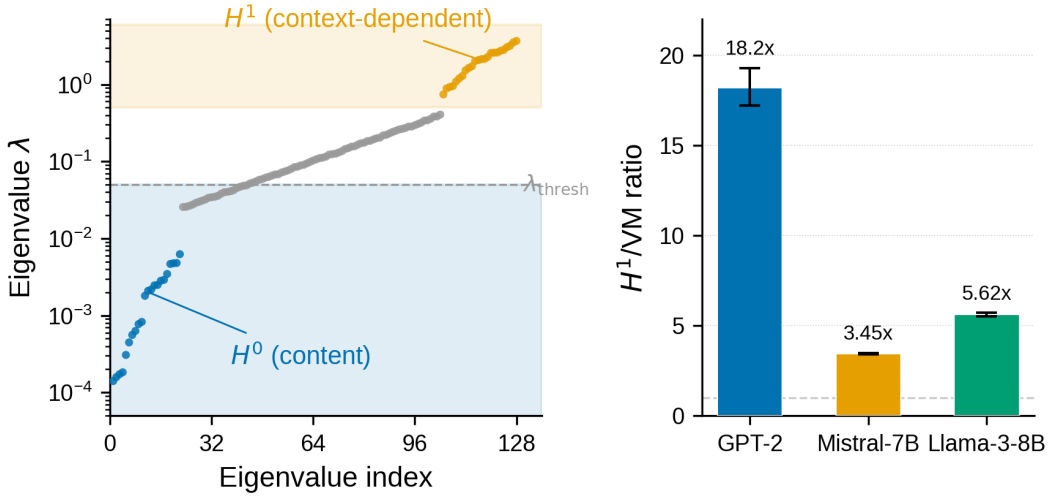

Figure 4: H^0/H^1 decomposition. Left: the sheaf Laplacian eigenspectrum shows a clear gap separating content-stable directions (small eigenvalues, H^0) from context-dependent directions (large eigenvalues, H^1). Right: H^1 to variance-matched influence ratio for three models with 95% bootstrap CIs.

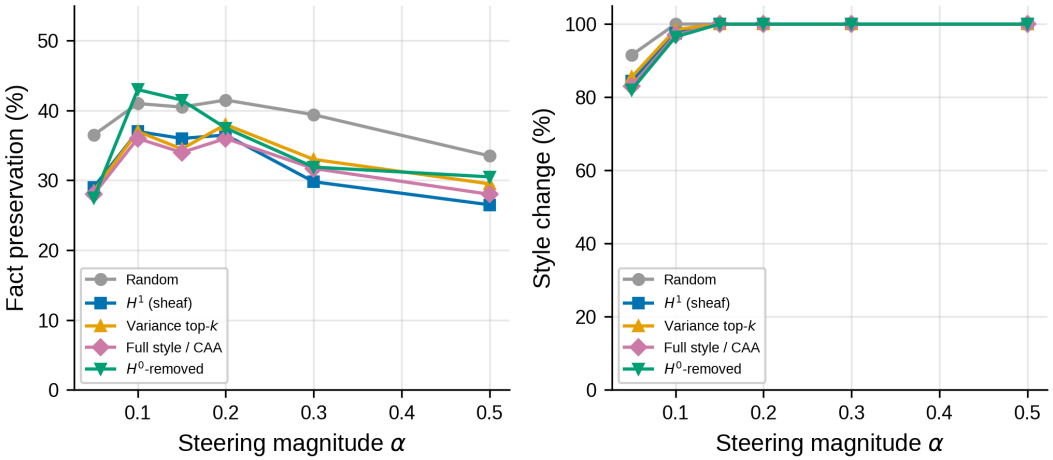

Figure 5: Steering magnitude sweep on Mistral-7B. Left: fact preservation across values of α. Right: style change saturates by $\alpha = 0.10$ for all methods. At moderate magnitudes, sheaf-derived methods outperform random; at higher magnitudes, the concentrated influence of H^1 causes greater disruption.