HIERARCHICAL GEOMETRY AND REFERENCE-DERIVED STEERING IN GEMMA-2-9B

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can represent hierarchical relationships geometrically, but the presence of this structure does not establish when it influences an answer. We examine this distinction in Gemma-2-9B across 46 noun hierarchies containing 690 concepts, tracing representations through all 42 transformer blocks. Hierarchical distance structure strengthens in early blocks, while reference-fitted principal components capture broad category distinctions more strongly than finer distinctions. We then add reference-derived category directions to individual concept-token states, excluding query leaves from their own reference fits and using the known target category to orient each edit. At B10, these interventions improve coarse and fine answer accuracy by 2.85 and 3.40 percentage points, respectively. Benefits diminish at later sites despite substantial retained category structure. In a paired cathedral example, a B34 edit strengthens the correct geometric preference while leaving the answer incorrect. State transfers identify the final prompt position as carrying a consequential difference from the successful early edit. Component interventions reveal contributions through this state from later attention and MLP computations, including B23 attention. These findings distinguish hierarchical representation from behavioral influence and show why evaluating geometric interventions requires tracing their effects through subsequent computation.

1 INTRODUCTION

Hierarchical categories connect specific concepts to broader classes. Understanding how language models represent and use these relationships can inform interpretation and targeted intervention. Geometric approaches have learned compact taxonomies in hyperbolic space (Nickel & Kiela, 2017), while structural probes have recovered syntactic tree distances from transformed contextual representations (Hewitt & Manning, 2019).

Park et al. (2025) relate categorical and hierarchical concepts to linear representations and orthogonal refinements under a specified causal inner product at the model's output interface. Nava & Wyart (2026) derive a statistical account in which word co-occurrence produces hierarchy-aligned eigenspaces with coarse-to-fine spectral organization. These accounts motivate examining how hierarchical organization changes across transformer depth and when its category directions influence the model's answers.

Geometric separation alone need not yield a correct prediction because an edit's influence depends on subsequent computation and the prediction position. Activation addition shows that contrastive directions can steer outputs (Turner et al., 2024), while activation-restoration experiments identify internal states contributing to predictions (Meng et al., 2022). Together, these approaches allow geometric organization and its behavioral consequences to be examined within the same task.

We study this connection in Gemma-2-9B using 46 WordNet hierarchies comprising 690 concepts. We follow native representations from the input through all 42 transformer blocks and final normalization, measuring agreement with tree structure and the concentration of broad and fine contrasts in reference-fitted principal directions. Using the known target category, we orient reference-derived concept-token edits and score complete coarse and fine answers. Queries remain excluded from their own reference fits. Paired state transfers and component interventions trace how an early edit affects the later answer-position state.

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.

Hierarchical distance structure strengthens beyond the input and is most pronounced in early internal layers, where category nudges also produce their largest observed accuracy gains. Broad contrasts concentrate more strongly in leading principal directions than finer distinctions. Substantial coarse concentration nevertheless remains late, and a late push can improve geometric category alignment without correcting the answer. Transfers identify the final colon in `Answer:` as a consequential state carrying the early intervention's effect. Attention at B23 contributes strongly to this tested pathway, alongside other attention and MLP components.

These findings connect hierarchy geometry to its behavioral influence across model depth. Stronger internal geometry supports further organization during processing while remaining compatible with a statistical origin in text. They motivate intervention designs that jointly consider category geometry and the computation carrying an edit to the prediction position.

2 METHODS

Hierarchies and representations. We constructed 46 disjoint binary noun hierarchies from WordNet 3.0 (Miller, 1995). Each contains 15 synsets at depths 0–3, with 1, 2, 4, and 8 nodes connected by 14 direct hypernym edges. Each nonroot node has one selected parent. Noun-sense counts and definition-based review determined fixed labels, with unresolved sense ties, instance relations, and named individuals excluded. Parent–child relations were reviewed and leaf ancestry was checked against competing categories. Synsets and normalized full labels occur once in the collection, yielding 690 concepts and 368 leaves.

We use pretrained Gemma-2-9B, with 42 blocks and hidden dimension $d = 3584$ (Gemma Team, 2024). Representations come from the final label subtoken at the scaled input E, complete block outputs $B_0, \ldots, B_{41}$, and final normalization N. Distance analyses average vectors from `The concept of {label}`, `The meaning of {label}`, and `A reference to {label}` before calculating distances. The additional input summary P averages all label-subtoken embeddings. Reference PCA and interventions use only the first template. Coordinates retain native scaling. Inference states are FP16, geometric calculations are FP64, and token log probabilities are computed in FP32 and summed in FP64.

Geometry and reference fits. Within each tree and stage, we correlate squared Euclidean distances with selected-tree path lengths using Spearman correlation over the 35 equal-depth pairs. Geometric quartet accuracy measures recovery of the reference split among the three pairings of each four-leaf set, using summed squared distances and fractional credit for exact ties. These measurements are averaged equally across trees. Centered, unwhitened PCA is fitted independently to 11 reference nodes per tree, fold, and stage. We measure the squared norm of category and sibling-leaf contrasts captured by each of the first ten supported PCs. Query leaves are projected only after fitting. A separate descriptive 3D illustration fits all 15 concepts' three-template means. Appendix A gives metric, normalization, and numerical-rank details.

Folds, prototypes, and category directions. The two fixed complementary folds take one reference and one query from each sibling-leaf pair. Each fold has four reference leaves and four queries, and every leaf is queried once. The seven internal nodes join the reference leaves for PCA. Queries never fit their own prototypes or axes. Let C_b denote broad category $b \in \{0, 1\}$, F_{bk} its fine categories, and r_{bk} the reference leaf below F_{bk}. Suppressing tree, fold, and stage indices, the single-template vectors $x_v \in \mathbb{R}^d$ define

$$\mu_b^C = \tfrac{1}{4}(x_{F_{b0}} + x_{F_{b1}} + x_{r_{b0}} + x_{r_{b1}}), \qquad \mu_{bk}^F = \tfrac{1}{2}(x_{F_{bk}} + x_{r_{bk}}), \tag{1}$$

$$d^C = \mu_0^C - \mu_1^C, \qquad d_b^F = \mu_{b0}^F - \mu_{b1}^F, \qquad u = d/\left\|d\right\|_2. \tag{2}$$

Figure 1(b) illustrates the fold and prototype construction. The coarse prototype gives equal weight to two fine-category nodes and two reference leaves. The fine prototype averages one fine-category node and its reference leaf. The root and broad nodes are not extra prototype terms. Category identity fixes the orientation.

Toward-correct activation addition. Contrastive activation addition supplies the methodological precedent (Turner et al., 2024; Rimsky et al., 2024). Our prototype weights and length rule are

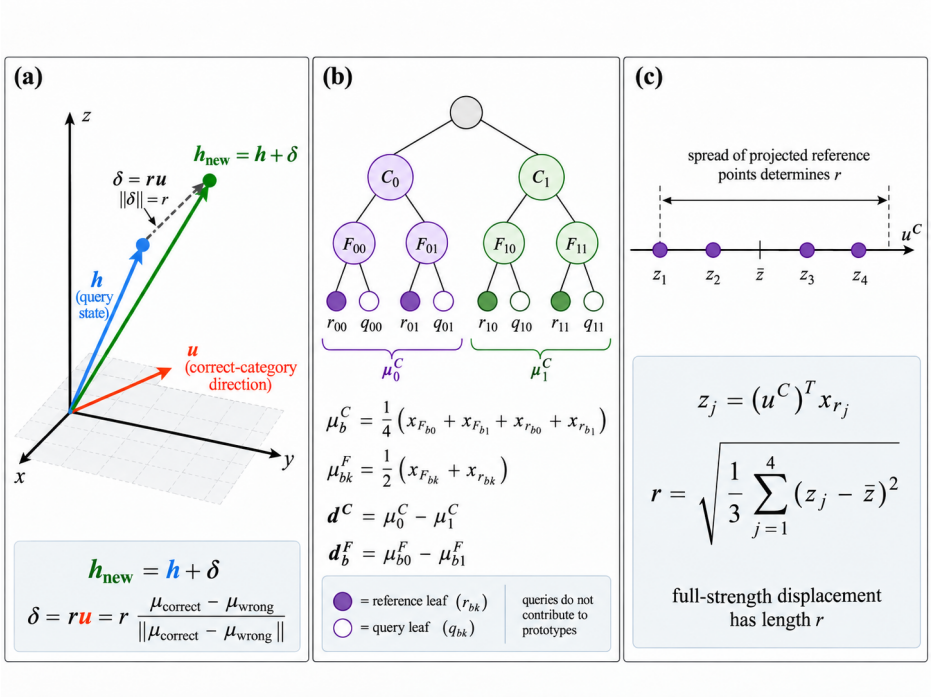

Figure 1: Reference-derived intervention. (a) Addition to the existing query state, with the figure's u already oriented toward the assigned category, corresponding to yu in Equation (4). (b) One reference fold, with filled reference leaves and open query leaves. Prototypes use the fine nodes and permitted reference leaves. (c) Reference projections determine the sample-standard-deviation length r, shared by coarse and fine edits. Positions are schematic, and lengths refer to the pre-rounding operation.

specific to this experiment. At each tree, fold, and site, both tasks use the sample standard deviation of the four reference-leaf projections onto the coarse direction,

$$z_j = (u^C)^\top x_{r_j}, \qquad \bar{z} = \tfrac{1}{4}\sum_{j=1}^{4} z_j, \qquad r = \sqrt{\tfrac{1}{3}\sum_{j=1}^{4}(z_j - \bar{z})^2}. \tag{3}$$

For the clean query state h and sign y toward its assigned category, the full-strength edit is

$$h_{\text{new}} = h + \delta, \qquad \delta = yr\,u = r\,\frac{\mu_{\text{correct}} - \mu_{\text{wrong}}}{\|\mu_{\text{correct}} - \mu_{\text{wrong}}\|_2}. \tag{4}$$

Figure 1(a) shows addition to the retained query vector, and panel (c) shows the length rule in Equation (3). Length r comes from references, not query-to-prototype distance. The known target determines the sign, making this answer-informed steering. A post-B_b edit changes only the final concept token and resumes at B_{b+1}, E is edited before block 0. Each edit starts from a clean run, with fixed parameters and other states unchanged at insertion. Post-cast displacements are checked. All 368 coarse leaves are supported. The fine analysis retains 24 leaves with undefined input contrasts at their baseline scores, leaving 344 pushed leaves within the same denominator. Appendix B specifies this fixed support rule.

Answer scoring and geometric boundary. Each leaf is queried with both option orders using the prompt

```
The concept of {leaf label}
Which category best describes this concept:  {option 1} or
{option 2}?
Answer:
```

Coarse choices are the two broad categories, fine choices are the two fine categories within the supplied correct broad branch. Each continuation is a leading space, the exact category label, and a final newline. We score the complete continuation without length normalization,

$$S(a \mid q) = \sum_{j=1}^{|a|} \log P_\theta(a_j \mid q, a_{<j}), \qquad m = S(a^+ \mid q) - S(a^- \mid q). \tag{5}$$

Positive margins receive accuracy credit 1, negative margins 0, and exact ties $1/2$. We average over orders and leaves within each tree, then equally over trees, giving 736 decisions per task. Corrections are strict negative-to-positive transitions and damage the reverse, tie transitions remain separate. A concept-only B_{41} edit cannot affect later prediction positions because subsequent operations are tokenwise.

To locate a query relative to its fixed prototypes, we use signed Euclidean distance to their separating hyperplane,

$$b(h) = \frac{\|h - \mu_{\text{wrong}}\|_2^2 - \|h - \mu_{\text{correct}}\|_2^2}{2\,\|\mu_{\text{correct}} - \mu_{\text{wrong}}\|_2}. \tag{6}$$

Positive b favors the assigned category geometrically. With $u_+ = yu$ pointing toward that category, any realized displacement a satisfies $b(h + a) - b(h) = a^\top u_+$, the ideal full push increases b by exactly r. This identity does not specify the model's answer rule or guarantee decreased absolute distance to the correct prototype.

State transfer and component tracing. Activation replacement tests which internal differences contribute to an answer (Meng et al., 2022; Zhang & Nanda, 2024). Let H^A and H^L be the full B34 states after ordinary concept pushes at B10 and B34, respectively. For a selected prompt-token set T, we form

$$H_j^{L \leftarrow A,T} = \begin{cases} H_j^A, & j \in T, \\ H_j^L, & j \notin T, \end{cases} \tag{7}$$

and score after B35 onward. Reverse replacement tests the corresponding loss in the early-push run. We test individual prompt tokens for cathedral and compare the final colon of `Answer:` with the strongest other single-token effect within each decision across all 46 trees. This colon is the prompt state predicting the first answer token, the outcome remains the complete-answer margin.

We then intervene on each attention or MLP residual write in B11–B34, after its post-module normalization and before residual addition. Interruption replaces an early-push write with the clean write, while restoration replaces a clean write with the early-push write. After recomputing through B34, only the resulting final-colon state is transferred into the fixed late-push recipient. If $\mathcal{M}_c(X)$ denotes this recipient's margin with colon state from source X, then

$$I_k = \mathcal{M}_c(A) - \mathcal{M}_c(A[k \leftarrow 0]), \qquad R_k = \mathcal{M}_c(0[k \leftarrow A]) - m_L. \tag{8}$$

Here 0 denotes a clean source, m_L the late recipient margin, and k the block and component. Positive I_k and R_k indicate loss on interruption and gain on restoration through this specified relay. The displayed trace replaces the component's write at the colon, Appendix D defines the original interval family, which also includes all-prompt writes. These are conditional pathway effects, not additive shares of the total correction.

Uncertainty and supporting analyses. All population estimates use equal tree weights and 1,000 paired whole-tree bootstrap draws, seed 20260910. Geometry and accuracy-gain intervals are pointwise. Component intervals are simultaneous over the full 384-endpoint family, including both token scopes. Geometry is compared with 200 depth-preserving node-assignment permutations shared across stages. Coarse and fine geometry–accuracy correlations use the mean unpushed geometry and post-nudge accuracy profiles over $E, B_0, \ldots, B_{40}$, recomputing both profiles in each bootstrap draw. They are descriptive associations. Sparse-autoencoder descriptions of ordinary-push changes for cathedral and their semantic-review procedure are confined to Appendix C, they do not supply the intervention direction.

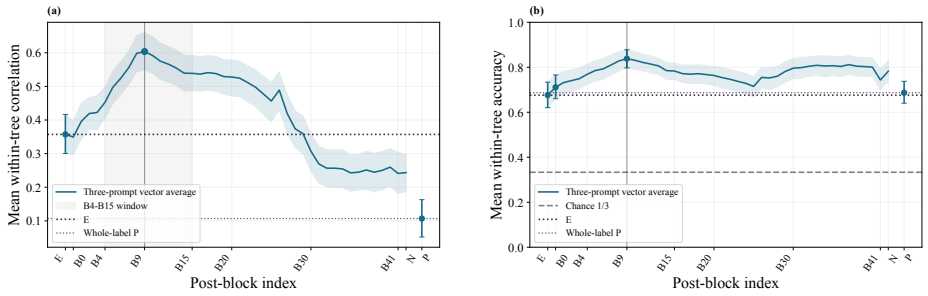

Figure 2: Hierarchy geometry across 46 trees. (a) Equal-depth Spearman correlation with tree distances. (b) Geometric quartet accuracy. Shading and error bars show pointwise 95% whole-tree bootstrap intervals. E and P are input summaries and N is final normalization. The highlighted window and observed maximum are descriptive.

3 Results

3.1 Hierarchy geometry strengthens early and remains detectable late

Mean equal-depth correlation increases from 0.357 at E to an observed maximum of 0.604 at B_9, with pointwise 95% interval $[0.549, 0.660]$, and is 0.592 at B_{10}. It declines to 0.241 at B_{41} and 0.244 at N, while the whole-label input average P is 0.107 (Figure 2). Quartet accuracy also peaks at B_9, rising from 67.61% at E to 83.82%. It remains 74.35% at B_{41} and 78.26% at N, showing that the late decline in distance correlation coexists with substantial branch grouping.

Both metrics exceed all 200 depth-preserving shuffled means at every stage, with Monte Carlo $p = 1/201$. Each contextual stage also exceeds the maximum-stage shuffled reference within its metric. This supports hierarchy-specific structure beyond the assignment control. The selected-stage values and intervals appear in Appendix Table 3.

3.2 Coarse distinctions concentrate more strongly in leading directions

Reference-fitted PCA separates the concentration patterns of broad, finer-category, and sibling-leaf distinctions. At B_9, PC1 captures 69.21% of the mean coarse contrast. At B_{10}, its first three PCs capture 87.77%, compared with 62.64% and 53.80% for the two fine-category contrasts (Table 1). Finer distinctions occupy additional directions rather than following a universal one-PC-per-level rule. Sibling-leaf contrasts are more dispersed, with only 39.12% and 38.89% contained in all ten reference PCs. Their remaining variation lies outside that reference span.

Table 1: Mean squared-contrast capture at B_{10} (%). All cells include 46 trees. Fine 1-a/b are within-branch category contrasts, Fine 2-a/b average the two sibling-leaf contrasts in each branch. Percentages measure contrast capture rather than explained variance. The full span contains reference-derived category contrasts by construction.

Contrast	PC1	PC2	PC3	PC1–3	PC1–10
Coarse	67.5	14.4	5.9	87.8	100.0
Fine 1-a	12.4	23.4	26.8	62.6	100.0
Fine 1-b	9.3	19.0	25.5	53.8	100.0
Fine 2-a	3.8	5.3	4.6	13.7	39.1
Fine 2-b	3.5	5.9	5.0	14.4	38.9

The invertebrate example makes this organization visible (Figure 3). Its B10 full-space equal-depth correlation is 0.780570, the separate all-concept 3D projection explains 56.61% of variance. Its reference-only coarse contrast has 90.70% capture in PC1, and PC2–PC3 capture 94.45% and 72.79% of its echinoderm and worm fine contrasts. The example illustrates a measured organization rather than establishing a universal three-dimensional hierarchy.

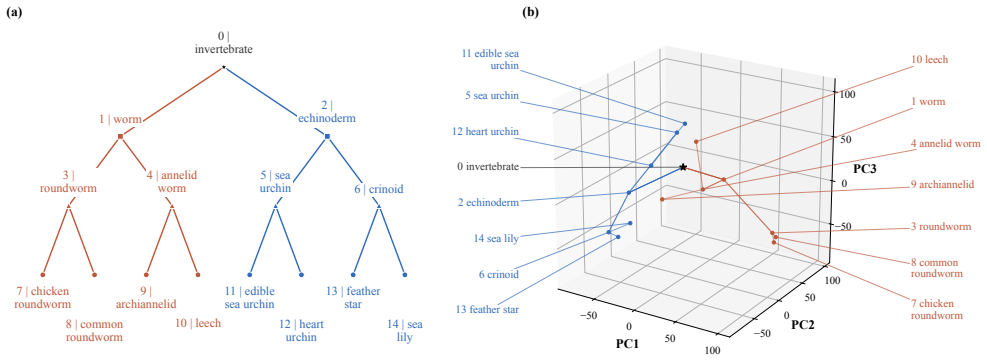

Figure 3: An individual hierarchy at B10. The supplied tree (a) and all-concept, unwhitened 3D PCA (b) share node numbers and branch colors. The projection retains 56.61% of variance, full-space hierarchy correlation is 0.780570. Edges are supplied, not inferred. The reference-only contrast analysis uses a separate fit.

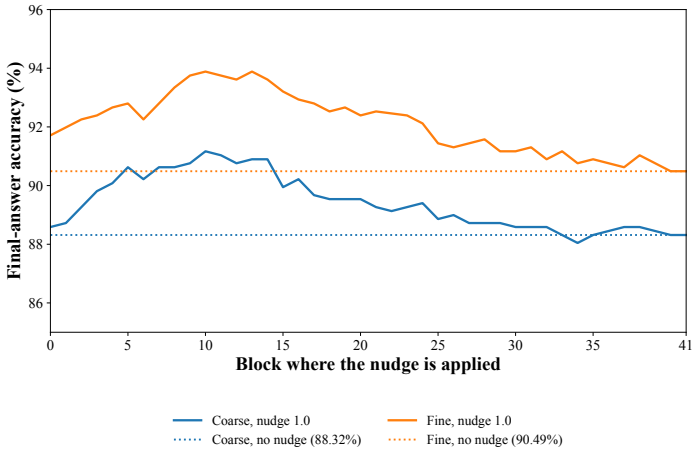

Figure 4: Final-answer accuracy under separate full-strength concept-token additions. Coarse and fine curves use both option orders and equal tree weights. Dotted lines show no-push baselines. Fine includes 48 fixed baseline-fallback decisions. Each point starts from the clean state, edits do not accumulate.

The all-stage heatmaps (Appendix Figures 9–10) show the prominent early organization around B8–B10 and its subsequent redistribution. Coarse concentration persists late. Across trees, PC1–PC3 retain 82.80% of the coarse contrast at B34, 81.28% at B41, and 81.04% at N. Thus, weaker late distance correlation does not mean the category distinction has disappeared. PC ranks remain local to each independently fitted basis.

3.3 The same category intervention has its largest gains near B10

The unmodified model answers 650 of 736 coarse decisions correctly (88.32%) and 666 fine decisions (90.49%). Full-strength B10 nudging increases coarse accuracy to 91.17%, with 21 corrections and no damage. Fine accuracy reaches 93.89% at both B10 and B13, through different transitions of 26 corrections and one damage at B10 versus 25 corrections and no damage at B13 (Table 2). These observed maxima identify a useful early region rather than a unique optimum.

Later gains diminish (Figure 4). B34 slightly reduces coarse accuracy, with one correction and three damaged decisions, while its fine gain is small. Both B34 gain intervals include zero. At B40, answer labels match baseline despite some margin changes, B41 leaves scores unchanged, as expected from its structural position.

Table 2: Selected answer outcomes, 736 decisions per row. Gain is in percentage points with pointwise paired-tree 95% intervals. C/D gives strict corrections and damage, listed rows have zero ties. Intervals do not adjust for selecting observed maxima.

Task	Condition	Accuracy (%)	Gain (pp), 95% interval	C/D
Coarse	No push	88.32	—	—
Coarse	B10	91.17	$+2.85\ [1.63, 4.21]$	21/0
Coarse	B34	88.04	$-0.27\ [-0.82, 0.27]$	1/3
Fine	No push	90.49	—	—
Fine	B10	93.89	$+3.40\ [2.17, 4.89]$	26/1
Fine	B13	93.89	$+3.40\ [2.17, 4.76]$	25/0
Fine	B34	90.76	$+0.27\ [-0.27, 0.82]$	3/1

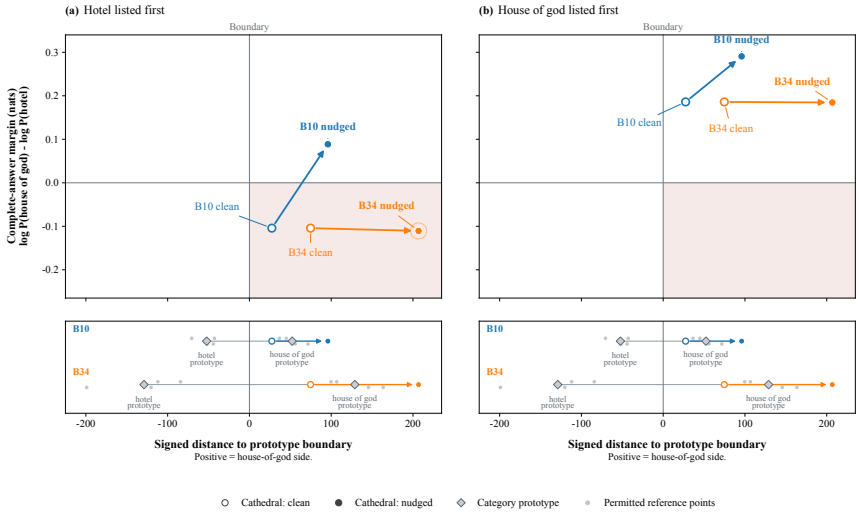

Figure 5: Local geometry and final answers for cathedral. Horizontal position is signed distance to the fixed prototype boundary, with positive values favoring house of god. Reference points and prototypes are shown separately below. Vertical position is the complete-answer margin. With hotel first (a), the B34 push moves farther onto the geometrically correct side while its answer margin remains negative. Reversing option order (b) changes answer scores despite the same concept geometry.

Across $E, B_0, \ldots, B_{40}$, the mean unpushed equal-depth tree-distance correlation has a strong Spearman association with post-nudge accuracy, with $R_C = 0.925$ (95% interval $[0.711, 0.944]$) and $R_F = 0.939$ ($[0.834, 0.946]$). The paired-tree bootstrap recomputes both mean profiles. These descriptive associations do not separate geometry from depth, edit length, or remaining computation. The full scatterplots are in Appendix Figure 8.

3.4 A GEOMETRICALLY FAVORABLE STATE NEED NOT CORRECT THE ANSWER

Cathedral provides a paired example using its original house-of-god-minus-hotel push (Figure 5). With hotel listed first, the B10 edit corrects the answer, while the B34 edit leaves a negative complete-answer margin. Cathedral is already on the house-of-god side of the B34 prototype boundary and moves farther onto that side after nudging. The failed answer therefore persists despite the intended improvement in this local geometric coordinate. This does not imply that the entire tree is perfectly represented or that nearest-prototype distance is Gemma's decision rule.

Both option orders have identical causal-prefix concept geometry, yet different answer margins, further limiting an explanation based on that state alone. At B34, the ordinary nudge also strongly increases several tentatively church-, building-, and religion-related SAE coordinates while the hotel-first answer remains incorrect, Appendix Tables 5–8 report these descriptive changes. Neither

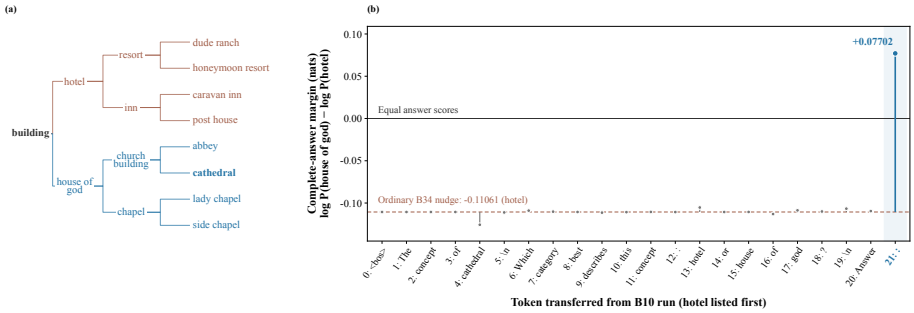

Figure 6: Individual-token transfers for cathedral. (a) The supplied building hierarchy. (b) Each point replaces one B34 prompt-token state in the late-push recipient by its early-push counterpart. Only the final colon transfer makes the complete-answer margin positive, reaching $+0.07702$ from -0.11061. Earlier tokens, including cathedral, do not individually restore the answer.

geometrically favorable position nor such feature activity alone establishes successful use at the prediction position.

3.5 The final prediction state carries a consequential early-push difference

We next transfer individual B34 prompt states from the B10-pushed run into the failed B34-pushed run. For cathedral's hotel-first question, the recipient margin is -0.11061. Every single-token transfer leaves it negative except replacement of the final colon in `Answer:`, which yields $+0.07702$ (Figure 6). This is a final margin, corresponding to a gain of 0.18763. Transferring the cathedral concept token alone does not restore the correction. The effective difference in this test is therefore carried at the prediction position by B34.

The population comparison supports a broader role for this position. Across all 46 trees, transferring the early-push colon into the late-push recipient increases mean margins by 0.16176 for coarse and 0.12347 for fine questions, with pointwise 95% intervals $[0.13527, 0.18973]$ and $[0.10342, 0.14458]$. Colon transfer corrects 17 coarse decisions while damaging one, and corrects 21 fine decisions while damaging one, relative to the B34-pushed recipient. Reversing the colon transfer lowers the early-push margin by 0.16145 and 0.12334. Relative to the strongest other single-token transfer within each decision, the colon gains are larger by 0.13285 for coarse and 0.09300 for fine questions, with simultaneous intervals $[0.10504, 0.16067]$ and $[0.06519, 0.12082]$ (Appendix D). The effect is concentrated at this tested position, but these interventions do not establish that it is the sole carrier of category information.

3.6 Later attention and MLP writes contribute through that state

Component interruption and restoration identify contributions to the B34 prediction-state effect (Figure 7). Among the tested colon writes in B11–B34, B23 attention has the largest mean contribution for both tasks. Its coarse interruption loss is 0.05355 with simultaneous 95% interval $[0.03755, 0.06955]$, and restoration gain is 0.05603 ($[0.04003, 0.07204]$). Fine effects are 0.03976 ($[0.02376, 0.05576]$) and 0.04051 ($[0.02451, 0.05651]$).

Additional contributions are distributed across depth. B25 and B28 attention and B26 MLP show positive effects, while B31 attention has opposing coarse effects. B23 is therefore a prominent contributor to the measured pathway, not a point at which all processing finishes. Only the modified source's colon state is relayed into the recipient, so the outcome concerns each component's influence through that specific state. Appendix D gives protocol details, the interval family, and reference-span directional controls.

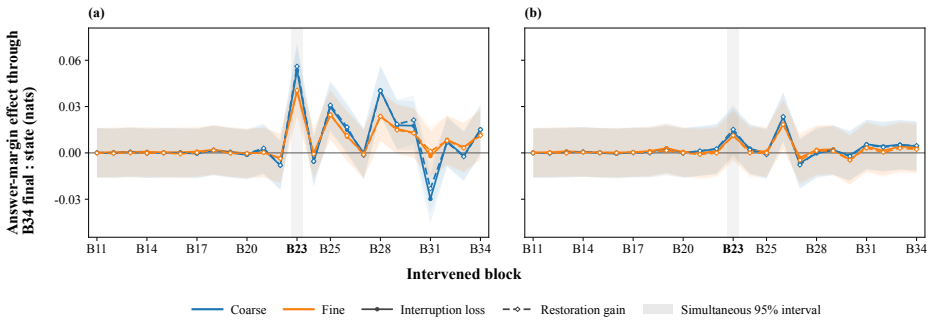

Figure 7: Component effects through the B34 final-colon state across 46 trees. Panels show attention (a) and MLP (b) writes at that token. Positive interruption loss means replacing an early-push write with its clean value reduces the relayed answer effect, positive restoration gain means inserting the early-push write into the clean source increases it. Intervals are simultaneous over the original 384 endpoints, including the all-prompt scope. Effects are conditional and need not add across components.

4 Discussion

The experiments connect measurable hierarchy geometry to the computation that makes a category edit useful. Early layers strengthen tree-distance structure and concentrate broad distinctions in leading directions. Applying the reference category contrast near B10 changes final answers more effectively than applying the same rule late. Yet the coarse contrast remains concentrated at B34, and cathedral can move farther onto its correct geometric side without changing its incorrect answer. These observations distinguish the presence of category structure from its effective influence on the prediction.

The transfer experiments provide a partial explanation of that dissociation. An early concept edit has time to change later prompt states, including the final state that predicts the answer. By B34, moving the concept token along its category contrast does not recreate those accumulated changes. Transferring the early run's final prediction state restores the cathedral correction and improves population margins, while reciprocal transfer removes part of the early benefit. Component interventions identify B23 attention and other attention and MLP writes as contributors through this state. The supported account is that distributed later processing contributes to the early edit's effect through a prediction state. It does not require loss of all hierarchy geometry or a single block containing the complete answer.

This account is compatible with co-occurrence-based origins of hierarchy geometry. Improved organization during forward computation adds a description of where such structure becomes behaviorally accessible, without isolating what training statistics produced it. For intervention design, the findings motivate considering both a reference-derived direction and the computational route by which an edit reaches prediction. They do not establish an autonomous correction method, since the intended category is supplied, or a universal best layer. Evidence comes from one model, fixed hierarchies and prompts, and inspected examples, depth, native edit size, and remaining computation are not independently matched. Finally, conditional state replacements identify contributions under the tested interventions, not a complete additive mechanism or uniquely necessary representation.

Future work should test whether these patterns extend to other models and previously unexamined hierarchies, while comparing intervention sites under matched displacement budgets. A further goal is to develop geometry-derived edits that improve answers without requiring the correct category in advance.

AI USE DISCLOSURE

The experimental design was developed by the authors. AI assistants supported code development and debugging, analysis of saved outputs, figure preparation, literature discovery, reference checking, and manuscript drafting and editing. The authors take responsibility for the final manuscript, experimental procedures, reported results, and citations.

REFERENCES

Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. arXiv preprint arXiv:2408.00118, 2024. URL https://arxiv.org/abs/2408.00118.

John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

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. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

Andres Nava and Matthieu Wyart. Hierarchical Concept Geometry in Language Models Emerges from Word Co-occurrence. arXiv preprint arXiv:2605.23821v1, 2026. URL https://arxiv.org/abs/2605.23821v1.

Maximilian Nickel and Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6338–6347. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

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*, pp. 76441–76463, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/be7430d22a4dae8516894e32f2fcc6db-Abstract-Conference.html.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation Addition: Steering Language Models Without Optimization. arXiv preprint arXiv:2308.10248v4, 2024. URL https://arxiv.org/abs/2308.10248v4.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

A GEOMETRY, PCA, AND PROFILE UNCERTAINTY

Distance measures. For each tree and stage, the three-template vectors are centered across the 15 concepts before their Gram is computed in FP64. This translation preserves squared distances. We compute $D_{uv} = G_{uu} + G_{vv} - 2G_{uv}$, clipping only negative roundoff after a tolerance check. Selected-tree distance counts unit hypernym edges. The equal-depth measure pools one pair at depth 1, six at depth 2, and 28 at depth 3, ranks the 35 pairs jointly using average ranks for ties, and correlates the ranks. Constant profiles remain undefined. All 46 trees have finite reported geometry values.

For each of the 70 four-leaf sets, the three representation costs are $D_{ab} + D_{cd}$, $D_{ac} + D_{bd}$, and $D_{ad} + D_{bc}$. The unique minimum of the corresponding tree-distance sums supplies the reference split. If k representation pairings tie for the minimum, credit is $1/k$ when the reference pairing is among them and zero otherwise. Scores average over quartets and then equally over trees. The chance reference is $1/3$.

Table 3: Selected geometry measurements from the saved 46-tree outputs. Intervals are pointwise 95% paired whole-tree bootstrap intervals. Quartet accuracy describes geometric grouping, not answer accuracy.

Stage	Equal-depth ρ	95% interval	Quartet (%)	95% interval
E	0.357	$[0.300, 0.417]$	67.61	$[62.07, 73.41]$
P	0.107	$[0.052, 0.163]$	68.70	$[64.04, 73.76]$
B_9	0.604	$[0.549, 0.660]$	83.82	$[79.75, 87.80]$
B_{10}	0.592	$[0.540, 0.648]$	83.14	$[79.10, 87.05]$
B_{41}	0.241	$[0.182, 0.303]$	74.35	$[70.25, 79.13]$
N	0.244	$[0.187, 0.296]$	78.26	$[73.66, 82.89]$

PCA and contrast capture. At each tree, fold, and stage, we center the 11 equally weighted reference vectors and diagonalize their small Gram $K = X_{R,c}X_{R,c}^\top$. If $K = U\,\mathrm{diag}(\eta)U^\top$, the unit native axes are $v_j = X_{R,c}^\top U_j/\sqrt{\eta_j}$. These are the axes of the sample covariance $X_{R,c}^\top X_{R,c}/10$, whose eigenvalues are $\eta_j/10$. The analysis computes the Gram rather than allocating the full covariance. Dividing by $\sqrt{\eta_j}$ constructs unit axes and does not whiten projected scores. Reference centering uses weight $1/11$, and query coefficients in the fitted axes are checked to be zero.

For the reference-derived contrasts d in Equation (2), the category heatmaps report

$$q_j^{\mathrm{cat}} = \frac{(v_j^\top d)^2}{\sum_{k=1}^{r_{\mathrm{PCA}}} (v_k^\top d)^2}. \tag{9}$$

The denominator is the full supported reference-span norm, checked against the native squared norm. Near-100% capture across the full span follows from reference construction, not compression. For an actual sibling-leaf pair (a, b), the denominator includes variation beyond that span,

$$q_{(a,b),j}^{\mathrm{leaf}} = \frac{[v_j^\top (x_a - x_b)]^2}{\|x_a - x_b\|_2^2}. \tag{10}$$

A query contributes to this descriptive measurement after fitting. The two leaf-pair percentages within a branch are averaged within each fold, then over folds within the tree, and finally equally over eligible trees. Category contrasts follow fold-then-tree averaging. Unavailable cells remain missing, and displayed axes are not renormalized. Cumulative values use common support.

Fine 1-a/b index the two fine-prototype contrasts in lexical broad-category order, Fine 2-a/b index the two branch summaries of sibling-leaf contrasts in saved node order. These conventions differ for some trees and do not imply a shared branch identity between suffixes. For invertebrate, Fine 1-a is echinoderm and Fine 1-b worm, while Fine 2-a is worm and Fine 2-b echinoderm. All five families are sibling distinctions, rather than parent–child displacements.

Numerical rank includes Gram eigenvalues above $10^{-10}\eta_1$. Adjacent eigenvalues differing by at most $10^{-12} + 10^{-8}\eta_1$ form tied groups, whose individual axes are marked unavailable. Zero contrasts

remain undefined. Of 2,200 pooled stage/family/PC cells, 49 have reduced support, all at E, every other cell includes all 46 trees. The 3D illustration instead fits the 15 three-template means, and its variance ratio $\sum_{j=1}^{3} \eta_j / \sum_j \eta_j$ is distinct from contrast capture.

Bootstrap and shuffled references. A fixed seed of 20260910 generates 1,000 paired samples of 46 whole trees with replacement. Each selected tree retains all its leaves, folds, orders, and stage measurements. Percentile intervals use the 2.5th and 97.5th percentiles with linear interpolation. This preserves dependence within trees, stages are not independently resampled. Reported geometry and accuracy-gain intervals are pointwise, including those at inspected profile maxima.

For geometric controls, 200 assignments permute node membership within each depth and tree, reusing assignments across all stages and both metrics. For observed mean $\bar{q}_s$, the one-sided Monte Carlo value is $(1 + \#\{k \mid \bar{q}^*_{s,k} \geq \bar{q}_s\})/201$. The contextual comparison uses the maximum permuted mean across B0–B41 and N, separately for each metric. E and P are outside this maximum-stage family. These shuffled assignments are geometric references, not unrun behavioral interventions.

The profile associations use mean equal-depth geometry and absolute post-nudge accuracy at E and B0–B40. Both profiles are rounded to 12 decimal places before average-rank Spearman correlation. Every bootstrap draw recomputes both means and their correlation. B41 is excluded as a concept-only structural no-effect endpoint, P and N have geometry but no corresponding local accuracy intervention.

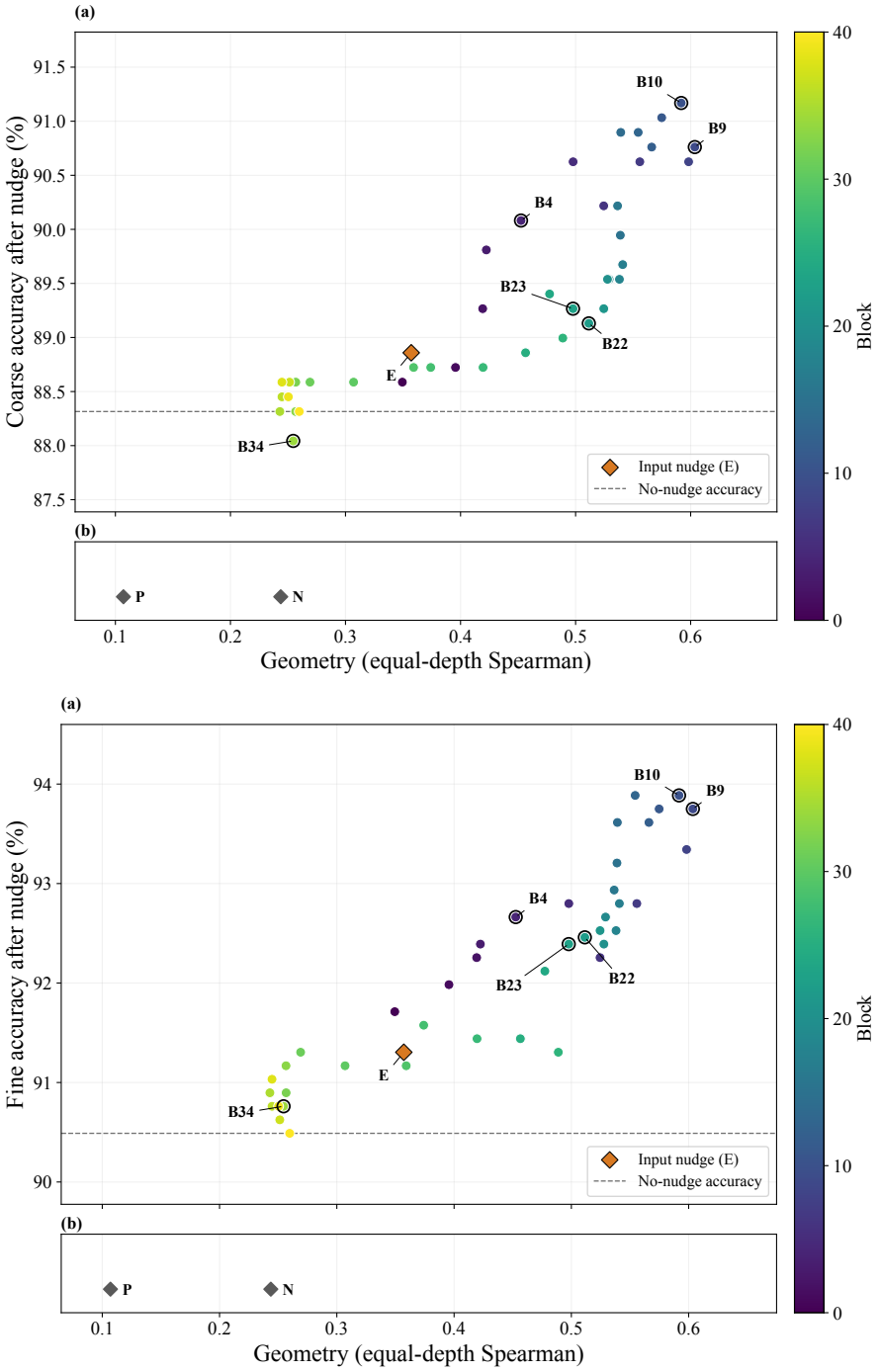

Figure 8: Geometry–accuracy associations, coarse above and fine below. Points are means across the same 46 trees, not independent leaves. The horizontal variable is the same clean hierarchy measure in both panels. Diamonds mark the input edit, P and N appear only as geometric references. The profile correlations are 0.925 ($[0.711, 0.944]$) and 0.939 ($[0.834, 0.946]$).

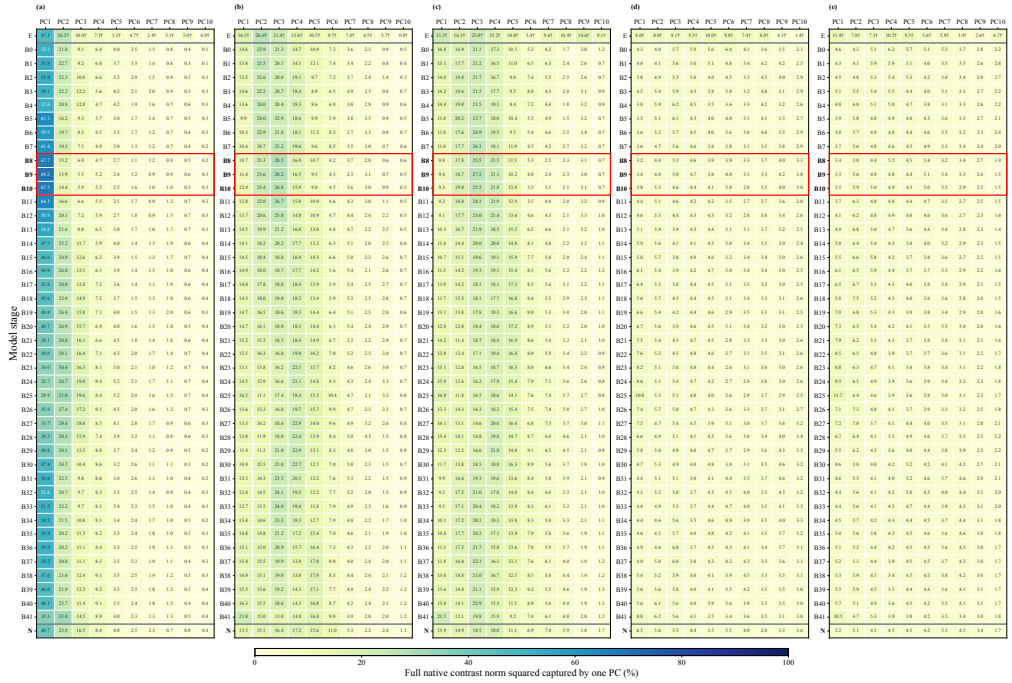

Figure 9: Complete 46-tree contrast spectrum in portrait layout. Rows are E, B0–B41, and N, columns are PC1–PC10. The panels show coarse, Fine 1-a/b, and Fine 2-a/b contrasts. Category capture is normalized by its supported reference-span norm, verified against the native norm, sibling-leaf capture uses the full native norm. Daggers indicate reduced eligible-tree support. The red window marks B8–B10 descriptively. The vector figure retains all cell values for magnification.

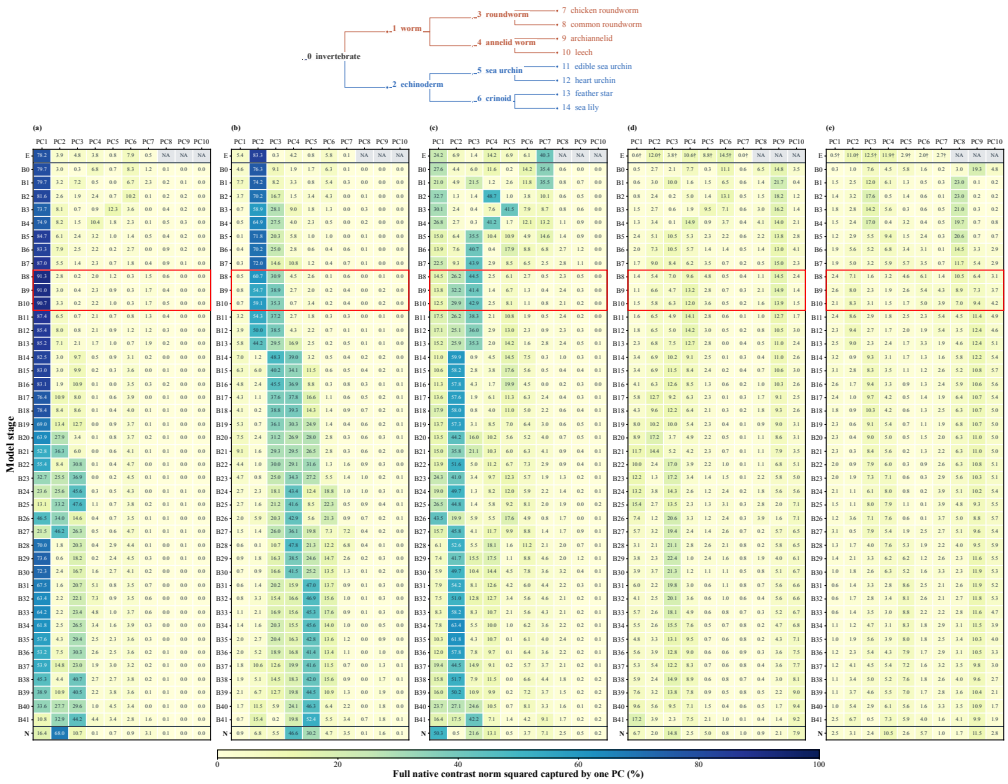

Figure 10: Invertebrate spectrum, with the supplied tree and all stages. Fine 1-a/b denote echinoderm/worm contrasts, while Fine 2-a/b summarize worm/echinoderm sibling leaves. At B10, coarse PC1 capture is 90.70%, PC2–PC3 capture 94.45% of the echinoderm fine contrast and 72.79% of the worm contrast. At N, coarse capture shifts to 16.39% in PC1 and 67.98% in PC2. Gray cells and daggers mark unavailable or partial support. These fits differ from the all-concept 3D visualization.

B INTERVENTION SUPPORT AND NUMERICAL VALIDATION

The frozen Gemma-2-9B revision is `33c193028431c2fde6c6e51f29e6f17b60cbfac6`. Native inference uses FP16 states, eager attention, evaluation mode, and no KV cache. The final subtoken is identified from the exact label's token offsets, rather than a fixed position across concepts. Fixed padded batches and masks preserve the recorded scorer. Complete-answer scoring includes the model's final normalization, tied output head, and final-logit softcap, as well as all continuation tokens and the newline. Candidate scores are kept in correct/competing order independently of displayed option order.

Directions and lengths are finite and nonzero above 10^{-10} at both E and the insertion site. For block edits, the concept disturbance induced by an input nudge must also be nonzero when it reaches the block. This propagated disturbance is an eligibility check, not the direction of the local push. Coarse support is 368 leaves at every site. Fine support is 344, the remaining 24 have zero input fine contrast and keep baseline scores at all sites. They contribute 48 of the 736 decisions per fine condition. Later-site fine directions can exist for these leaves even though the fixed eligibility rule excludes their edits.

Each block intervention clones the clean block-output state, forms the concept edit in FP64, and casts it for delivery. The reference casting path is $Q_{16}(Q_{32}(h^{64} + \delta^{64}))$. Realized displacement can therefore differ from the ideal $yr\,u$. Checks compare delivered states, unedited entries, finiteness, prefix equivalence, direction norms, and replayed complete-answer scores. State tolerances are relative L2 0.005 and maximum absolute difference 0.1, score comparisons use absolute tolerance 0.05, relative tolerance 0.001, and strict margin difference below 0.005. Delivered-norm checks use $0.05 + 0.01\,\|\delta\|_2$. These numerical checks are separate from statistical uncertainty and do not imply bitwise reproducibility across devices.

Independent site edits use clean prefixes and resume the remaining suffix rather than accumulating nudges. At B41, later normalization and output mapping act tokenwise, so a concept-only edit cannot change the scored answer positions. Saved B41 score differences are exactly zero. At B40, unchanged answer labels do not imply identical answer margins.

Boundary interpretation for cathedral. Table 4 records the plotted signed positions and complete-answer margins. Prototype contributors are resort, inn, dude ranch, and caravan inn for hotel, and church building, chapel, abbey, and side chapel for house of god, with weight $1/4$ each. Cathedral is excluded. The saved building tree is the supplied taxonomic construction, context-dependent readings of chapel nodes limit generalizing the illustration to every ordinary-language use. This caveat does not change the measured state-transfer result for cathedral.

Table 4: Exact cathedral comparison, rounded from saved values. Signed positions are native distance units and favor house of god when positive. Margins are nats and favor that answer when positive. The clean margin is shared between sites. B10 and B34 positions use separately defined axes and are not a cross-layer trajectory.

Quantity	Condition	B10	B34
$b(h)$	Clean	27.43259	74.76694
$b(h)$	Full push	95.92471	206.83383
Hotel prototype position	Fixed	-52.28646	-128.90538
House-of-god prototype position	Fixed	$+52.28646$	$+128.90538$
Margin, hotel first	Clean	-0.10422	-0.10422
Margin, hotel first	Full push	$+0.08860$	-0.11061
Margin, house of god first	Clean	$+0.18592$	$+0.18592$
Margin, house of god first	Full push	$+0.29065$	$+0.18463$

C DESCRIPTIVE SAE CHANGES FOR CATHEDRAL UNDER THE ORDINARY NUDGE

We encode the saved final cathedral-token residual state before and after the original full-strength ordinary nudge using the released Gemma Scope residual SAEs (Lieberum et al., 2024). B10 uses `layer_10/width_16k/average_l0_113` and B34 uses `layer_34/width_16k/average_l0_114`, from `google/gemma-scope-9b-pt-res`, pinned to revision `f9b689815814972562d28082f9f7d65d7e01fdc8`. The released FP32 JumpReLU encoding is

$$a = xW_{\text{enc}} + b_{\text{enc}}, \qquad z = \text{ReLU}(a) \odot \mathbf{1}\{a > \theta\}, \qquad \widehat{x} = zW_{\text{dec}} + b_{\text{dec}}. \tag{11}$$

Input bias subtraction, unit normalization, and additional scaling are not applied. For each block, we rank $\Delta_i = z_{i,\text{full}} - z_{i,\text{clean}}$ across all 16,384 coordinates and retain the ten largest increases and ten largest decreases. These are signed activation changes, not fold-change significance tests. They concern the ordinary full-strength state, not a smaller movement-matched edit or a gradient intervention.

Recovering tentative meanings from feature identities. Every lookup uses the block, exact SAE folder, and feature ID together. We reviewed corresponding Neuronpedia records, comparing existing GPT-4o-mini descriptions with high-activation natural-text contexts, peak tokens, and activation values. Exact duplicate token windows were collapsed, three paraphrased examples per feature were retained with provenance, giving 120 retained examples for the 40 listed IDs. Descriptions were summarized from those examples rather than inferred from cathedral's activation alone. Mixed and uncertain meanings remain marked. The public records exposed no explanation-validation scores, so the descriptions are not treated as validated semantic labels.

Public repository and folder metadata matched the pinned local SAEs, but the API exposed no checkpoint hash or revision, preventing independent verification of public decoder-byte identity. Features may respond to several contexts, and retained examples can share documents. These constraints apply even when a label is intuitively related to churches. IDs and activation magnitudes are specific to each SAE and cannot be equated across blocks.

Tables 5–8 preserve all 40 IDs, tentative descriptions, and saved activations. A zero denotes an inactive coordinate in that state, not absence of an entire concept. The reconstruction remainder $x - \widehat{x}$ is not assigned a semantic feature label. The tables show changes accompanying the ordinary interventions, without establishing that any listed feature causes the answer or exhausts the representation. B10 corrects the hotel-first answer, whereas B34 does not, despite increases in several church-related coordinates at both sites.

Table 5: B10 ten largest increases under the full ordinary nudge. Values are SAE activation units, rounded to three decimals, $\Delta = \text{full} - \text{clean}$.

Feature ID	Tentative semantic meaning	No push	Full push	Δ
2197	Church architecture	25.697	41.061	+15.364
6074	Church as community / institution	8.842	20.192	+11.349
6718	Mass / church services	0.000	10.174	+10.174
6447	Buildings being built / architectural history	5.289	12.757	+7.468
11454	Pilgrimage / religious places, mixed	2.682	10.144	+7.462
6135	Religion / religious institutions	9.550	16.890	+7.340
5167	Notices / letters, mixed	0.000	7.161	+7.161
13643	Buildings	0.000	7.104	+7.104
12011	Rooms / interior spaces	0.000	6.983	+6.983
6978	Saints / religious celebrations	0.000	6.723	+6.723

Table 6: B10 ten largest decreases under the full ordinary nudge. All listed coordinates become inactive. Meanings remain tentative.

Feature ID	Tentative semantic meaning	No push	Full push	Δ
854	Unclear / mixed	3.883	0.000	−3.883
3935	Possibly phrases such as give rise / play host	3.768	0.000	−3.768
15966	Mostly text-boundary marker, uncertain	3.475	0.000	−3.475
10301	Ferries / ferry operations	3.421	0.000	−3.421
12889	Unclear / mixed word fragments	3.330	0.000	−3.330
13079	Schools / educational institutions	3.296	0.000	−3.296
4341	Crops / raw materials	3.193	0.000	−3.193
15296	Large size / magnitude	2.943	0.000	−2.943
1516	Unclear / mixed	2.875	0.000	−2.875
9256	Projects / organizations, mixed	2.837	0.000	−2.837

Table 7: B34 ten largest increases under the full ordinary nudge. Several church-related coordinates increase even though the hotel-first answer remains incorrect.

Feature ID	Tentative semantic meaning	No push	Full push	Δ
5384	Church services / community	38.624	62.446	+23.822
1819	Church wording, mainly Scientology	51.655	71.343	+19.687
10447	Church architecture / clergy history	58.033	77.290	+19.257
3746	Medical / personal-name fragments, mixed	0.000	15.561	+15.561
8869	Liberty / charity words and names, mixed	0.000	15.487	+15.487
12153	Panel / label wording, mixed	0.000	14.467	+14.467
7809	Building word, literal and figurative	9.800	23.371	+13.571
8042	Chairs / desks in narratives	0.000	12.659	+12.659
584	Music / instrument contexts	0.000	12.373	+12.373
3529	Christian teaching / sermons	0.000	12.150	+12.150

Table 8: B34 ten largest decreases under the full ordinary nudge. Nine become inactive, feature 2625 remains active. Feature IDs refer only to the stated block-specific SAE.

Feature ID	Tentative semantic meaning	No push	Full push	Δ
13349	Leadership styles / management	10.452	0.000	−10.452
6210	Mathematical terminology	9.203	0.000	−9.203
10637	Everyday / travel contexts, unclear	9.180	0.000	−9.180
8642	Career / education contexts, tentative	8.737	0.000	−8.737
13539	Name fragments, unclear	8.286	0.000	−8.286
7895	Mixed historical / biological contexts, unclear	7.944	0.000	−7.944
6054	Operating levers / machinery	7.670	0.000	−7.670
10911	Mixed technical / historical topics, unclear	7.544	0.000	−7.544
8284	Code definitions / macros	7.372	0.000	−7.372
2625	Detection wording across domains	14.581	9.242	−5.339

D PREDICTION-STATE TRANSFERS AND COMPONENT DIAGNOSTICS

Matched population and token sets. All 46 trees, 368 leaves, and both answer orders are retained. Fine support is 344 pushed leaves with the 24 fixed baseline-fallback leaves kept in the denominator. Cases are not restricted to B10 corrections or B34 failures. The donor has an ordinary full-strength post-B10 concept edit, the recipient has an ordinary full-strength post-B34 concept edit. Both are evaluated at B34 using the same numerical setup before state exchange. All other states remain those of the recipient at the exchange boundary. Reverse exchange replaces early-run states with late-run counterparts. Both complete continuations are rescored after the remaining blocks, final normalization, and output head.

For each presented-order decision, exhaustive single-token transfer considers every actual prompt token, including the options. If $T_j = m(L \leftarrow A, \{j\}) - m(L)$ and $U_j = m(A) - m(A \leftarrow L, \{j\})$, the primary localization contrasts are $T_c - \max_{j \neq c} T_j$ and $U_c - \max_{j \neq c} U_j$, where c is the final colon. Maxima are computed within each decision before averaging. Table 9 gives the estimates. Colon transfer strictly exceeds the other-token maximum for 555/736 coarse decisions and 481/736 fine decisions, so average dominance does not imply dominance in every case.

For cathedral's hotel-first prompt, the final colon is token position 21, ID 235292. Its transfer yields margin $+0.07701874$, compared with recipient -0.11061096 and early donor $+0.08859730$. The gain is 0.18762970, and the donor margin is not reproduced exactly. Every other individually transferred prompt token leaves a negative margin. This experiment identifies a consequential state for the tested correction, not proof that semantic information exists only at that position.

Table 9: All46 final-colon transfer and reciprocal-undo effects in nats. Absolute effects use pointwise 95% intervals. Differences from the strongest other token use simultaneous intervals across the four primary task/action contrasts.

Task	Endpoint	Effect	95% interval
Coarse	Transfer gain	0.16176	[0.13527, 0.18973]
Coarse	Undo loss	0.16145	[0.13494, 0.18889]
Fine	Transfer gain	0.12347	[0.10342, 0.14458]
Fine	Undo loss	0.12334	[0.10335, 0.14491]
Coarse	Transfer minus strongest other token	0.13285	[0.10504, 0.16067]
Coarse	Undo minus strongest other token	0.13216	[0.10434, 0.15997]
Fine	Transfer minus strongest other token	0.09300	[0.06519, 0.12082]
Fine	Undo minus strongest other token	0.09286	[0.06504, 0.12068]

Component writes and the colon relay. At each B11–B34 block, attention or MLP output is intercepted after its post-module normalization and before residual addition. The residual skip remains intact. The two scopes are the final colon alone and all real prompt positions. Interruption starts from the early-push source and substitutes the corresponding clean write. Restoration starts from a separate clean source and substitutes the early-push write. It is not an algebraic reversal of interruption. Remaining computation is rerun through B34, only the resulting colon is then inserted into the unchanged ordinary B34 recipient.

Let $m_{\mathrm{relay},A}$ be the recipient margin after the unmodified early source's colon is transferred. Interruption loss subtracts the interrupted relay margin from $m_{\mathrm{relay},A}$. Restoration gain subtracts the ordinary B34 recipient margin m_L from the restored relay margin. The recipient's colon is otherwise clean because its local concept edit occurs after B34, which also explains the clean-colon baseline. The final outcome still sums the complete answer, rather than reading a single logit at the patched state.

The two-panel main figure displays 192 colon endpoints with the original simultaneous intervals across 384 endpoints, comprising 24 blocks, two components, two scopes, two actions, and two tasks. Figure 7 shows the colon-specific trace. B23 attention has the largest displayed positive means, but the intervals are not pairwise tests of a uniquely best block. For B26 MLP, interruption/restoration means are $0.02323/0.02345$ coarse and $0.01847/0.01838$ fine, with positive simultaneous intervals. B31 attention has negative coarse interruption -0.02979 ($[-0.04579, -0.01379]$) and restoration -0.02310 ($[-0.03911, -0.00710]$). B23 MLP intervals include zero.

Along-category and sideways controls. The saved expanded control includes all 46 trees. At B10, four random unit directions are constructed in the span of the 11 centered reference vectors and orthogonalized to the task's category contrast. They use the same nominal displacement length as the category edit. The intended category coordinate is preserved for sideways edits within the native-precision tolerance, the rule and rounding correction use reference geometry, not answer outcomes. Orthogonality to one contrast does not preserve all pairwise distances, all semantic properties, or whole-tree geometry.

The same colon relay and B23 attention interruption/restoration compare the along-category edit with the mean of four sideways outcomes. Controls are averaged within observations before orders, leaves, and trees, not treated as independent samples. B23 was selected in the earlier 33-tree analysis and held fixed for the additional 13 trees. The pooled 46-tree control therefore extends coverage but is not independent confirmation of that selection. All six along-minus-sideways contrasts have positive simultaneous intervals (Table 10). Mean sideways total-relay effects are negative, -0.06968 coarse and -0.03954 fine, so the contrasts reflect both beneficial along-category edits and harmful sideways edits. Random or orthogonal edits are not assumed behaviorally inert.

Table 10: All46 along-category minus mean-four-sideways effects in nats. One simultaneous 95% family covers all six endpoints. The controls share the reference span and nominal length while differing in category-axis alignment.

Task	Endpoint	Difference	95% interval
Coarse	Total colon relay	0.23143	$[0.18786, 0.27500]$
Coarse	B23 attention interruption	0.08354	$[0.03997, 0.12711]$
Coarse	B23 attention restoration	0.08274	$[0.03917, 0.12631]$
Fine	Total colon relay	0.16301	$[0.11944, 0.20658]$
Fine	B23 attention interruption	0.05567	$[0.01210, 0.09924]$
Fine	B23 attention restoration	0.05573	$[0.01216, 0.09930]$

Paired uncertainty and completion checks. All pathway draws resample 46 whole trees jointly, 1,000 times with seed 20260910. For simultaneous intervals over a specified family, we take the 95th percentile of the bootstrap maximum absolute centered error and add/subtract that common width from every estimate. The four token-localization contrasts have width 0.02781716, the 384 component endpoints 0.01600244, and the six directional-control contrasts 0.04357076. These are separate inferential families. They are not individual-token, order-level, or block-level resamples.

Saved token-transfer, component-trace, and directional-control caches have complete numerical and analysis flags, all 46 retained tree IDs, empty exclusion lists, and complete scoring masks. Identity checks preserve tree, leaf, fold, task, order, and site mappings. Numerical checks cover clean/source replay, causal-prefix equality across continuations, finite outputs, unchanged unpatched positions and residual skips, and correct boundary placement. The report reuses completed scores and introduces no new inference.