FREQUENCY, DIRECTION, AND FEATURE SHARING: DISSECTING CO-ACTIVATION IN SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Co-activation in sparse autoencoders (SAEs) is often used to infer concept hierarchies, but shared features and parent–child direction require different evidence. We show that on non-tied pairs, reciprocal coverage orients concepts solely by support-set size. In a complete noiseless chain model with an unknown full-rank dictionary, every ordering of L nonzero states admits an exact representation, and an L_0 penalty favors frequent states at shallow depths rather than identifying semantic order. On 5,572 WordNet pairs with Llama-3.1-8B/JumpReLU, direction accuracy is 79.1% when the child-to-parent corpus-frequency ratio is below one and 25.3% otherwise. In a separate matched-control analysis, a lexical-frequency baseline scores 66.1% versus 61.1% for the support rule, while hierarchical pairs show greater excess overlap than frequency-matched random controls (Cohen's $d = 0.91$). Replacing JumpReLU with TopK at fixed encoder weights and approximately matched mean activation counts reduces the frequency-regime gap from 53.8 to 6.4 points. Our results distinguish frequency-associated orientation from semantic sharing and give sufficient conditions for latent-order recovery within the chain model.

1 INTRODUCTION

Sparse autoencoders (SAEs) expose sparse features in language-model activations (Cunningham et al., 2024; Bricken et al., 2023; Gao et al., 2025; Templeton et al., 2024). Hierarchy-aware dictionaries sharpen a natural question: do relations among learned features reveal relations among concepts (Bussmann et al., 2025; da Costa et al., 2025; Cao et al., 2026)? Shared activation for *animal* and *dog* may indicate common content; orienting *animal* as the parent is a separate inference. This paper identifies what each observation can establish.

Reciprocal coverage is a revealing case. Although it compares two conditional overlaps, its non-tied direction algebraically reduces to a comparison of support sizes. We study the consequence at three levels: the statistic itself, latent-order identifiability with an unknown dictionary, and measured SAE supports under frequency controls and a fixed-weight gate replacement. Figure 1 illustrates why overlap and orientation can diverge even when the shared features are unchanged.

In a complete noiseless chain, all $L!$ orders of the observed nonzero states have exact dictionary representations; sparse code length ranks them by state probability. In trained SAEs, taxonomic pairs share more features than frequency-matched controls, while the support-based direction tracks lexical frequency. On 5,572 WordNet pairs, the rule scores 79.1% below a child-to-parent corpus-frequency ratio of one but 25.3% above it; replacing only the inference-time gate sharply reduces this contrast. The corpus ratio indexes the empirical benchmark, whereas state probability governs the model's sparsity cost.

We contribute (i) an exact size-based characterization of reciprocal-coverage direction; (ii) a latent-order ambiguity theorem, probability-dependent cost result, and explicit sufficient recovery conditions; and (iii) an SAE evaluation that distinguishes orientation from sharing while isolating the effect of feature selection within fixed-weight checkpoints.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

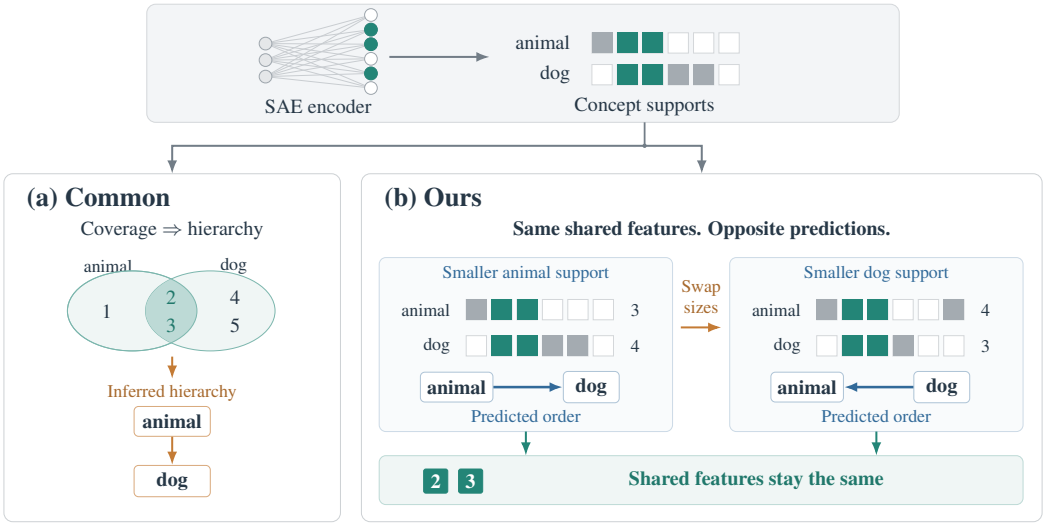

Figure 1: **Overlap and direction answer different questions.** Both examples share features $\{2, 3\}$ and have $|S_P||S_C| = 12$, so excess overlap is unchanged. Changing only nonshared memberships changes $(|S_P|, |S_C|)$ from $(3, 4)$ to $(4, 3)$ and reverses reciprocal coverage. Teal marks shared features; gray marks other active features.

2 Related Work

SAE features and selection. SAEs expose sparse directions in language-model activations (Cunningham et al., 2024; Bricken et al., 2023); scaling studies extend the practice to larger dictionaries and models (Gao et al., 2025; Templeton et al., 2024). Gated and JumpReLU variants alter feature selection (Rajamanoharan et al., 2024a;b), while feature absorption can redistribute related information (Chanin et al., 2025). These results motivate measuring the gate and support construction rather than reading direction directly from co-activation.

Learning hierarchical sparse representations. Matryoshka SAEs organize features across nested widths (Bussmann et al., 2025); matching-pursuit, Tree-SAE, and hierarchical-SAE approaches seek multilevel feature organization (da Costa et al., 2025; Cao et al., 2026; Luo et al., 2026). Hierarchical group penalties provide another route to structured sparse coding (Jenatton et al., 2011; Bach et al., 2012). Such training objectives differ from our post-hoc question: whether reciprocal coverage of already learned supports identifies a parent–child direction.

Identifiability of latent representations. Dictionary and factor recovery require assumptions on atoms, codes, or observations (Spielman et al., 2012; Arora et al., 2013; Fu et al., 2019). Unsupervised representation learning likewise needs inductive restrictions to prefer one latent explanation (Locatello et al., 2019). Our chain construction isolates ambiguity in the ordering of observed states under an unknown unrestricted dictionary, while its recovery results show what additional information removes that ambiguity.

Geometry and additional information. Geometric analyses ask what hierarchical structure is present in activations (Park et al., 2025). Observational equivalence in causal models and identification through auxiliary variation or interventions distinguish structure from direction (Verma & Pearl, 1990; Hyvärinen et al., 2019; von Kügelgen et al., 2023; Buchholz et al., 2023). Our amplitude, intervention, and dictionary conditions make this distinction explicit within the defined chain model.

3 PRELIMINARIES

SAE representations. For $h \in \mathbb{R}^d$, an encoder gives $f(h) \in \mathbb{R}^m_{\geq 0}$ and a decoder reconstructs $\hat{h} = Df(h) + b$. We compare candidate representations with

$$\mathcal{J} = \mathbb{E}\|h - \hat{h}\|_2^2 + \lambda \mathbb{E}\|f(h)\|_0. \tag{1}$$

Fixed-count selection makes the penalty constant; variable-count gates retain it.

Concept supports. For a word w in T templates, define

$$S(w) = \left\{ j : \sum_{t=1}^{T} \mathbf{1}[f_j(h_{w,t}) > 0] \geq K, \quad g_j \leq \tau \right\}, \tag{2}$$

where g_j is global feature firing rate. Aggregated support size differs from per-token activation count $\|f(h)\|_0$.

Direction and sharing. For nonempty supports S_P, S_C, set $I = |S_P \cap S_C|$ and

$$d(P, C) = \frac{I}{|S_P|} - \frac{I}{|S_C|}, \tag{3}$$

$$o(P, C) = I - \frac{|S_P||S_C|}{M}, \tag{4}$$

where M is the number of eligible features. Predict P for $d > 0$, reverse for $d < 0$, and abstain at zero. Excess overlap uses the expected intersection of uniform sets with these sizes; it does not adjust heterogeneous feature propensities.

4 PROBLEM FORMULATION

We distinguish empirical orientation scored against semantic labels from latent-order identification using observations alone.

Definition 1 (Complete noiseless chain model)**.** Let $D = [v_1, \ldots, v_L] \in \mathbb{R}^{d \times L}$ have full column rank. For depth $k \in \{0, \ldots, L\}$, let $z^{(k)}$ be the binary prefix code with its first k entries equal to one. Draw K with $\mathbb{P}(K = k) = p_k$, where $p_k > 0$ for $k \geq 1$, and observe

$$x = Dz^{(K)}, \qquad x_k = \sum_{i=1}^{k} v_i, \qquad x_0 = 0. \tag{5}$$

The dictionary and assignment of nonzero observations to depths are unknown; atom norms, signs, and inner products are otherwise unrestricted.

The prefix constraint leaves $L + 1$ states. Reordering need not preserve normalization or orthogonality, so the equivalence result uses unrestricted atoms and complete noiseless chains rather than arbitrary graphs.

Two-state notation and empirical proxy. For $L = 2$, write $x_a = a$ and $x_b = a + b$ with exclusive probabilities $p_a, p_b > 0$ and ratio $r = p_b/p_a$. Here p_a is parent-only state mass, not the parent component's marginal firing probability $p_a + p_b$. For WordNet parent–child pairs, we separately use the lexical ratio $r_{\text{corpus}} = \text{freq}(C)/\text{freq}(P)$; it is not the latent ratio r.

Information available to an estimator. A latent-order estimator sees samples of x, but not true codes, the generating dictionary, semantic labels, or interventions. It must identify which named observed state is shallower. In the empirical task, WordNet supplies the reference order and SAE supports supply a prediction.

5 Analysis Framework and Theoretical Results

The framework separates a support statistic from latent-order identification. Figure 2 shows how paired contexts yield distinct direction and sharing readouts. We then characterize exact observational ambiguity, the preference induced by sparsity, and information that identifies order. Proofs appear in Appendix A.

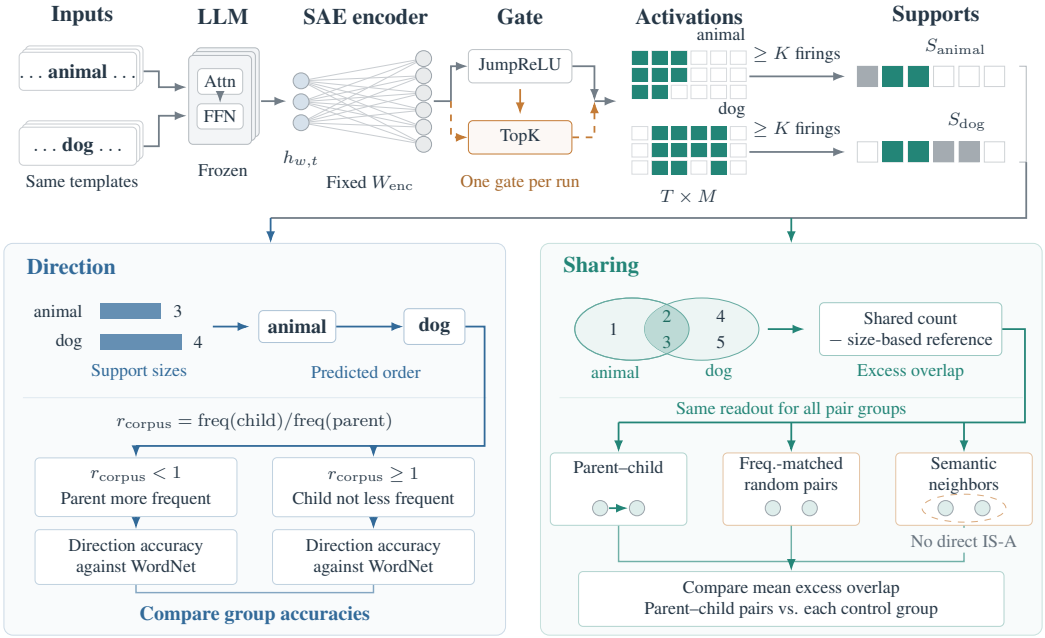

Figure 2: **From paired contexts to separate direction and sharing readouts.** Shared templates feed a frozen model and SAE encoder; JumpReLU and TopK select features in separate runs. Direction is scored against WordNet by corpus-frequency regime, while excess overlap is compared across hierarchical pairs and controls. The toy grids use $T = 3$ and $K = 2$.

5.1 Direction reduces to support size

Proposition 2 (Reciprocal-coverage cancellation)**.** *For nonempty S_P, S_C with $I > 0$,*

$$\operatorname{sign} d(P, C) = \operatorname{sign}(|S_C| - |S_P|). \tag{6}$$

If $I = 0$ or $|S_P| = |S_C|$, the coverage rule ties.

Shared-feature identity and intersection size affect whether the rule abstains, but not the direction of a non-tied verdict.

5.2 Observational ambiguity with an unknown dictionary

Theorem 3 (Two-state latent-order ambiguity)**.** *Let a, b be linearly independent and $p_a, p_b > 0$ with $p_a + p_b \leq 1$. The observation law supported on $\{0, a, a + b\}$ with masses $1 - p_a - p_b, p_a, p_b$ has two exact chain explanations:*

$$D_{\text{I}} = [a, b], \qquad z(x_a) = (1, 0), \qquad z(x_b) = (1, 1), \tag{7}$$

$$D_{\text{II}} = [a + b, -b], \qquad z'(x_a) = (1, 1), \qquad z'(x_b) = (1, 0). \tag{8}$$

The shallower nonzero state differs. At any sample size, observation-only estimators have worst-case error at least $1/2$ over this two-model family.

Any fixed feature-extraction procedure on the same observations preserves equality of the two data laws; the ambiguity concerns the latent explanation, not whether the observed points differ.

Theorem 4 (Exact chain representations)**.** *Within Definition 1, a law with L linearly independent nonzero support points has exactly $L!$ ordered depth-indexed dictionary representations. For each permutation π of those points, the dictionary is uniquely*

$$w_k = x_{\pi(k)} - x_{\pi(k-1)}, \qquad x_{\pi(0)} := 0, \tag{9}$$

with probability $p_{\pi(k)}$ assigned to depth k.

The permutation acts on observed prefix sums rather than the original atoms. Figure 5 in Appendix A illustrates the two-state case.

5.3 Frequency selects a low-cost representation

Proposition 5 (Frequency-dependent L_0 preference)**.** *For the two exact representations of Theorem 3,*

$$C_{\mathrm{I}} = p_a + 2p_b, \quad C_{\mathrm{II}} = 2p_a + p_b, \quad C_{\mathrm{I}} - C_{\mathrm{II}} = p_a(r-1). \tag{10}$$

Model I has lower expected code length for $r < 1$, Model II for $r > 1$, and they tie at $r = 1$. More generally, the cost of ordering π is $\sum_{k=1}^{L} k p_{\pi(k)}$, minimized by placing larger-probability states earlier.

Equal reconstruction error means an L_0 penalty ranks these exact representations by state probability, not by externally supplied semantic order.

Normalized penalties and lossy alternatives. With unit reference atoms, $\rho = \langle a, b\rangle$, and $c = \|a+b\|_2$, normalized-decoder L_1 costs cross at $r^* = c/(2-c)$, equal to $1+\sqrt{2}$ when $\rho = 0$. A one-atom candidate along $a+b$ saves an expected $\min(p_a, p_b)$ active feature but incurs error $p_a(1-\rho)/2$. Under Eq. (1), it beats the better exact candidate when

$$\lambda \min(p_a, p_b) > \frac{p_a(1-\rho)}{2}. \tag{11}$$

These comparisons concern normalized or lossy candidate classes, not the exact binary-code orbit; derivations are in Appendix A.

5.4 Recovery requires specified additional information

Proposition 6 (Sufficient recovery conditions)**.** *The two-state order is identified under any one of these distinct additions:*

- *(a)* ***Amplitude variation and state grouping.*** *The separately observed populations generate $x = s_A a$ and $x = s_A a + s_B b$. The positive amplitudes are independent, have finite nonzero variances, and a, b are linearly independent; conditional covariance ranks one and two identify the parent-only population.*
- *(b)* ***Causal intervention access.*** *With independent nondegenerate Bernoulli variables U_A, U_B, the structural equations are $Z_A = U_A$ and $Z_B = Z_A U_B$. Coordinate-level interventions and readout distinguish the upstream variable: $\mathrm{do}(Z_A = 0)$ suppresses Z_B, whereas $\mathrm{do}(Z_B = 0)$ preserves Z_A.*
- *(c)* ***Dictionary information.*** *The generating full-rank dictionary is known, or the unknown two-atom dictionary is restricted to nonzero mutually orthogonal atoms.*

These conditions add variation, causal access, or a dictionary restriction absent from the exact observational construction.

6 Experiments

We ask whether trained SAEs exhibit the separation between support direction and sharing, and how that direction responds to feature selection. The empirical corpus-frequency ratio is descriptive; it is not the latent state-probability ratio in Section 5.

6.1 SETUP AND REPORTING CONVENTIONS

The principal setting uses Llama-3.1-8B with the Llama Scope layer-16 JumpReLU SAE (width 32,768) (He et al., 2024; Rajamanoharan et al., 2024b). Additional settings use GPT-2-small/OpenAI TopK and Gemma-2-9B/GemmaScope (Gao et al., 2025; Gemma Team et al., 2024; Lieberum et al., 2024); these cross-setting contrasts vary multiple factors. The GPT-2 and Gemma evaluations contain 5,554 and 6,633 valid pairs, respectively, and differ in base model, layer, training recipe, and gate. WordNet supplies 5,572 principal parent–child pairs (Miller, 1995), and lexical frequencies use 200M C4 tokens (Raffel et al., 2020).

Unless stated otherwise, supports use $T = 20$ templates, $K = 3$, and $\tau = 0.10$ in Eq. (2). Thresholds were selected in-sample; Appendix B.2 reports the full grid. Accuracy conditions on non-tied predictions, with decision counts reported. Principal direction CIs use 10,000 pair-bootstrap resamples; the significance test clusters on both words. Define

$$\Delta = \mathrm{Acc}(r_{\mathrm{corpus}} < 1) - \mathrm{Acc}(r_{\mathrm{corpus}} \geq 1). \tag{12}$$

Control and recovery analyses are separate runs with their own denominators. The main extraction gives 3,327 positive, 2,141 negative, and 104 tied verdicts, for 60.8% pooled non-tied accuracy.

6.2 SYNTHETIC CHECKS AND THEIR SCOPE

Synthetic experiments use a two-state chain in $d = 64$ with ten seeds per training cell. An 8×7 normalized-L_1 grid varies $\lambda \in \{0.05, \ldots, 0.4\}$ and $\rho \in \{0, \ldots, 0.6\}$; its crossover approaches the exact-reconstruction value $1 + \sqrt{2}$ at $\rho = 0$ as $\lambda \to 0$. This numerical grid and the exact coefficient-cost comparison address distinct penalty regimes.

A 320-SAE sweep covers four sparsity families and eight ratios $r \in \{0.5, 1, 1.5, 2, 2.414, 3, 4, 6\}$. Among L_1 runs for which the support certificate decides, direction accuracy is 100% at $r = 0.5$ and 0% at $r \geq 1.5$. Table 1 reports the trained code-count signatures, which depend on the selection rule; counts alone do not determine the decoder. With independent $U(0.6, 1.4)$ amplitudes and observed state groups, noiseless covariance-rank recovery is 100% across tested ratios (50 seeds each); intervention simulation succeeds in 50 configurations of its structural model.

Table 1: **Synthetic code-count signatures over 320 runs.** Columns count active coordinates on parent-only and joint states; toy TopK uses $k = 1$.

Training family	$(1, 2)$	$(2, 1)$	$(1, 1)$	Other
L_1	0	0	80	0
TopK ($k = 1$)	0	0	78	2
JumpReLU	36	6	24	14
L_0-style	0	0	72	8

6.3 DIRECTION TRACKS CORPUS-FREQUENCY REGIMES

The support rule yields 79.1% accuracy for $r_{\mathrm{corpus}} < 1$ and 25.3% for $r_{\mathrm{corpus}} \geq 1$, a gap of $\Delta = 53.8$ percentage points (95% CI $[51.4, 56.2]$; two-way cluster-robust $p = 8.8 \times 10^{-132}$). The extreme bins in Figure 3a are 84.3% ($2462/2922$, $r_{\mathrm{corpus}} < 0.5$) and 17.0% ($204/1203$, $r_{\mathrm{corpus}} \geq 2$). The intermediate bins score 57.3% ($396/691$) and 40.6% ($265/652$); a 5,422-pair unique subset gives $\Delta = 54.4$ points. Table 2 compares the principal direction baselines.

M1 and the size-only M2 nearly coincide, as Proposition 2 predicts; sharing affects M1's abstention but not a non-tied sign. Mean token activation count exhibits a similarly large regime gap.

6.4 SENSITIVITY TO SUPPORT CONSTRUCTION AND LAYER

A separate 20-template pass sweeps $K = 1$–5 and $\tau \in \{0.05, 0.10, 0.15\}$. For $K = 2$–5, the gap remains 47.5–53.8 points across cutoffs; $K = 1$ gives 22.6–24.1 points. At $\tau = 0.10$, mean support size falls from 211.4 features at $K = 1$ to 58.6 at $K = 3$. Table 3 shows the $\tau = 0.10$ slice; Appendix B.2 contains all 15 cells.

Table 2: **Direction baselines in the principal 20-template run.** Accuracy conditions on decided pairs; Δ is in percentage points.

Method	Decided	Overall	$r_{\text{corpus}} < 1$	$r_{\text{corpus}} \geq 1$	Δ
M0: enrichment argmax	5,572	49.9	51.4	46.9	4.5
M1: reciprocal coverage	5,468	60.8	79.1	25.3	53.8
M2: support size	5,471	60.8	79.1	25.3	53.9
M3: weighted inclusion	5,568	55.4	66.3	34.4	31.9
Mean token L_0 direction	5,572	60.7	79.7	23.8	55.9

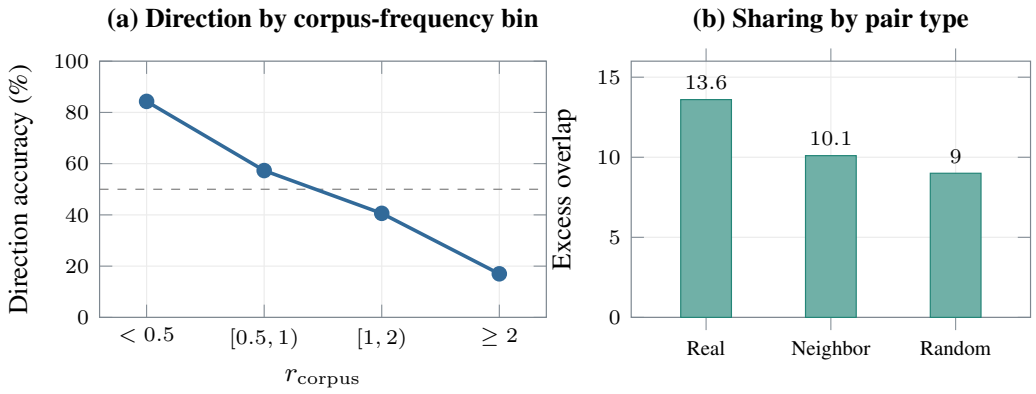

Figure 3: **Direction and sharing give distinct readouts.** (a) Direction accuracy across corpus-frequency bins; the dashed line marks 50%. (b) Mean excess overlap for WordNet pairs and two controls.

Table 3: **Support-threshold sensitivity at $\tau = 0.10$.** This is a separate encoding pass; accuracies condition on decided pairs.

K	Decided	Overall	$r_{\text{corpus}} < 1$	$r_{\text{corpus}} \geq 1$	Δ
1	5,534	46.40	54.47	30.79	23.68
2	5,472	56.80	73.38	24.54	48.84
3	5,471	61.05	79.29	25.46	53.83
4	5,424	62.59	80.53	27.48	53.05
5	5,393	63.23	80.29	29.63	50.67

At Llama layers 8, 16, and 24, the gaps are 47.3, 53.8, and 38.3 points; overall accuracies are 59.3%, 60.8%, and 56.7%. The corresponding gap CIs are $[44.7, 49.7]$, $[51.4, 56.2]$, and $[35.7, 40.9]$. These sweeps measure sensitivity, while the gate comparison below holds weights fixed.

6.5 Frequency controls and semantic sharing

Random controls match both log word frequencies within ± 0.3 and exclude WordNet relations. Taxonomic-neighborhood controls pair siblings of the parent and child, filter direct IS-A edges, and match log-frequency ratio within ± 0.5. The ordered control slots are not verified semantic directions. Table 4 compares orientation and sharing across groups.

Table 4: **Similar slot orientation, different sharing.** Pair counts include ties; control M1 rates are slot agreement, not semantic accuracy.

Group	Pairs	M1 (%)	Δ (pp)	Mean o	Jaccard
WordNet parent–child	5,572	61.1	54.1	13.6	.157
Frequency-matched random	55,700	61.7	48.3	9.0	.099
Taxonomic neighborhood	26,404	61.3	45.0	10.1	.111

A higher-frequency-word baseline scores 66.1% on real pairs, versus 61.1% for M1 in this separate control run. The real–random slot-agreement difference is -0.56 points (90% CI $[-2.7, 1.6]$; TOST $p = 3 \times 10^{-4}$ at a ± 5-point margin). Sharing instead rises from 9.0 excess features for random controls and 10.1 for neighborhood controls to 13.6 for WordNet pairs. The real–random effect is Cohen's $d = 0.91$; the frequency-adjusted real–neighborhood contrast is $+3.7$ features (95% CI $[3.4, 4.0]$). These are distinct measurements of direction and overlap under the specified controls.

6.6 Controlled gate and width comparisons

At Llama Scope layer 16, two separately trained checkpoints have widths 32K and 128K. Within each width, we hold encoder weights and preactivations fixed and replace native JumpReLU selection with TopK calibrated to mean activation count (Figure 4). At 32K, mean token L_0 changes from 43.671 to 44.000 while Δ falls from 53.8 to 6.4 points; at 128K, it changes from 40.370 to 40.000 while Δ falls from 47.8 to 5.9. Mean child-minus-parent support size contracts from $+5.8$ to -0.1 and from $+3.3$ to $+0.4$ features, respectively.

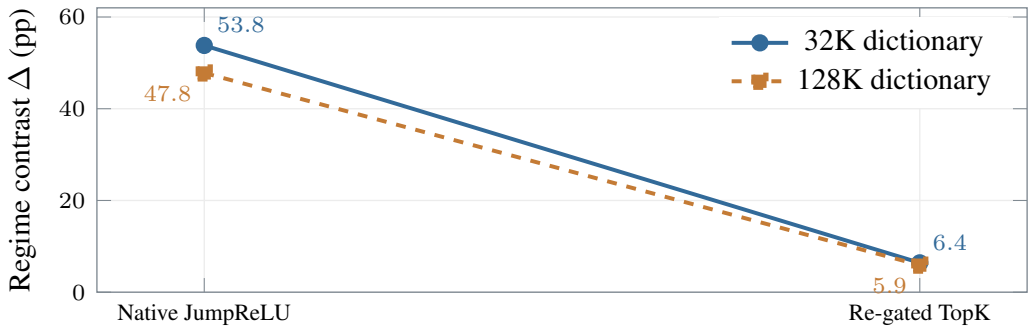

Figure 4: **A fixed-weight inference-time gate comparison.** Each line holds encoder weights and width fixed while approximately matching mean token L_0; the width contrast uses separately trained checkpoints.

Fixed per-token counts do not force equal aggregated supports; the measured contraction accompanies the gate swap. Reconstruction cosine rises from .7991 to .8205 at 32K and from .8067 to .8264 at 128K. Cross-setting gaps are -28.8 points for GPT-2 and $+0.7$ for Gemma; their model, layer, and training differences make them descriptive rather than gate-only contrasts (Appendix B.4).

6.7 Recovery reduces reversal but does not establish identification

A separate 30-template readout selects each word's top three enriched features and measures effective rank over their union. It decides 99.9% of pairs, versus 96.8% for M1. At $r_{\text{corpus}} \geq 1$, accuracy rises from 26.7% for M1 to 49.7% for focused rank (95% CI $[47.4, 51.9]$); the systematic reversal disappears, while above-chance orientation remains unestablished. The full comparison is in Appendix B.6.

7 Discussion and Limitations

The formal ambiguity and empirical frequency association answer different questions. Matched controls distinguish sharing from orientation, while fixed-weight re-gating localizes the latter to feature selection in the tested checkpoints. Corpus frequency remains a proxy rather than latent-state probability.

The chain model becomes identifiable with state-grouped amplitude variation, coordinate-level interventions, or dictionary restrictions. For trained SAEs, direction should be benchmarked against lexical frequency and sharing measured against matched controls, rather than treating one coverage score as a hierarchy readout.

8 Conclusion

Support-based direction and feature sharing are distinct inference problems. Reciprocal coverage reduces to support-size ordering, while complete noiseless chains admit $L!$ exact latent orders whose sparsity costs favor common states at shallow depths. In trained SAEs, direction is strongly frequency-associated and gate-sensitive, even as taxonomic pairs exhibit elevated feature sharing. Evaluating these readouts separately clarifies what co-activation demonstrates and what additional information is needed to identify semantic order.

AI Assistance

AI assistance was used in preparing this work to reorganize the manuscript, revise the presentation of theoretical statements and proofs, and prepare explanatory figures and LaTeX. This was an editorial revision: no additional model experiments were run, and all empirical values are retained unchanged from the supplied manuscript rather than generated by AI. The authors reviewed each AI-revised statement and proof against the original results to confirm that the revisions preserved the intended claims, and verified that the empirical values carried over from the source manuscript were unchanged. The authors take full responsibility for the final content of this work, including any material produced with the aid of generative AI.

Ethics Statement

This study analyzes model representations using public language-model resources and lexical relations. It does not involve human-subject experiments or private user data. Interpretation of individual SAE features and their semantic relations remains uncertain; the results should not be taken as evidence of uniquely recovered human concepts.

Reproducibility Statement

Appendix A provides proofs under explicit assumptions. Evaluation settings, reported aggregate measurements, and the distinction between separate runs are specified in Section 6 and the experimental appendix. Reproducing the empirical results requires the model checkpoints, exact pair lists, templates, tie-handling rules, and analysis code in addition to these aggregates. The analysis and evaluation code, data generators, and figure-generation scripts are provided in the anonymized supplementary material; the WordNet pair lists and corpus-frequency statistics can be regenerated

from the included scripts, and the empirical experiments use the publicly released Llama Scope and OpenAI GPT-2 SAE checkpoints.

REFERENCES

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, 2013.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

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.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *International Conference on Machine Learning (ICML), PMLR 267*, pp. 6077–6101, 2025.

Tue Minh Cao, Hoang X. Nhat, Raed Alharbi, Phi Le Nguyen, and My T. Thai. Tree SAE: Learning hierarchical feature structures in sparse autoencoders. *arXiv preprint arXiv:2605.07922*, 2026.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 (ICLR)*, 2024.

Vinicius Oliveira da Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba E. Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, and Wing-Kin Ma. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Processing Magazine*, 36(2):59–80, 2019.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, et al. Llama scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868, 2019.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

Tom Lieberum, Senthooran Rajamanoharan, Simon Conly, Jonah Smith, Nicolas Sonnerat, Andonia Varma, János Kramár, Aidan Dragan, Vishal Shah, and Neel Nanda. GemmaScope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019.

Yifan Luo, Yang Zhan, Jiedong Jiang, Tianyang Liu, Mingrui Wu, Zhennan Zhou, and Bin Dong. From atoms to trees: Building a structured feature forest with hierarchical sparse autoencoders. *arXiv preprint arXiv:2602.11881*, 2026.

George A Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b. doi: 10.48550/arXiv.2407.14435.

Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory (COLT), JMLR Workshop and Conference Proceedings*, volume 23, pp. 37:1–37:18, 2012.

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.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pp. 255–270, 1990.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

A Proofs and Scope of the Formal Results

A.1 Reciprocal-coverage cancellation

For positive set sizes,

$$d(P, C) = I \frac{|S_C| - |S_P|}{|S_P||S_C|}.$$

When $I > 0$, the prefactor is positive, proving Proposition 2. For $I = 0$, both coverages are zero; equal sizes also give a tie. Empty supports require an explicit abstention convention because the ratios are undefined. The proof concerns the sign only, not the numerical magnitude of the coverage difference.

A.2 Two-state observational ambiguity

Let $\mu = (1 - p_a - p_b)\delta_0 + p_a\delta_a + p_b\delta_{a+b}$. Under Model I, depth one has mass p_a and depth two mass p_b. Under Model II these depth masses are exchanged, and

$$[a + b, -b](1, 0)^\top = a + b, \qquad [a + b, -b](1, 1)^\top = a.$$

Both therefore induce μ. Both dictionaries have full rank because a, b are independent. Writing

$$T = \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}, \qquad D_{\mathrm{II}} = D_{\mathrm{I}} T,$$

exhibits an invertible, non-monomial transformation. Nonnegative codes do not require nonnegative decoder atoms.

Let $\theta = +1$ mean that the observed point a has depth one and $\theta = -1$ mean that $a + b$ has depth one. For a possibly randomized estimator based on n observations, let q be its probability of returning $+1$ under the common sample law $\mu^{\otimes n}$. Its risks in the two models are $1 - q$ and q, so their maximum is at least $1/2$. A fair randomized guess attains $1/2$ on this two-point family. This proves Theorem 3.

Applying a common measurable data-processing or randomized learning procedure to identically distributed samples preserves equality in law. Thus observationally computed marginal supports, conditional supports, or higher-order statistics do not identify which of these two generating explanations is true. Oracle latent codes, a dictionary supplied with identifying semantics, additional training data linked to the target, or interventions are different information settings. The result neither compares their risks nor claims that all real-world distributions are maximally ambiguous.

A.3 Counting exact chain representations

Let $X = [x_1, \ldots, x_L]$. Full rank of D implies full rank of X, because taking cumulative sums corresponds to multiplication by an invertible triangular matrix. For any permutation π, Eq. (9) defines a full-rank dictionary: it is obtained from X by a permutation and an invertible difference transform. Its depth-k prefix sum telescopes to $x_{\pi(k)}$. Assigning depth probability $p_{\pi(k)}$ gives the same observation law.

Conversely, a competing representation in Definition 1 has exactly L distinct nonzero prefix sums, all with positive mass. Equality of observation laws requires those sums to be precisely the given support points, so they have a unique order π. Each atom must be the difference of two successive prefix sums, fixing the dictionary uniquely. Two different orders yield different prefix sequences and hence different ordered dictionaries. There are exactly $L!$ orders, proving Theorem 4.

The word "ordered" matters: columns are indexed by depth, not counted after quotienting out arbitrary latent relabelings. The proof also excludes absent states, rank-deficient dictionaries, extra unused atoms, and noise. Enumeration is an illustration of this finite construction, not a substitute for the converse argument.

A.4 Frequency ordering and normalized L_1 costs

The first exact code uses one feature at x_a and two at x_b; the second exchanges these lengths. Equation (10) follows directly. For a general chain, cost is $\sum_k k p_{\pi(k)}$. If $i < j$ but $p_{\pi(i)} < p_{\pi(j)}$,

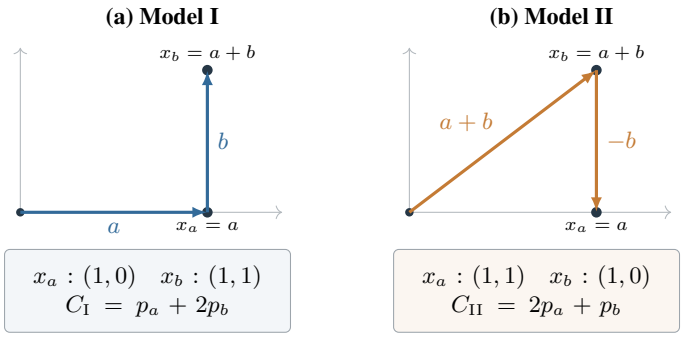

Figure 5: **Equivalent observations, different code lengths.** Both models reconstruct the same points with the same probabilities, while expected sparsity cost can prefer one ordering. The geometry is schematic.

swapping these two assignments changes cost by

$$(j-i)\big(p_{\pi(i)} - p_{\pi(j)}\big) < 0.$$

Removing all inversions sorts masses in descending order. Tied masses permit multiple minimizers. This proves Proposition 5 without making a statement about optimization dynamics.

For the normalized L_1 comparison, Model I has unit columns a, b and coefficients $(1,0)$ and $(1,1)$. Model II has unit columns $(a+b)/c, -b$ and coefficients $(c,1)$ and $(c,0)$. Consequently,

$$\mathbb{E}\|f\|_{1,\mathrm{I}} - \mathbb{E}\|f\|_{1,\mathrm{II}} = -cp_a + (2-c)p_b.$$

Since $0 < c < 2$ for $\rho \in (-1,1)$, the sign changes at $r = c/(2-c)$. At $\rho = 0$, this equals $\sqrt{2}/(2-\sqrt{2}) = 1+\sqrt{2}$. Without column normalization, arbitrary dictionary scaling changes coefficient L_1 costs, so this convention cannot be omitted.

A.5 A LOSSY SINGLE-ATOM COMPARISON

For unit reference atoms, set $s = a + b$. Projection of a onto the line spanned by s is $s/2$, since

$$\frac{\langle a, s\rangle}{\|s\|_2^2} = \frac{1+\rho}{2+2\rho} = \frac{1}{2}.$$

The squared residual is $\|a - s/2\|_2^2 = (1-\rho)/2$; the joint state s is reconstructed exactly. Both projection coefficients are positive, so the expected L_0 is $p_a + p_b$. The better exact candidate has expected L_0 equal to $p_a + p_b + \min(p_a, p_b)$ and zero error. Subtracting objectives gives Eq. (11). The calculation holds for this fixed single-atom direction and these two exact candidates. It does not optimize over every rank-one direction, encoder, or dictionary.

A.6 RECOVERY UNDER ADDITIONAL ASSUMPTIONS

Amplitude variation. Conditional on the parent-only population,

$$\Sigma_{\mathrm{parent}} = \mathrm{Var}(s_A)aa^\top.$$

Conditional on the joint population, independence gives

$$\Sigma_{\mathrm{joint}} = \mathrm{Var}(s_A)aa^\top + \mathrm{Var}(s_B)bb^\top.$$

The first matrix has rank one, and the second rank two by linear independence and positive variances. Their ranks identify which group is parent-only. With constant amplitudes, both within-group covariances instead vanish. With generic additive noise, rank ordering need not survive without a noise model. With unknown state grouping, the proposition does not supply a mixture-identification procedure.

Causal intervention. Under $Z_A = U_A$ and $Z_B = Z_A U_B$, the intervention $\mathrm{do}(Z_A = 0)$ replaces the first equation, after which the unchanged second equation yields $Z_B = 0$. Under $\mathrm{do}(Z_B = 0)$ the second equation is replaced, leaving $Z_A = U_A$ unchanged. The asymmetry follows from these structural equations. Logical implication alone specifies an observational constraint and does not determine the effects of interventions.

Known dictionary. For a full-rank known dictionary D, the coefficients are uniquely $D^\dagger x$, where $D^\dagger = (D^\top D)^{-1} D^\top$. Across the two nonzero states, the parent component is nonzero in both and the child-specific component only in the joint state. This statement concerns coefficients in a known basis; it makes no claim that the parent vector belongs to the intersection of the two one-dimensional spans of the observed vectors.

Unknown but orthogonal atoms. If $a \perp b$ and both are nonzero, the order $(a, a + b)$ has increments a, b with zero inner product. The reverse order has increments $a + b, -b$, whose inner product is $-\|b\|_2^2 < 0$. It is therefore inadmissible in the orthogonally restricted model. These arguments prove Proposition 6. They also show precisely why the unrestricted ambiguity theorem is not contradicted.

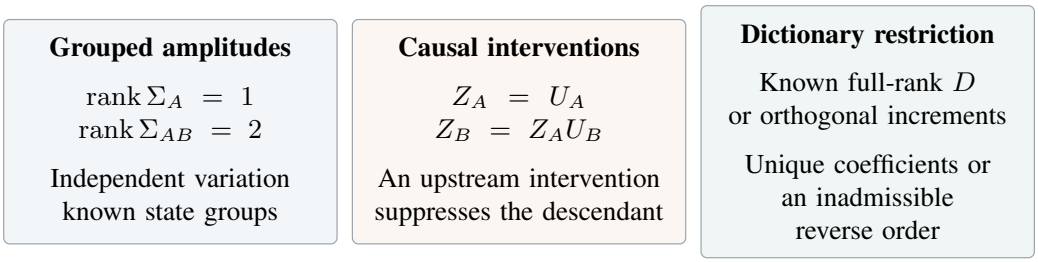

Figure 6: **Three distinct sufficient conditions, not one universal recovery algorithm.** This figure summarizes the assumptions used in Proposition 6 and its proof. The statements concern their formal settings; the real-data rank result in Table 10 remains near chance for $r_{\mathrm{corpus}} \geq 1$.

B Additional Empirical Results and Interpretation

B.1 Principal direction and layer results

Table 2 reports the principal run's method comparison. The 95% bootstrap CIs for Δ are $[1.8, 7.3]$ for M0, $[51.4, 56.2]$ for M1, $[29.3, 34.5]$ for M3, and $[53.6, 58.2]$ for mean token-L_0 direction. The raw count statistic is not necessarily identical to the aggregated support-size rule.

The reported layer contrasts are $+47.3$, $+53.8$, and $+38.3$ points at Llama layers 8, 16, and 24. These measurements show a non-monotone depth profile within one family. They do not identify a unique semantic layer. The permutation analysis establishes a departure from random label assignments, but a frequency-confounded rule can also depart from that null; the test does not by itself isolate hierarchy-specific information.

B.2 Full support-threshold grid

Table 5 gives all 15 cells of the separate support-construction sweep. The decision count varies with both thresholds, so comparison of accuracies across cells also changes the evaluated subset. The $K = 3, \tau = 0.10$ cell decides 5,471 pairs in this pass, three more than the principal extraction; the two runs are not pooled.

Table 5: **Full K–τ sensitivity grid.** Each row uses the same 20 templates but a separate encoding pass from the principal run. Accuracies condition on decided pairs; Δ is in percentage points.

τ	K	Decided	Overall	$r_{\text{corpus}} < 1$	$r_{\text{corpus}} \geq 1$	Δ
.05	1	5,528	46.38	54.58	30.48	24.10
.05	2	5,481	56.14	72.71	23.91	48.80
.05	3	5,445	59.61	77.40	25.00	52.40
.05	4	5,360	61.72	79.15	27.66	51.50
.05	5	5,310	61.68	78.10	29.35	48.76
.10	1	5,534	46.40	54.47	30.79	23.68
.10	2	5,472	56.80	73.38	24.54	48.84
.10	3	5,471	61.05	79.29	25.46	53.83
.10	4	5,424	62.59	80.53	27.48	53.05
.10	5	5,393	63.23	80.29	29.63	50.67
.15	1	5,532	46.02	53.73	31.10	22.62
.15	2	5,499	56.25	72.45	24.99	47.46
.15	3	5,466	60.74	78.87	25.30	53.57
.15	4	5,414	62.54	80.14	27.87	52.28
.15	5	5,395	62.98	79.84	29.54	50.31

Table 6: Reported control-analysis aggregates. "Slot agreement" for controls is not semantic direction accuracy. Neighborhood controls select related words but are not independently matched on semantic similarity.

Quantity	Real pairs	Random controls	Neighborhood controls
Pairs	5,572	55,700	26,404
Excess overlap	13.6 [13.5, 13.8]	9.0 [9.0, 9.0]	10.1 [10.1, 10.2]
Jaccard	.157 [.155, .159]	.099 [.098, .099]	.111 [.111, .112]
M1 accuracy / slot agreement (%)	61.1	61.7	61.3
Regime contrast (pp)	54.1	48.3	45.0

B.3 Control measurements

Reported intervals are 95% bootstrap intervals, with rounding inherited from the reported aggregates. Their narrowness does not remove dependence among pairs sharing words. Matching uses seed 42; multi-seed matching stability is not established.

For the real-versus-random comparison, standardized mean differences in log parent-slot frequency, log child-slot frequency, and log ratio are 0.057, 0.019, and 0.026. These balance diagnostics do not automatically apply to the neighborhood controls. Eligible-feature count is $M = 32{,}730$. The excess-overlap correction preserves set cardinalities but assumes uniform feature sampling; it does not control every marginal feature propensity.

The available direction logistic regression, clustered on the first word, reports hierarchy coefficient -0.069 (SE 0.060, $p = 0.258$) and log-ratio coefficient -1.913 (SE 0.034), with $N = 59{,}898$ and pseudo-$R^2 = 0.264$. These are model-specific conditional associations, not a proof that no estimator can extract incremental information from SAE representations.

The available real-versus-random excess-overlap regression reports a real-pair coefficient of 4.68 (SE 0.15), controlling both log word frequencies, with $N = 59{,}898$ and $R^2 = 0.139$. The separately reported real-versus-neighborhood adjusted contrast is 3.7 (95% CI $[3.4, 4.0]$). The effect size $d = 0.91$ belongs to the former comparison and must not be attached to the latter. Neither adjustment establishes equal semantic similarity between groups.

B.4 GATE AND WIDTH MEASUREMENTS

Table 7: Four-arm gate/width comparison at Llama layer 16. Mean L_0 is per token; support sizes aggregate templates. Cosine measures reconstruction direction and is not a complete reconstruction-error assessment.

Arm	Mean L_0	$\lvert S_P \rvert$	$\lvert S_C \rvert$	Δ (pp)	Acc. (%)	Cosine
32K JumpReLU	43.671	49.3	55.1	53.8	60.8	.7991
32K TopK ($K = 44$)	44.000	58.3	58.2	6.4	49.8	.8205
128K JumpReLU	40.370	49.0	52.3	47.8	56.5	.8067
128K TopK ($K = 40$)	40.000	54.5	54.9	5.9	52.7	.8264

The two gates share weights within a width. Selection by TopK is calibrated to the native mean activation count; it does not preserve the entire distribution of counts or the identity of selected features. The cross-width comparison uses separately trained published checkpoints with matching reported non-width hyperparameters.

Table 8: Cross-setting direction contrasts. These rows vary multiple factors and must not be interpreted as a gate-only or width-only experiment.

Setting	Gate	Δ (pp)	95% CI
Llama Scope L16, 32K	JumpReLU	53.8	$[51.4, 56.2]$
GPT-2 OpenAI L6, 32K	TopK	-28.8	$[-31.6, -26.0]$
GemmaScope L16, 128K	JumpReLU	0.7	$[-1.9, 3.3]$

B.5 CONDITIONAL SUPPORTS AND FEATURE SELECTION

The conditional-support experiment identifies a word's most frequently active feature across 30 templates. It assigns the same feature to both words in 3,721 of 5,572 pairs, leaving 1,851 evaluable pairs. This feature-assignment collision is specific to the raw argmax procedure; it is not a direct measurement of global dictionary absorption. The enrichment-based selection used elsewhere has a reported collision rate of only 0.4%, illustrating the dependence on the selection rule.

Table 9: Conditional-support results on the selected evaluable subset. Accuracy conventions and coverage differ from the main M1 evaluation.

Frequency bin	Pairs	Accuracy (%)	Non-tie accuracy (%)	Ties (%)
$r_{\text{corpus}} < 0.5$	1,099	78.6	83.0	5.3
$0.5 \leq r_{\text{corpus}} < 1$	207	57.0	61.5	7.2
$1 \leq r_{\text{corpus}} < 2$	185	37.3	39.4	5.4
$r_{\text{corpus}} \geq 2$	360	25.0	26.4	5.3

These statistics also vary with corpus-frequency regime. They do not contradict Theorem 3, which does not distinguish marginal from conditional statistics obtained by the same observational pipeline. Nor do they prove that adding corpus co-occurrence would necessarily identify a latent hierarchy without further assumptions.

B.6 FOCUSED-RANK READOUT

The separate 30-template analysis selects each word's top three enriched features and measures effective rank over the union of the selected subspaces. It decides 99.9% of pairs, versus 96.8% for M1. For $r_{\text{corpus}} \geq 1$, accuracy is 49.7% (95% CI $[47.4, 51.9]$) versus 26.7% ($[24.7, 28.8]$) for M1; for $r_{\text{corpus}} \geq 2$, it is 49.8% ($[47.0, 52.6]$) versus 20.7%. The rank readout uses more than support information and is not governed by the observation-only lower bound.

Table 10: **Focused effective rank in the separate 30-template analysis.** All accuracies are percentages. Δ uses the two broad frequency regimes; these results are not pooled with the main 20-template run.

Method	Overall	$r_{\text{corpus}} < 0.5$	$r_{\text{corpus}} \geq 2$	$r_{\text{corpus}} \geq 1$	Δ (pp)
M1: support orientation	57.0	76.8	20.7	26.7	45.9
R4: focused effective rank	57.9	64.1	49.8	49.7	12.4

B.7 Full-space amplitude statistics

Table 11: Additional 30-template amplitude/count predictors. Different descriptive signals need not share the same frequency dependence.

Method	Overall (%)	$r_{\text{corpus}} < 0.5$	$r_{\text{corpus}} \geq 2$	Reported Δ (pp)
Effective rank	44.7	53.3	29.9	17.4
Singular-value ratio	48.5	56.4	34.3	16.2
L_0 coefficient of variation	39.3	28.3	60.6	-24.4
Mean L_0	60.7	84.8	15.8	55.3

The direction of the frequency association changes for the coefficient of variation, but poor aggregate performance prevents treating that observation as a general recovery result. The focused rank experiment in Table 10 restricts the feature subspace and improves the high-ratio result toward chance. Population covariance rank in Proposition 6 and empirical effective rank of selected SAE features are different objects.