CONTEXT IS KING: HOW IN-CONTEXT SPECIFICATION SHAPES THE GEOMETRY OF CONCEPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models place structured concepts on geometrically faithful manifolds: weekdays lie on a circle, months on another, often read as a fixed world-model the network stores and looks up. Prior work shows that context can affect entity-token geometry while a strong pretrained prior remains prominent. We show where context takes over: at the final prompt token, after the model has read the rule and query, geometry follows the in-context specification and dominates the prior. A declarative rule changes not just the ordering of entities but the geometry's relational topology: the same tokens form a cycle or a hierarchy with tree depth and branch relationships on command. This change cannot be explained by simple relabeling and holds even for arbitrary nouns randomly assigned to the structure's nodes. Under conflicting specifications, capable models in both tested families (Gemma, Qwen) show representational similarity of 0.6–0.9 to the imposed structure versus near-zero to the pretrained order, across the concept sets tested. To investigate the computation underlying these geometric findings, we use activation patching and find that replacing one entity's activation with another redirects the answer to the donor's successor under the context-specified order. Sweeping the intervention across layers localizes where entity identity enters this computation. Smaller models can already form rough context-specified geometry, while larger models more consistently suppress the pretrained order and follow the active relation under patching. This family-dependent trend cautions against transferring mechanistic conclusions across model sizes. Whether the model builds this geometry anew or reconfigures a stored one remains open; operationally, the geometry it uses is the one the context specifies. Code and artifacts: https://github.com/anonauthoriclr/context-is-king.

1 INTRODUCTION

The geometric structure of concept representations is drawing increasing attention in interpretability research. A prominent answer is the linear representation hypothesis: individual features can be encoded along directions in activation space (Gurnee & Tegmark, 2024; Arditi et al., 2024). Work has since expanded from individual features to how *related* concepts sit together: weekdays and months form circles, while categorical and hierarchical concepts form simplices and polytopes (Engels et al., 2025; Park et al., 2025b). Such structure is naturally read as a stored world-model: geometry learned during pretraining and retrieved when a concept is invoked.

Park et al. (2025a) show that models exposed to random walks on a graph develop entity-token representations that reflect its structure. When that structure conflicts with strong semantic priors, it appears in the representations but does not dominate the pretrained geometry. We identify a transition across token positions. Under a conflicting declarative order, the representation at the entity's own token retains the pretrained structure, whereas the final prompt token, after the rule and query have been read, represents the imposed structure (Figure 1). Changing the specification from a cycle to a hierarchy changes this downstream state's relational topology, not merely its ordering (Section 5.2).

This bears on a basic question about how language models represent concepts: is a concept's relational structure an intrinsic property the model stores and retrieves, or is it assembled from the context on demand? Prior work on concept geometry and causally validated world models moti-

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.

vates a stored-world-model reading: the recovered geometry is learned in pretraining and available to downstream computation (Engels et al., 2025; Gurnee & Tegmark, 2024; Park et al., 2025b; Li et al., 2023; Nanda et al., 2023). Our results complicate that picture: the geometry a model *uses* is the one the context specifies, and the stored structure, though real, gives way the moment the context contradicts it. This goes beyond behavioral instruction-following: the geometry's organization changes on command, forms even for arbitrary nouns with no shared relational order, and activation patching localizes where entity identity enters the context-specified successor computation. We establish this in two parts: context sets the geometry the model *represents* (Section 5), and entity identity is causally bound into the context-specified relation (Section 6). Our contributions:

1. *We identify where context overrides pretrained geometry.* Under conflict, pretrained structure persists at the entity token, while the final prompt state follows the specified relations. Activation patching localizes where entity identity enters the context-dependent successor computation (Sections 5.1 and 6).
2. *We show that context changes relational topology, not merely order.* The same tokens form either a cycle or a hierarchy with depth and branch structure, even when the tokens are arbitrary nouns (Section 5.2).
3. *We distinguish geometry formation from reliable use across model sizes.* Smaller models can form rough context-specified geometry, while larger models more consistently suppress the pretrained order and follow the active relation under patching. This family-dependent pattern cautions against transferring mechanistic conclusions across sizes (Section 6.2).

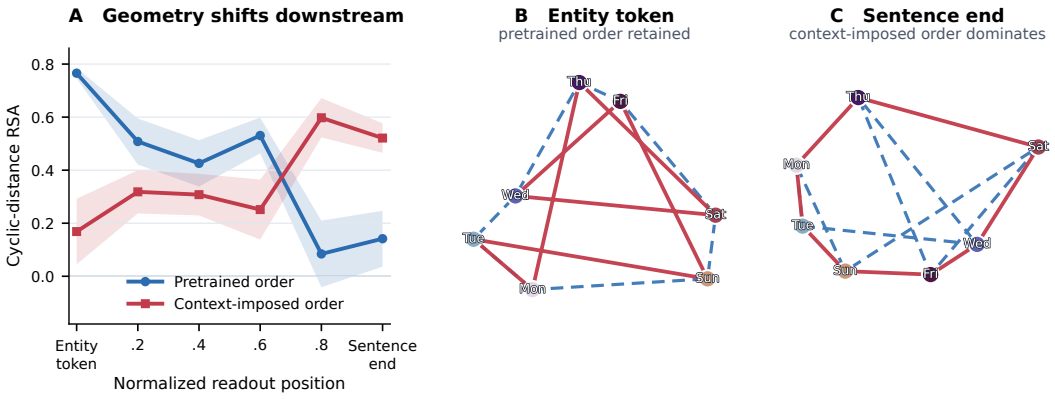


Figure 1: **Context takes over downstream.** Under a conflicting weekday redefinition, pretrained geometry persists at the entity token but gives way to the imposed order by the final pre-generation state. **A:** Mean representational similarity analysis (RSA), with 95% CIs over 20 random weekday orders. **B,C:** Entity centroids at the entity token and final prompt token, respectively, for one illustrative order, traced in the pretrained (blue dashed) and imposed (red solid) orders. Gemma-31B.

2 Background and Related Work

We connect our question of how context shapes relational geometry to five lines of work.

Geometry of concepts. Periodic concepts occupy low-dimensional cyclic manifolds (Engels et al., 2025), spatial and temporal properties are linearly decodable (Gurnee & Tegmark, 2024), and categorical/hierarchical concepts form simplices and polytopes (Park et al., 2025b). These are readouts of pretrained structure. Individual directions can causally mediate behavior (Arditi et al., 2024), while manifold steering ties curved representational geometry to behavior (Wurgaft et al., 2026). We ask whether context can override this pretrained relational structure by specifying conflicting relations among the same entities.

Context-induced geometry. Park et al. (2025a) teach in-context graphs through random-walk exemplars and find contextual structure in entity-token representations, while the weekday prior

remains prominent in the leading principal components. We recover that persistence at the entity token but identify a downstream transition: after the model integrates the relation and query, the pre-generation state is dominated by the imposed structure. Thus the pretrained structure can persist at the entity token while the final prompt state represents the context-specified relations. Our declarative setup further separates this transition from exemplar learning, varies the specified relation between cycle and hierarchy, and causally localizes identity-to-relation binding. Complementary work studies geometric changes during prediction and classification. Hosseini et al. (2026) find that neural trajectories straighten as contextual information accumulates, correlating with improved prediction. Xiong et al. (2026) find that in-context examples reorganize representations so that classes become more separable. We instead examine how an explicit relational specification overrides a conflicting pretrained relation.

Causally-validated world models. In Othello-GPT, a language model trained on sequences of Othello moves, probes recover the game's board state from internal representations, and targeted edits to those representations alter predicted moves (Li et al., 2023; Nanda et al., 2023).

Activation patching. Another way to test the causal role of internal representations is activation patching: replacing internal states with those from another run and measuring the effect on predictions (Zhang & Nanda, 2024). Kim (2025) uses this method to distinguish task-level information from specific input–output bindings. We apply entity-substitution patching to orders specified in the prompt that conflict with pretrained relations. The answer follows the donor entity's successor under the specified order, ruling out a context-insensitive successor lookup. Sweeping intervention depth and model size reveals where this context-dependent binding occurs and how consistently it appears across scales.

Scale and emergence. Emergence debates (Wei et al., 2022; Schaeffer et al., 2023) turn on thresholded benchmark scores. We compare representational geometry and activation patching across model sizes: smaller models already show rough context-specified geometry, while larger models more consistently suppress the pretrained order and follow the specified relation under patching (Section 6.2). This echoes, in the representation, the gap Lepori et al. (2026) report between encoding a structure and deploying it, and bears on whether circuit-analysis methods scale to larger models (Lieberum et al., 2023).

3 Framework

We frame the paper around one question: can a context-specified relational geometry override the pretrained one, and if so, where in the computation does it take over? This section defines the objects the question is about and the space of specifications we impose; how we *measure* the geometry is deferred to Section 4.

3.1 Definitions

Entities. Individual items, such as weekdays, months, or arbitrary nouns. We denote the entity set by $E = \{e_1, \ldots, e_N\}$; for weekdays, $E = \{\text{Monday}, \ldots, \text{Sunday}\}$.

In-context specification. A declarative rule, given without exemplars, that specifies relations among entities as either an order ("the only valid order is: Wednesday, Monday, ...") or a hierarchy ("P is the parent of A and B"). In conflict conditions, it also states that the normal relations do not apply.

Concept geometry. For each entity e_i, let c_i be its activation vector averaged across queries at a fixed layer and token position. The pairwise dissimilarities $D_{ij} = d(c_i, c_j)$ characterize the entities' representational geometry. We distinguish open, cyclic, and hierarchical structure. We characterize cycles by loop topology and hierarchies by depth and branch relations.

The geometry the model *uses*. We use this phrase operationally for geometry measured at the final prompt token, immediately before generation, where the relation and query have been integrated into the state consumed by next-token prediction. The entity-token state provides a contrasting readout and can retain the pretrained order (Figure 1). Activation patching separately tests how entity identity enters the context-conditioned successor computation (Section 6.1).

Congruent vs. conflicting specification. A specification is congruent when its relations agree with the pretrained structure and conflicting when they contradict it. Conflict directly tests whether the model follows the stored or stated relation.

3.2 Specification taxonomy

Every experiment is one point in a common space: impose structure X on entities Y, congruent or conflicting, in surface form Z:

- **Structure type:** an *order* (open line, or closed cycle once a wrap relation is stated) or a *hierarchy* (binary tree).
- **Relation to the prior:** *congruent* or *conflicting*. A *scramble* is a fresh random permutation of the order.
- **Entities:** concepts spanning strong cyclic priors (weekdays, months, clock hours), a weak cyclic prior (zodiac), a non-cyclic prior (musical notes, linear by pitch), and no relevant prior (arbitrary tokens with no shared relational prior). Section 5.1 quantifies these differences.
- **Surface form:** a numbered **list** or **adjacency** edges, with or without the **wrap** edge that closes a loop; form $\times$ wrap (a 2×2) separates *structure* from *surface form* (Figure 2).

Our central test imposes a *conflicting order* on a familiar cyclic concept. Section 5.2 extends the analysis to other specifications and entity sets.

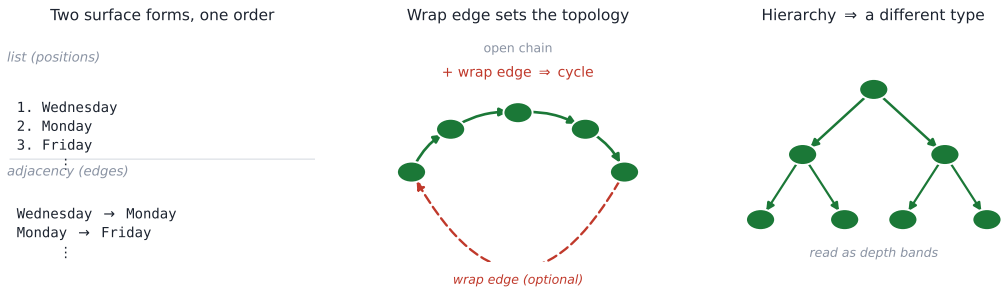


Figure 2: **What a specification is.** The same order stated as a numbered list or as adjacency edges (left); adding a *wrap* edge closes the chain into a cycle (middle; the open, no-wrap case is a weaker baseline); a declarative *hierarchy* is a different structure type, read as depth layers (right).

4 Experimental Setup

Here we fix how we *measure* the geometry: the probe, the prompt regimes, the models, and the sampling. Its causal counterpart, activation patching, is presented with its results in Section 6.1.

Relational-geometry measure. Representational similarity analysis (RSA) compares the model's pairwise entity distances with those predicted by a candidate relational structure (Kriegeskorte et al., 2008). For each entity e_i, we average its final-prompt-token residual across query templates at one late layer to obtain a centroid c_i, preventing any one phrasing from defining the geometry. The templates range from k-step queries (e.g., "1 step after Monday") to bare mentions (Appendix A). After mean-centering, $\tilde{c}_i = c_i - \bar{c}$, we define the model dissimilarity as $D_{ij} = 1 - \cos(\tilde{c}_i, \tilde{c}_j)$. The RSA score is the Spearman correlation between these dissimilarities and target distances T_{ij}; for a cycle, $T_{ij} = \min(|i-j|,\ N-|i-j|)$:

$$\mathrm{RSA} = \rho_{\mathrm{Spearman}}\big(\{D_{ij}\}_{i<j},\ \{T_{ij}\}_{i<j}\big).$$

Mean-centering is essential: the residual stream is anisotropic, and subtracting the shared direction is what exposes the ordering (Appendix H). The specific layer, $L = \mathrm{round}(0.75\, n_{\mathrm{layers}})$, is chosen from a full-layer sweep that shows a broad late-network plateau, so the result does not hinge on the exact choice (Appendix B).

Why RSA. Concept-geometry work usually reads structure off PCA plots or fitted probes and SAEs (Engels et al., 2025; Gurnee & Tegmark, 2024); we instead score the representational dissimilarity matrix (RDM) against a *theory-specified* template with a permutation null, the testable form of what such work shows by eye. Across eight models, higher RSA accompanies fewer crossings in two-dimensional PCA plots and higher successor accuracy. A separate persistent-homology analysis supports the distinction between cyclic and hierarchical specifications (Appendix I).

Regimes. We read geometry from the state that generates the behavior, the residual stream at the last prompt token, before generation, under three prompt regimes that share the rule and query and differ in the response instruction and thinking mode (Appendix A):

- *direct*: answer immediately;
- *scripted reasoning*: we instruct the model to follow the specified order one step at a time before giving its answer;
- *free-form*: the model reasons in its native thinking mode, without an added reasoning instruction.

The imposed-order structure forms in all three (Section 5.1).

Models. We study eight instruction-tuned models across three families and roughly an order of magnitude in scale: Gemma-4 (E2B, E4B, 12B, 31B), Qwen-3.5 (4B, 9B, 27B), and Llama-3.1-8B as a cross-architecture capability floor. The scaling claim spans all eight; but because a clean, usable map forms only with scale (Section 6.2), our main analyses (causal patching, robustness controls) concentrate on the two large models that reliably build it, Gemma-31B and Qwen-27B. For the causal analysis we additionally sweep the full Gemma and Qwen ladders (2B–31B and 4B–27B) to locate where the effect emerges with scale. Throughout, a paired value a/b is Gemma-31B / Qwen-27B.

Sample sizes and intervals. Geometry RSA is averaged over 10 scrambles per concept, the imposed hierarchy over 6, each with per-scramble 95% confidence intervals. For the causal arm we report per-scramble values. Activation patching in the direct regime uses 20 imposed orders $\times$ 42 ordered recipient–donor entity pairs for each of Gemma-31B and Qwen-27B, plus a fixed natural-order control. Scale-ladder and chain-of-thought sample sizes are reported in Appendix E.

5 Context Sets the Geometry the Model Uses

Our representational evidence is twofold: the geometry the model uses follows the stated relations rather than the pretrained prior (Section 5.1), and the specification replaces loop topology with a hierarchy encoding tree depth and branch relationships (Section 5.2).

5.1 Geometry Follows the Specification, With or Without a Prior

Under a conflicting specification, the measured relational geometry follows the imposed order while the pretrained organization is suppressed (Figure 3). In Gemma-31B and Qwen-27B, imposed-order RSA reaches $+0.87/+0.67$, while natural-order RSA is $-0.03/-0.01$. Under free-form prompting, imposed-order RSA still exceeds natural-order RSA by 0.58 in Gemma-31B. Imposed-order RSA is high across all five concepts, regardless of their pretrained-prior strength, and weaker in Qwen than Gemma. This dominance does not hold uniformly across model sizes: Qwen-9B retains the natural weekday ring (imposed-order RSA $+0.10$ versus natural-order RSA $+0.44$), while Llama-8B does not recover the imposed structure. As a control, we verified that natural weekday geometry remains significant at the same final-token readout without a relational specification in Gemma-31B and Qwen-27B (Appendix K).

The overridden prior itself varies widely: with *no* specification the natural ring is strong for weekdays/months/clock ($\approx$ 0.6–0.8), weak for zodiac ($\approx$ 0.3), and negative for musical notes (a linear pitch scale). Context imposes its structure across this whole range (Figure 4 and Table 7), and on arbitrary tokens forms the specified structure without a matching prior (Table 9); whether the induced shape is a ring or a tree is set by the specification (Section 5.2).

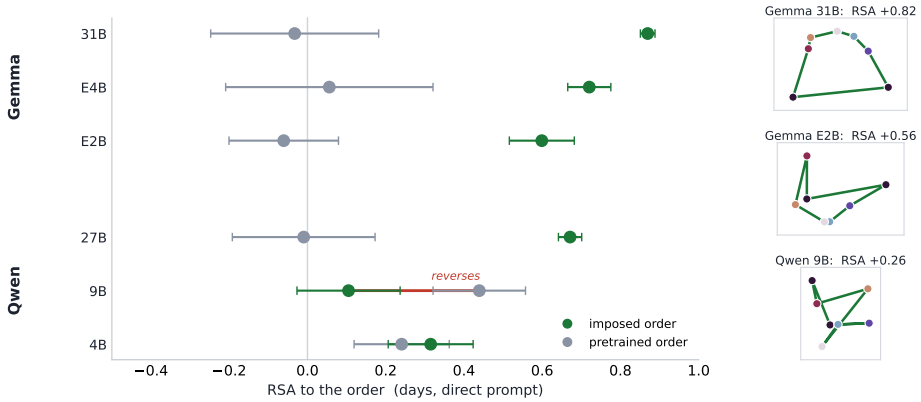


Figure 3: Imposed-order RSA (green) vs. natural-order RSA (gray) per model in the same probed state (*left*; days, direct prompt, per-scramble 95% CIs), and imposed-order geometry for three models (*right*; annotated RSA). Per-model values in Appendix G.

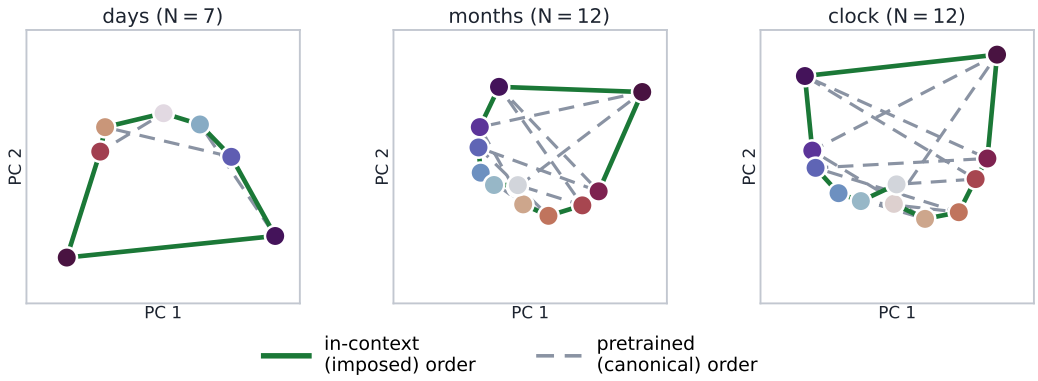


Figure 4: Entity centroids for three cyclic concepts under a conflicting in-context order (Gemma-31B; 2D PCA of the last-token pre-generation state; one representative scramble, colored by imposed position). Each path is traced in the pretrained order (dashed) and the imposed order (solid). Imposed vs. pretrained RSA (full residual space, not the 2D projection shown; mean over 10 scrambles): days $+0.87/-0.03$, months $+0.81/+0.05$, clock hours $+0.82/-0.05$.

Artifact controls. The dominance survives the checks that could show it inflated by the metric or summoned by our question (full procedures in Appendices F to H):

- **Fixed geometry with new labels.** A high imposed-order score could arise if the points stayed fixed while only their labels changed. Instead, geometry matches the natural order under a natural specification and the imposed order under a scrambled specification; among all 5040 weekday orders, the imposed one fits best (Appendix H).
- **Residual-stream anisotropy.** A narrow activation cone could produce positive RSA without relational structure. With the measured activations held fixed, the imposed order yields significantly higher RSA than expected for a randomly permuted order, across all ten scrambles, five concepts, and both large models ($p < 0.001$ for the twelve-point concepts and $p \leq 0.008$ for the seven-point concepts; Appendix H).
- **Prompt position.** The geometry could encode where items appear in the prompt rather than their stated relations. It does not: hierarchy depth appears although it is never listed explicitly, while the first and last items of an imposed cycle become adjacent despite being far apart in the list (Section 5.2 and Appendix F).

- **Query-induced structure.** The structural question itself could summon the ordering. Yet the ordering appears under bare mentions and stories, remains stable across queries asking for one to six steps ahead, and emerges only after the model reads the redefinition (Table 6 and Figs. 1 and 9).

5.2 The Specification Changes Relational Topology

Section 5.1 showed that context determines which order the geometry follows. Here we ask whether it can change a stronger property: the representation's topological signature. Relabeling a fixed ring can reorder tokens, but it cannot remove the loop and create hierarchical depth and branch structure. We test this in three stages: an imposed hierarchy, a cycle-to-hierarchy switch on the same tokens, and a cycle-closure test.

An imposed hierarchy on arbitrary tokens. A declarative binary tree on arbitrary tokens forms a depth-stratified manifold: under neutral queries the centroid RDM tracks node depth with RSA near $+0.9$ at depth-2, holding to $+0.73/+0.63$ (Gemma/Qwen) at depth-4 (Figure 5A,B). Same-depth nodes band together.

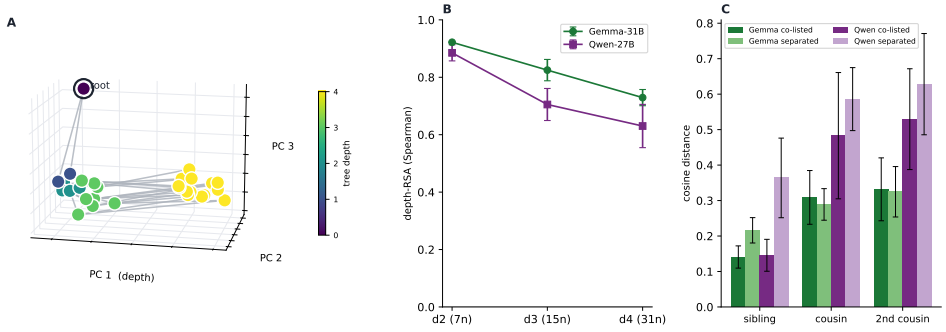


Figure 5: **Imposed hierarchy on arbitrary tokens under neutral queries.** *(A)* Entity-centroid PCA for a depth-4 tree (31 nodes; color = depth; depth-RSA $+0.73$). *(B)* Depth-RSA from depth-2 to depth-4. *(C)* Distances between sibling, cousin, and more distantly related leaves in the depth-3 tree; lighter bars show a control that lists parent–child links separately and in shuffled order. Error bars in B and C are 95% CIs across six scrambles.

Distances also reflect branch relationships: among leaves at the same depth, siblings are closer than cousins, and more distant relatives are farther apart (Figure 5C). This does not establish a faithful tree embedding. A co-occurrence control preserves depth-RSA but shows that siblings' proximity partly reflects being listed together in the prompt (Figure 5C; Appendix A).

The hierarchy is not induced by the structural query or inherited from token similarity. The hierarchy remains strong even under neutral queries ("Consider the item e"), reaching depth-RSA near $+0.9$ without structural questions such as "Who is the parent of e?" (see Appendix A for further examples). Across 12 randomized token-to-node assignments, token pairs become closer when assigned as siblings than when not. The effect is also capability-gated: the 2B model does not form the hierarchy. Different queries shift the tree's location in activation space while largely preserving its relational organization (Appendix G).

Same entities, different shape on command. We hold the *entities* fixed and vary only the declared relation. At the same final pre-generation readout, the seven weekdays follow their natural cycle with no relational specification (RSA $+0.57/+0.60$). An imposed cycle aligns them to the specified cycle (RSA $+0.72/+0.71$), whereas an imposed tree stratifies the same days by depth (depth-RSA $+0.67/+0.68$, cyclic structure gone); both effects hold in Gemma-31B and Qwen-27B (Figure 6). The tokens never change, so the specification rather than the concept changes the representation's topological signature. A complementary persistent-homology analysis is reported in Appendix I. An imposed *line* opens the ring only partially and capability-dependently (Appendix K).

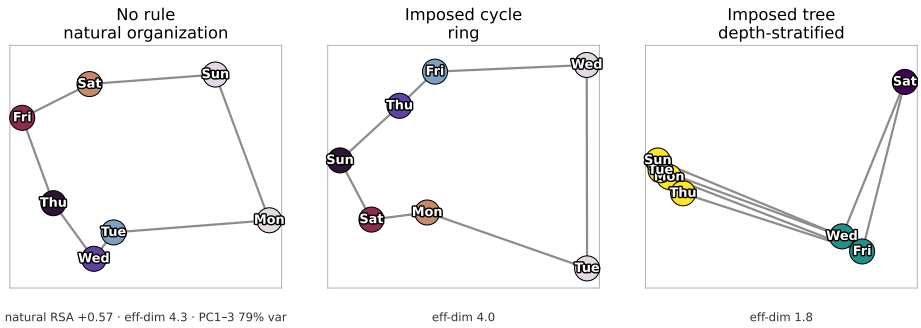


Figure 6: Seven weekdays under no rule, an imposed cycle, and an imposed tree (Gemma-31B; orthographic views of the first three PCs). All use the final pre-generation readout. The no-rule panel averages queries without an imposed order; imposed panels show one scramble. Statistics are computed in full space; cross-family values appear in Appendix K.

Additional analyses show that explicitly stating the last-to-first wrap relation makes the geometry more ring-like than leaving it unstated, with controls for prompt format and endpoint co-occurrence (Appendix F).

6 Context-Specified Relational Binding Is Causally Localized and Scale-Gated

Section 5 showed that context sets the geometry represented at the pre-generation state; this geometry could still be an epiphenomenal signature. We next use activation patching to test a narrower causal claim: whether replacing an entity's representation redirects the answer to the replacement entity's successor under the specified relation, and how this effect varies with model size (Sections 6.1 and 6.2); the full intervention controls are detailed in Appendix E.

6.1 Activation Patching Redirects the Answer

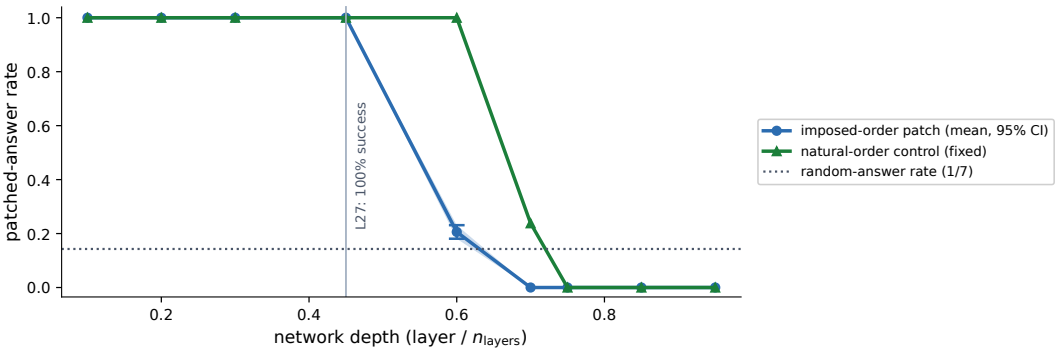


Figure 7: **Entity identity causally redirects the context-specified successor** (Gemma-31B, days; 20 imposed orders $\times$ 42 ordered pairs, plus a fixed natural-order control). Entity-slot patching follows the context-defined successor across depth. Shading and error bars show 95% CIs across imposed orders; the dotted line marks chance $(1/7)$, and the vertical line marks layer 27, the deepest tested layer with 100% imposed-order patch success.

We test the map by *intervention*. At one layer, we overwrite the queried entity's residual with a donor entity's activation from the same context, $h_{E_{\text{orig}}} \leftarrow h_{E_{\text{patch}}}$. Over 42 ordered pairs, we score whether the answer becomes the donor's successor under the imposed or pretrained order (chance $1/N \approx 0.14$). This yields a clean crossover (Gemma-31B; Figure 7): the imposed context

produces the imposed-order successor (1.00), while the natural context produces the pretrained-order successor (1.00).

We next ask which prompt supplies the successor relation after patching. We patch between prompts that assign the donor entity different successors. Following the donor's order would indicate that the patch transfers the relation; following the recipient's order would indicate that the recipient context supplies it. At layers 6–18, every effective patch follows the recipient relation; at layer 27 the rate is 0.88, and none follows the donor relation. At later layers, even patching between prompts with the same order no longer reliably transfers identity, so we cannot distinguish identity from relation transfer (Appendix E, Figure 12). Thus the intervention localizes a context-dependent identity-to-relation computation rather than a fixed successor lookup; it does not intervene on the RSA-defined manifold itself.

Under direct prompting, only the larger models (Gemma-12B/31B and Qwen-27B) meet the criteria in Table 4 for reliable patching under both imposed and natural orders at the same layer. Separate chain-of-thought tests show that, in Gemma-12B/31B and Qwen-27B, entity patching still redirects the final answer after step-by-step reasoning (Appendix E). Whether the relation is constructed anew or configures a stored mechanism remains open.

6.2 Forming Is Cheap; Faithful Use Strengthens with Scale

Larger models more consistently combine clear context-specified geometry with reliable use of the specified relation. Smaller models show a less consistent combination, with substantial differences across model families and concepts (Tables 4 and 7).

Behavioral validation. Under direct prompting, capable models reach 1.00 accuracy at k=1, while less capable models achieve lower accuracy. For $k \geq 2$, direct-prompt accuracy falls while step-by-step reasoning succeeds, consistent with hop-by-hop traversal of a pre-generation structure (Appendix C, Appendix D).

7 Limitations

Our causal evidence is a *localization*: entity-substitution patching identifies the context-defined computation and its depth, but does not establish the RSA-defined manifold as its causal carrier; manifold steering (Wurgaft et al., 2026) is out of scope. We measure finite entity sets at one late layer and the pre-generation token, and RSA captures rank rather than metric shape; metric-shape claims therefore rely on Appendix I.

Coverage is set by our model range: the effect scales across the seven-model Gemma and Qwen ladders, with the causal signature characterized on both (Table 4). This is an existence proof over these models, not a universal law. Manifold steering, richer relations (partial orders and DAGs), and the geometry during native reasoning remain for future work.

8 Conclusion

Across two model families, pretrained organization persists at the entity token but gives way downstream to context-specified relational geometry. The same entities form a cycle or hierarchy on command, and entity identity is causally bound to the specified relation, with clean binding emerging only at scale. Whether this downstream geometry is built or reconfigured remains open; either way, recovered "world-model" geometry must be interpreted relative to context and readout position. More broadly, representation geometry need not only describe stored knowledge; it can map an ongoing computation, revealing where context replaces the prior and directing causal tests of the mechanism that produces the answer.

REPRODUCIBILITY STATEMENT

Every number in this paper comes from a scripted pipeline over open-weight models, and the release accompanying this submission contains the code and figure-ready cached artifacts needed to regenerate every figure, together with per-scramble or aggregate outputs supporting each table. Prompt schemas, exact query paraphrases, and regime instructions are given in Appendix A; the activation-patching procedure, including the multi-token entity handling and the layer choice, is specified in Appendix E; and the confound controls (anisotropy, crossover, serial position, and the wrap-versus-co-occurrence test) are given in Appendices F and H. Headline comparisons are reported with per-scramble 95% confidence intervals rather than pooled means, and the permutation null used to test them is described in Appendix H. All models are publicly available checkpoints identified in Section 4.

AI USE STATEMENT

The authors originated the central research hypothesis and made all final methodological and scientific judgments. Under author supervision, generative AI tools assisted with (i) literature search and novelty stress-testing; (ii) refinement of experimental protocols and controls, implementation and execution of analysis code, and figure generation; and (iii) preliminary analysis and interpretation of outputs, manuscript drafting, and adversarial review. The authors approved experimental designs before execution, reviewed the code, regenerated and checked quantitative results, read and verified every cited work, and adjudicated all interpretations. All reported measurements were computed by the released pipeline over model activations; no reported value was inferred or fabricated by a generative model. The authors take responsibility for the final work, including its text, claims, measurements, and released artifacts.

REFERENCES

A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

J. Engels, E. J. Michaud, I. Liao, W. Gurnee, and M. Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025.

K. Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 55–65. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

W. Gurnee and M. Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

E. A. Hosseini, Y. Li, Y. Bahri, D. Campbell, and A. K. Lampinen. Context structure reshapes the representational geometry of language models. *arXiv preprint arXiv:2601.22364*, 2026.

C. Kim. Task schema and binding: a double dissociation study of in-context learning. *arXiv preprint arXiv:2512.17325*, 2025.

N. Kriegeskorte, M. Mur, and P. Bandettini. Representational similarity analysis: connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

M. A. Lepori, T. Linzen, A. Yuan, and K. Filippova. Language models struggle to use representations learned in-context. *arXiv preprint arXiv:2602.04212*, 2026.

K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg. Emergent world representations: exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

T. Lieberum, M. Rahtz, J. Kramár, N. Nanda, G. Irving, R. Shah, and V. Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in Chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

B. Lin and N. Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, 2024. arXiv:2309.11028.

N. Nanda, A. Lee, and M. Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

C. F. Park, A. Lee, E. S. Lubana, Y. Yang, M. Okawa, K. Nishi, M. Wattenberg, and H. Tanaka. In-context learning of representations. In *International Conference on Learning Representations*, 2025a.

K. Park, Y. J. Choe, Y. Jiang, and V. Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, 2025b.

R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ITw9edRDlD.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

D. Wurgaft et al. Manifold steering reveals the shared geometry of neural network representation and behavior. *arXiv preprint arXiv:2605.05115*, 2026.

H.-D. Xiong, Ji-An Li, R. C. Wilson, K. Lee, and X.-X. Wei. Large language models reorganize representational geometry during in-context learning. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2605.28854.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

A PROMPT TEMPLATES

Most prompts concatenate a *rule* (the redefined order or hierarchy), a *query*, and, where applicable, a regime-specific instruction (Section 4). Prompt schemas follow; repeated entity lines are abbreviated with ellipses, and placeholders appear in *italics* inside square brackets.

Main conflicting-order rule (cyclic concepts). The main multi-scramble geometry experiment uses a numbered order followed by an explicit cycle:

```
You are operating in a REDEFINED calendar.  The ONLY valid
order is below; the normal order does NOT apply.
Order:
1.  [entity at position 0]
2.  [entity at position 1]  ...
Full cycle (wraps):  [order repeated], joined by ''→''
Convention:  start = position 0; 'k steps after X' = item at
position k of the walk.
```

The reported multi-scramble runs use this rule zero-shot, without worked input–output examples. Controlled variants remove the explicit cycle chain or state the order as shuffled adjacency edges. The adjacency form is:

```
[entity 0] is immediately followed by [entity 1]
[entity 1] is immediately followed by [entity 2]  ...
(wrap) [last entity] is immediately followed by [entity 0]
```

For activation patching, the imposed arm uses the numbered order without the explicit cycle chain; the natural control omits the redefinition and asks the same canonical $k{=}1$ query. The closure factorial independently crosses numbered-list vs. adjacency form with the wrap relation absent or present; in the numbered form the wrap is added as "The order wraps around: [*last*] is immediately followed by [*first*]." These four cells form the $2{\times}2$ (Appendix F). The endpoint co-occurrence controls append one of the following sentences to the numbered rule and its convention:

- *Wrap by position*: "The order is circular: one step after the last position (position 6) returns to position 0."
- *Co-mention without a relation*: "For reference, two of the items above are [*last*] and [*first*]."

Here position 6 is the final position of the seven-item order.

Query paraphrases (cyclic). The relational-geometry measure marginalizes the query over the six paraphrases in Table 1 *and* over the step count $k \in \{1, \ldots, 6\}$ (36 queries per entity). Direct behavioral accuracy is also evaluated across all six hop counts (Appendix C); the entity-substitution patch uses one canonical $k{=}1$ query (Appendix E). These are the *structural* step-queries; beyond them the probe also reads *non-structural* templates that ask nothing about the order (a bare mention or a short story; Table 6).

Table 1: Query paraphrases marginalized over by the relational-geometry measure (entity *e*, step count *k*); the probe averages over these and over $k \in \{1, \ldots, 6\}$.

	Query (parametric in *k*)
1	What is *k* steps after *e*?
2	*k* steps after *e* is?
3	From *e*, advance *k* steps. Which item?
4	Starting at *e*, move *k* steps forward. Result?
5	*e* plus *k* steps =?
6	Advance *k* from *e*. Which one?

Imposed-hierarchy rule (tree).

```
You are operating under a REDEFINED hierarchy.  The ONLY
valid parent/child relations apply below; the normal
meanings of these words do NOT apply.
Hierarchy:
- P is the parent of A and B.  ...
Convention:  the root has no parent; 'steps up to the root'
= number of parent-links from the item to the root.
```

The co-occurrence control (Section 5.2) instead states one edge per line ("*P* is the parent of *A*.") and shuffles them so siblings are not co-listed. Tree queries are *relational* ("Who is the parent of *e*?"; "How many steps from *e* up to the root?") or *neutral* ("Consider the item *e*."; "Take note of the item *e*."). Neutral queries ask nothing structural and receive no trailing answer-format instruction.

Arbitrary tokens. The arbitrary-token conditions assign the imposed relation to diverse everyday English nouns with no shared cyclic or hierarchical order (Apple, River, Chair, Cloud, Tiger, Bridge, Lamp, . . . , extended to twelve for the larger cycle and to thirty-one for the depth-4 tree). Noun-to-position assignments are re-randomized across scrambles, so a consistent match to the imposed topology cannot be explained by a fixed pretrained ordering of the nouns.

Prompt regimes. The three regimes (Section 4) use these trailing instructions:

- *direct*: "Answer with ONLY the name, nothing else."
- *scripted reasoning*: "Work through it ONE STEP AT A TIME along the order above, then end with a line exactly: FINAL: *name*".
- *free-form*: no trailing instruction; native thinking enabled.

B LAYER-SWEEP ROBUSTNESS

The shared probe fraction, $0.75\, n_{\text{layers}}$, was selected from diagnostic full-layer sweeps on Gemma-31B and Qwen-27B and then fixed for every model and concept; it was not tuned per condition. The sweeps use one weekday order, three query paraphrases over all six hops, and a prompt with worked examples. Figure 8 reports the resulting all-layer curves. The imposed-order RSA is near zero or noisy in the early and middle network and then rises to a broad plateau across the upper layers, while the natural-order RSA falls to zero over the same range. The probe layer sits inside this plateau and below each model's own peak (Gemma-31B 0.86 at 0.75 vs. a 0.88 peak; Qwen-27B 0.82 vs. 0.91), so the choice is conservative rather than peak-seeking. The dominance of imposed-order over natural-order RSA persists across this plateau.

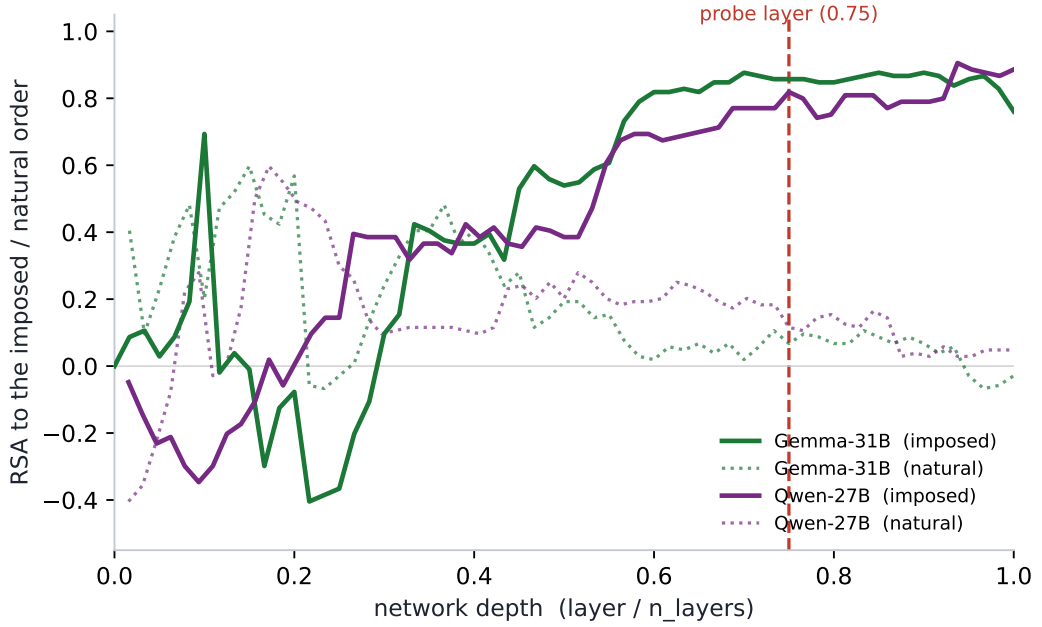


Figure 8: Full-layer sweep (days): imposed-order RSA (solid) and natural-order RSA (dotted) vs. network depth for Gemma-31B and Qwen-27B; the probe layer (0.75) marked dashed.

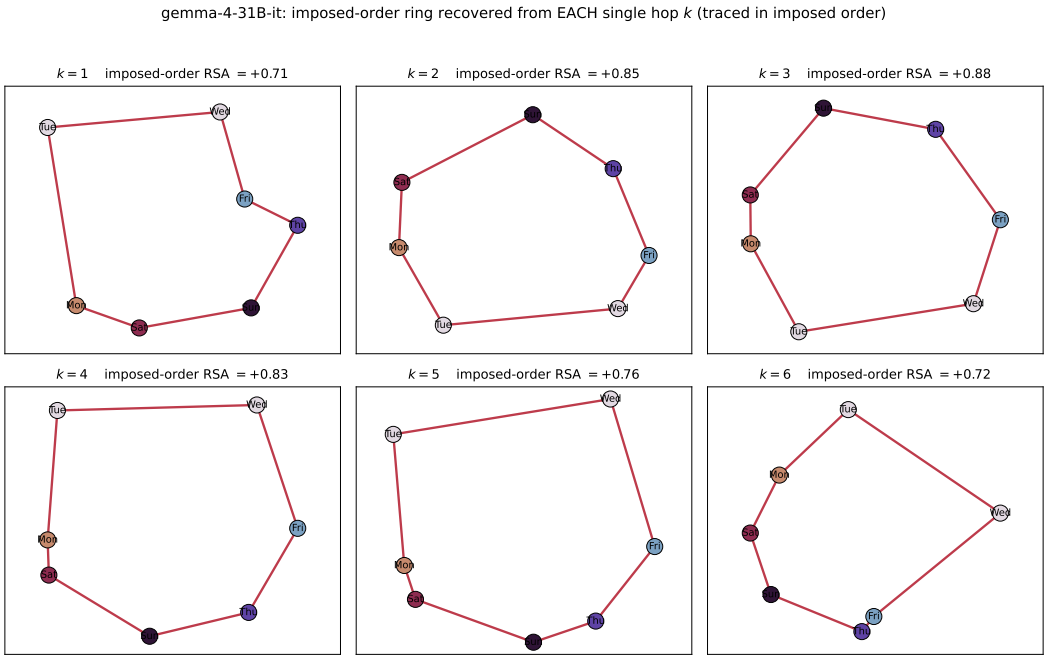


Figure 9: Imposed ring recovered from each hop value k separately (Gemma-31B; centroids from one k at a time, traced in the imposed order; imposed-order RSA $+0.71$ to $+0.88$ across k).

C BEHAVIORAL VALIDATION ACROSS THE LADDER

Figure 10 plots, per model and at matched hop counts k=1 to 6, accuracy on the natural order and on the imposed order (direct regime; days, list format, bare-name answers, two scrambles). Testing the natural order at every k provides a baseline for multi-step accuracy. Gemma-12B/31B and Qwen-27B remain near ceiling on the natural order, while their imposed-order accuracy falls sharply at $k \geq 2$. Their difficulty therefore concerns following the new order, rather than multi-step traversal in general. Llama-8B and Qwen-4B also lose accuracy on the natural order as k increases, so their failures are not specific to the imposed order. Geometry, rather than accuracy, is the primary measure.



Figure 10: Accuracy by hop count k=1–6 per model in the *direct* (no-reasoning) regime (days, list, bare-name answers, two scrambles): natural order (dashed, per-k ceiling) vs. imposed order (solid), with the $1/7$ chance line. Direct-prompt accuracy degrades with k even for the capable models; the reasoning regimes recover ceiling (Figure 11).

D BEHAVIORAL REGIMES AND THE GENERATION BUDGET

Figure 11 and Table 2 report imposed-order accuracy under the three regimes, with the fraction of generations reaching the 2048-token cap. In Gemma-31B and Qwen-27B, *direct*-prompt accuracy degrades with hop count while *scripted* reasoning recovers ceiling (Gemma-31B 1.00, Qwen-27B 0.99); *free-form* recovers ceiling in Gemma-31B (1.00).

The one asymmetry is Qwen's free-form regime (0.89), where 81% of generations reach the 2048-token cap. Scripted reasoning avoids truncation and reaches 0.99, showing that guided traversal remains successful. We do not infer how a larger generation budget would affect free-form accuracy.

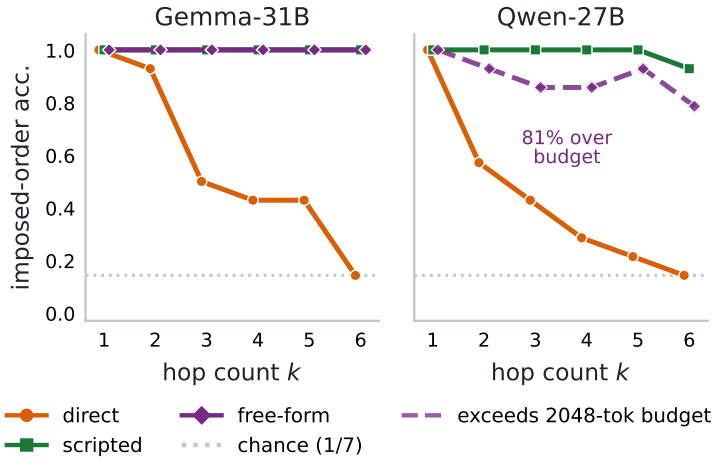


Figure 11: Imposed-order accuracy by hop count under three prompt regimes (days; $1/7$ chance dotted): direct, scripted reasoning, and free-form. Qwen free-form runs past the 2048-token budget on 81% of generations; its scripted reasoning reaches 0.99.

Table 2: Imposed-order accuracy by prompt regime (days, list, bare answers). Direct generation allows 16 new tokens; scripted and free-form generation allow 2048. "free-form over budget" is the fraction of free-form generations that reach the 2048-token cap. Gemma terminates within budget and solves both reasoning regimes; Qwen's free-form reasoning over-runs the budget but its scripted reasoning does not.

	imposed accuracy			
model	direct	scripted	free-form	free-form over budget
Gemma-31B	0.57	1.00	1.00	0%
Qwen-27B	0.44	0.99	0.89	81%

E ACTIVATION-PATCHING DETAILS

Causal window across depth. Entity-substitution patching is swept over nine layers in Gemma-31B and Qwen-27B. In imposed contexts, patch success remains 1.00 through layer 27 in Gemma-31B and layer 29 in Qwen-27B, then falls sharply. We interpret cross-order outcomes only where patching a different entity under the same order still redirects the answer to that entity's successor (the positive control). The imposed arm uses 20 independent orders per model; the fixed natural-order control is repeated six times for Gemma and twice for Qwen. The *wrong-slot control* injects the donor's final entity-token activation at the token immediately before the query entity. It checks whether the answer becomes the donor's successor even when the activation is inserted at the wrong position. We use this control in the cross-model same-layer depth summary below, not to locate where relational binding occurs.

Source activation and injection. For each entity we cache, in a single forward pass over the corresponding context prompt, the residual-stream activation at that entity's token position(s) at every swept layer; this patch activation therefore already carries the current context. To patch $E_{\text{orig}} \leftarrow E_{\text{patch}}$ we overwrite, via a forward hook at one layer, the residual at the original entity's position(s) with the patch entity's, then generate and check whether the answer becomes the patch entity's successor *under the current context's order*.

Cross-order identity-versus-relation control. We patch between two prompts that specify different orders and score the donor entity's successor under both the donor and recipient orders. At layers 6, 12, and 18, the same-order positive control transfers identity on every order pair, and the

cross-order patch follows the recipient successor on every pair. At layer 27 it follows the recipient successor at 0.88 (95% CI 0.85–0.92). It never follows the donor successor at any informative layer. From layer 36 onward, the same-order positive control largely or completely fails, so these layers cannot distinguish identity from relation transfer (Figure 12).

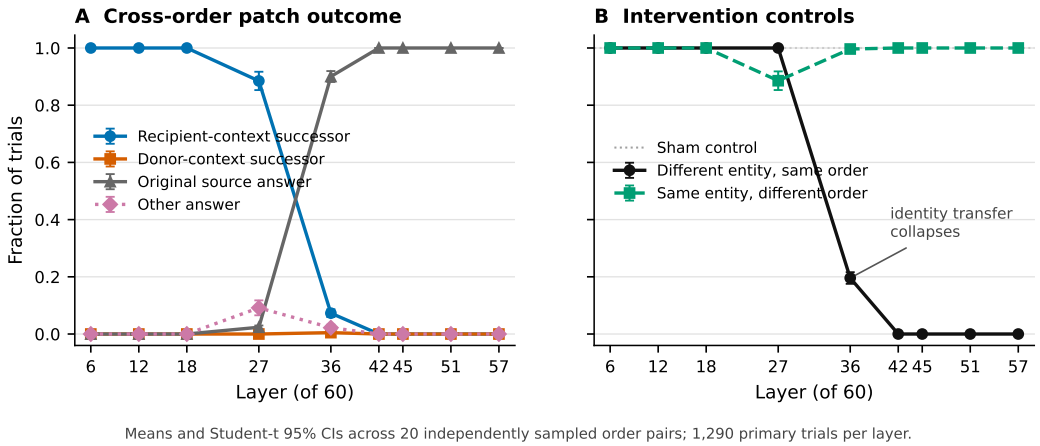


Figure 12: Cross-order entity-slot control (Gemma-31B, days; 20 independent order pairs). *(A)* Outcomes after patching a donor entity between prompts specifying different orders. *(B)* Intervention controls: a different-entity, same-order patch tests identity transfer; a same-entity, different-order patch tests whether the recipient answer is preserved. The sham replaces an entity's activation with its own activation from the same prompt and preserves the correct answer at 1.00 throughout. Means and 95% CIs use the order pair as the statistical unit.

Alternative-successor scoring in the same-context sweep. Because a donor entity can have the same successor under the imposed and pretrained orders, we restrict that comparison to trials where its successors differ. At the clean same-layer locus in each model, 666 of 840 trials across 20 imposed orders are diagnostic: all 666 follow the imposed successor and none follows the pretrained alternative. The remaining 174 share a successor and are correct under both orders.

Multi-token entities. We overwrite matched entity sub-tokens, right-aligned to the end of each entity span; when the original and donor differ in length, this uses the final $\min(m_{\text{orig}}, m_{\text{donor}})$ positions. This alignment rule concerns the intervention site. Relational geometry is measured separately at the final pre-generation residual.

Normalization and the norm confound. We inject the *raw* patch residual, with no rescaling. The concern that a flip could be driven by activation *norm* rather than content is controlled directly by the random-vector control: a Gaussian vector rescaled to the per-row norm of the real patch, injected at the same slot, flips the answer only at chance (0.14), whereas the real patch flips it at 1.00. Norm alone therefore does not produce the effect.

Chain-of-thought scale check (Gemma). In a two-order sweep, neither E2B nor E4B shows the full double dissociation (peak imposed/natural success $0.46/0.00$ and $0.00/0.00$), whereas 12B and 31B reach $1.00/1.00$. With two orders per checkpoint, this establishes a small-to-large contrast, not a precise onset scale.

Chain-of-thought patch (Qwen-27B). Across three imposed orders (42 source–donor pairs each), patching at layers 6–19 redirects every visible final answer to the context-defined successor. All 102 trials whose imposed and pretrained successors differ follow the imposed successor; none follows the pretrained alternative. At layers 45–48, many patched generations do not reach a visible final answer; at still later layers, final answers reappear but donor identity no longer transfers. This arm therefore does not localize a precise transition depth. A deterministic parser reads only the

post-reasoning response; independent judging of the first two orders gives the same 1.00 live-layer success.

Step-count sweep. A companion patch targets the step-count token rather than the entity: for the query "advance k steps from E", we overwrite the k-digit residual $k_{\text{orig}} \leftarrow k_{\text{patch}}$ and check whether the answer becomes the k_{patch}-hop successor, restricted to pairs correct at both counts. Across the ladder (days, k up to 6, 2 scrambles), the step count is causally used cleanly in the larger models (patch success 1.00 for Gemma-12B/31B and 0.96 for Qwen-27B) and partially in the smaller ones (Qwen-9B 0.84, Qwen-4B 0.69, Gemma-E2B 0.62, and E4B 0.46), where the original step count can still influence the answer. Together with the entity patch, this shows that the computation uses both the queried entity and the requested step count.

Mechanism class. We patch the full residual stream, not attention or MLP outputs separately, so we localize the identity-to-relation computation by *depth* (Appendix E) rather than by component; attention-vs-MLP attribution is left to future work.

Swept layers. The sweep is at depth fractions $\{0.1, 0.2, 0.3, 0.45, 0.6, 0.7, 0.75, 0.85, 0.95\} \times n_{\text{layers}}$, with 0.75 also used for the geometry readout. Table 3 gives the absolute indices per model. These fractions are fixed a priori (the same grid for every model) and we report the patch and control success across the whole sweep (Figure 7) rather than at a single layer chosen on the outcome, so the reported dissociation is not selected post hoc.

Table 3: Absolute layer indices swept by the activation-patching arm, per model (the depth fractions of Appendix E resolved to integer layers).

model	n_{layers}	swept layer indices
Gemma-E2B	35	4, 7, 10, 16, 21, 24, 26, 30, 33
Gemma-E4B	42	4, 8, 13, 19, 25, 29, 32, 36, 40
Gemma-12B	48	5, 10, 14, 22, 29, 34, 36, 41, 46
Gemma-31B	60	6, 12, 18, 27, 36, 42, 45, 51, 57
Qwen-27B	64	6, 13, 19, 29, 38, 45, 48, 54, 61

The crossover across scale. Table 4 reports the two arms of the double dissociation across the ladder, using the common two-scramble subset per model–concept cell and all ordered unequal entity pairs: 42 for weekdays and 132 for months. We report three summaries of the crossover (Section 6.1). *Imp* and *Nat* are patch-success rates under imposed and natural orders, respectively, each at its strongest layer. Success means answering with the donor entity's successor under the corresponding order. *Same-layer depth* is the deepest normalized depth where both rates reach 0.9 and wrong-slot control success remains below 0.2. The wrong-slot control inserts the donor activation immediately before the query entity, as defined above. The smaller models do not meet this joint criterion. Qwen-4B/9B reach 1.00 natural-order patch success at their best layers, but lower imposed-order success. Gemma-E2B has no natural-order patch success; Gemma-E4B has partial success (0.48 for days and 0.67 for months). Only Gemma-12B/31B and Qwen-27B meet the same-layer criterion. Some smaller-model rates lie near the cutoffs, so these results do not establish a sharp size threshold.

Table 4: **Causal patch success across the scale ladder** (direct prompt, common two-scramble subset per model–concept cell; 42 ordered pairs per weekday scramble and 132 per month scramble). *Imp*/*Nat*: imposed- and natural-context arms (own-map success), each at its strongest layer. *Same-layer depth*: deepest normalized depth with Imp, Nat ≥ 0.9 and wrong-slot control success < 0.2; "—" means no such layer, and "n/r" marks a model×concept not run (Gemma-12B was run on days only). Chance is $1/7 = 0.14$ (days), $1/12 = 0.08$ (months). Llama-8B builds no usable map and is omitted.

	days			months		
model	Imp	Nat	same-layer depth	Imp	Nat	same-layer depth
Gemma-E2B	0.90	0.00	—	0.78	0.00	—
Gemma-E4B	0.89	0.48	—	0.95	0.67	—
Gemma-12B	1.00	1.00	0.46	n/r	n/r	n/r
Gemma-31B	1.00	1.00	0.45	1.00	0.98	0.45
Qwen-4B	0.73	1.00	—	0.74	1.00	—
Qwen-9B	0.89	1.00	—	0.88	1.00	—
Qwen-27B	1.00	1.00	0.45	1.00	0.98	0.45

F WRAP RELATION, CLOSURE, AND ENDPOINT CO-OCCURRENCE

Measuring cyclic closure. We gauge how ring-like a geometry is with two graded indices rather than a binary line/cycle verdict: *dRSA* = RSA(cyclic) − RSA(linear), signed (higher = more circular, negative = less); and a wrap-pair *closure ratio*, the last↔first distance relative to typical adjacent-pair distance (≈1 when tightly closed, ≫1 when open). dRSA is global over all entity pairs, while the closure ratio isolates the relation that differs between the open and wrapped specifications. Figure 13 also shows the corresponding centroid layouts.

The wrap relation strengthens closure. Holding tokens and task fixed, stating the *wrap* makes the geometry more ring-like by degree: in Gemma-31B it lowers the closure ratio from 2.09 to 1.06 for the numbered form and from 2.07 to 1.25 for adjacency edges, while dRSA changes sign in both forms (Figure 13). The same effect appears on arbitrary tokens with no pretrained ring to inherit (Table 9), while its magnitude across Gemma-31B and Qwen-27B is capability-dependent.

Relation versus co-occurrence. Because the wrap edge names the two endpoints together and the closure metric reads exactly that pair, the effect could in principle be local token co-occurrence. Co-mentioning the endpoints without relating them leaves the geometry open in both models (Table 5). Stating that the last position returns to position 0, without naming the endpoints together (exact prompts in Appendix A), partially closes Gemma but not Qwen; directly stating endpoint adjacency closes Gemma and moves Qwen toward closure. Thus co-occurrence alone does not explain the effect, but relation-only closure is not uniform across families.

Table 5: Wrap relation vs. endpoint co-occurrence (days, 3 scrambles). Closure ≈ 1 is tightly closed and ≫ 1 open. Co-mentioning the endpoints without a relation leaves both models open. Wrap-by-position partially closes Gemma but not Qwen; direct endpoint adjacency closes Gemma and shifts Qwen toward closure.

	Gemma-31B		Qwen-27B	
condition	closure	dRSA	closure	dRSA
list (no wrap)	2.09	−0.06	1.96	−0.29
co-mention (no relation)	2.19	−0.10	1.84	−0.25
wrap by position	1.47	+0.06	1.85	−0.30
adjacent wrap	1.09	+0.16	1.36	−0.08

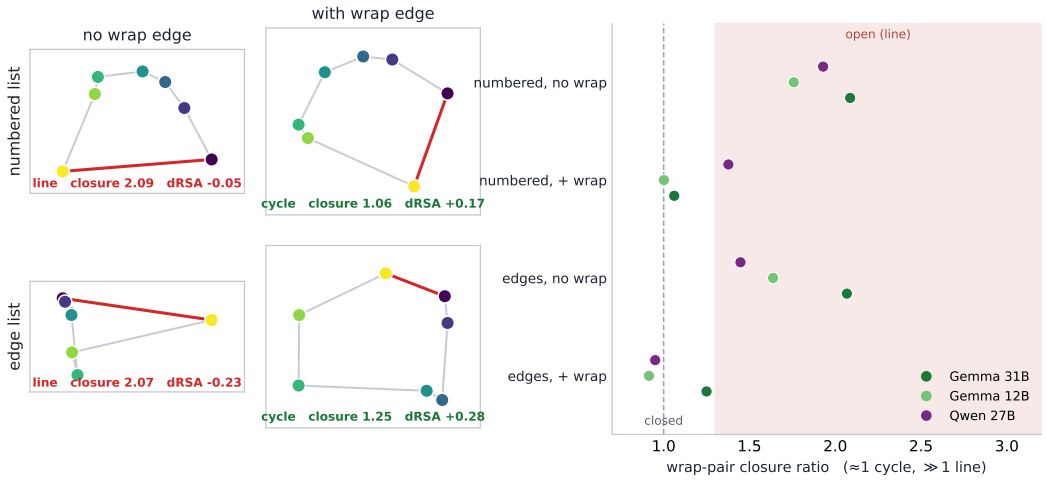


Figure 13: Form × wrap closure test. *Left:* Gemma-31B centroid layouts from scramble 0 (PC1–PC2, colored by imposed position, wrap edge red); labels report mean closure and dRSA over two scrambles. *Right:* mean wrap-pair closure ratio over two scrambles for three models. The wrap effect is strongest in Gemma and weaker in Qwen.

G ADDITIONAL EVIDENCE FOR CONTEXT-SPECIFIED GEOMETRY

Table 6: Imposed-order RSA on days (Gemma-31B, Qwen-27B). *Top:* under non-structural probes (bare mention, story), imposed−natural RSA. *Bottom:* imposed-order RSA by hop count, with the k=1-vs-full-k RDM agreement.

	Gemma-31B	Qwen-27B
non-structural probe: imposed−natural RSA		
neutral mention	+0.52	+0.41
story	+0.46	+0.41
imposed-order RSA by hop count		
k=1 only	+0.67	+0.64
k=6 only	+0.71	+0.54
full k	+0.81	+0.56
k=1 vs. full-k RDM (ρ)	0.74	0.93

Table 7: Per-model imposed-order vs. natural-order RSA (direct prompt, means over 10 scrambles): the evidence behind the dominance (Section 5.1) and within-family scale claims. Figure 3 plots the days columns.

	days		months	
model	imposed	natural	imposed	natural
Gemma-31B	+0.87	−0.03	+0.81	+0.05
Gemma-12B	+0.81	−0.03	+0.54	+0.14
Gemma-E4B	+0.72	+0.06	+0.58	+0.31
Gemma-E2B	+0.60	−0.06	+0.59	+0.23
Qwen-27B	+0.67	−0.01	+0.71	+0.12
Qwen-9B	+0.10	+0.44	+0.60	+0.19
Qwen-4B	+0.31	+0.24	+0.46	+0.38

The per-model imposed-order vs. natural-order RSA is given in Table 7. Additional support comes from per-concept, per-scramble plots (centroids connected in the imposed order), and by the two

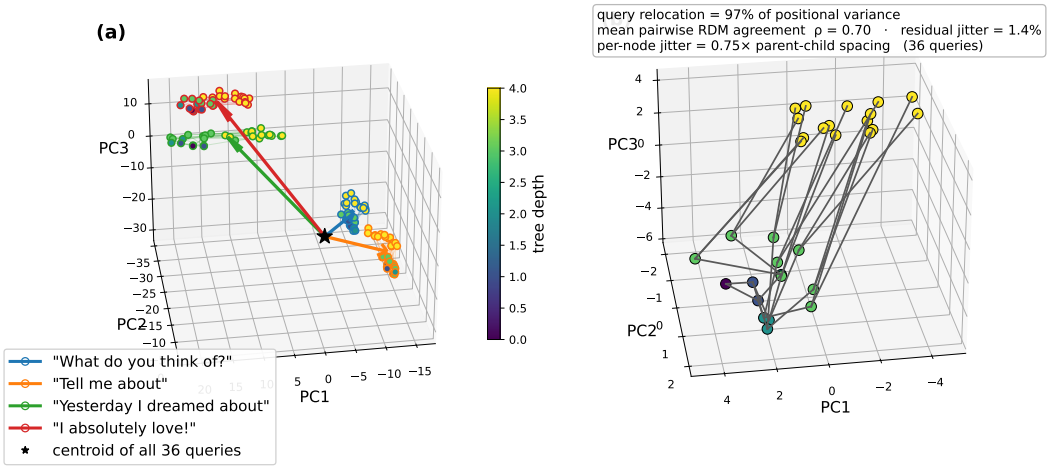


Figure 14: Imposed depth-4 tree across 36 query types (Gemma-31B). *(a)* Four query types place the tree at different locations (arrows = query vectors from the 36-query centroid; color = depth). *(b)* After recentering, the 36 geometries retain a shared relational organization (mean pairwise RDM agreement $\rho = 0.70$); query identity accounts for 97% of positional variance.

confound controls (the crossover and the anisotropy-matched null) whose full procedure and numbers are given in Appendix H. Per-scramble confidence intervals for all headline numbers accompany the release.

RSA is not a projection artifact. RSA is computed in the full residual space; the 2D and 3D plots are illustration. Table 8 shows that for the displayed geometries 2D, 3D, and full-dimensional RSA agree, so the low-dimensional views preserve the broad comparison. Adding the third PC does not reveal a clear ring in Qwen-9B or Qwen-4B: imposed-order RSA remains low in 2D, 3D, and the full space (0.23–0.32). Gemma-E2B shows stronger alignment (0.55–0.64). Gemma-31B and Qwen-27B show the clearest cycle-like layouts and more concentrated variance: their top three PCs explain 0.88 and 0.83, compared with 0.63–0.71 for the smaller models in Table 8.

Table 8: Imposed-order RSA at increasing projection dimension (days, scramble s_0), with the variance captured by the top three PCs. 2D/3D/full agree; the clean cycle and concentrated variance coincide at the largest models.

model	RSA 2D	RSA 3D	RSA full	top-3 var.
Gemma-31B	+0.81	+0.85	+0.82	0.88
Gemma-E2B	+0.55	+0.64	+0.56	0.71
Qwen-27B	+0.78	+0.71	+0.65	0.83
Qwen-9B	+0.28	+0.23	+0.26	0.71
Qwen-4B	+0.32	+0.31	+0.24	0.63

Topology on arbitrary tokens. Table 9 gives the 2×2 on arbitrary tokens with no shared cyclic order, the control behind the claim in Section 5.2 that the wrap statement, rather than a fixed pretrained ordering of the nouns, drives the closure. Adding the wrap closes the loop (dRSA ≥ 0, closure $\approx$1) and removing it leaves the geometry open, less circular (dRSA < 0, closure $\gg$1), in both surface forms.

Table 9: Topology on arbitrary tokens (Gemma-31B): dRSA and wrap-pair closure for the four specifications. In this control, open-versus-closed geometry follows the wrap relation rather than surface form or a fixed pretrained ordering of the nouns.

specification	dRSA	closure
numbered list, no wrap	-0.12	2.7
numbered list, +wrap	$+0.16$	1.06
edges, no wrap	-0.27	3.3
edges, +wrap	$+0.31$	1.1

H CONFOUND CONTROLS IN DETAIL

This appendix gives the full procedure and numbers for the relabeling and anisotropy confounds summarized in Section 5.1 (the third, serial position, is handled by the tree and wrap evidence of Section 5.2 and Appendix F); numbers are Gemma-31B on days unless noted, from the same conflict-state residuals used throughout. Both controls address the residual stream's strong *anisotropy*: the entity centroids share a dominant direction, so raw cosines are high and barely distinguish the orderings (imposed-adjacent pairs average 0.986 vs. 0.979 natural-adjacent, a 0.007 gap). The reported structure lives not there but in the *mean-centered* RDM, which removes the shared direction before measuring proximity; every RSA in the paper uses it. The controls verify the resulting signal is neither a renaming of a fixed geometry (Control 1) nor an artifact of the cone (Control 2).

Control 1: relabeling (does the geometry actually move?). *Worry:* a high imposed-order RSA might only reflect renaming which point we call "position 3," with a fixed set of centroids scoring equally against any declared order. *Test:* probe two identical-token contexts (natural, scrambled) and score each context's centroids against both templates; a context-dependent change in geometry predicts a crossover, whereas a fixed geometry with relabeled points predicts none. *Result:* the crossover is sharp: scrambled context imposed-RSA $+0.84$/natural -0.04, and natural context imposed -0.10/natural $+0.87$. Thus, the same tokens give opposite geometries. Ranked by RSA to the scrambled-context geometry against all $7! = 5040$ orderings, the imposed cycle is maximal; its 14 rotation- and reflection-equivalent orderings share the maximum (top 0.28%).

Control 2: anisotropy (is the RSA an artifact of the cone?). *Worry:* in a narrow cone even an unstructured arrangement can give a positive RSA, so we need a null with the data's cone geometry but no imposed structure. *Test (label-shuffle null):* shuffle which entity label sits on which real centroid vector, mean-center, and recompute the imposed-order RSA, 2,000 times per scramble across ten orders; this preserves the cone, norms, and pairwise angles while destroying the geometry–order correspondence, and we report the empirical tail probability rather than a Gaussian z. *Result:* the observed RSA clears this null on every concept in Gemma-31B and Qwen-27B: all ten scrambles are significant for days, notes, months, zodiac, and clock. RSA against the natural order remains near zero (-0.05 to $+0.12$).

Caveat (the smallest concept). At $N = 7$, cyclic symmetry alone contributes $14/5040 = 0.0028$ of the exact permutation mass. Nonzero exceedance counts are therefore expected even for a perfectly matched cycle; every weekday scramble remains significant at $p \leq 0.008$. At $N = 12$, the corresponding symmetry mass is $24/12! \approx 5 \times 10^{-8}$.

I RSA CARRIES STRUCTURE: CONVERGENT VALIDITY

We compare RSA with PCA appearance and successor accuracy across eight models, then report a separate persistent-homology test of cyclic versus hierarchical specifications. Prior circular-manifold work used sparse autoencoders and PCA inspection (Engels et al., 2025); RSA provides a quantitative comparison with a specified relational template and a permutation test.

- **PCA crossings.** We connect entity centroids in imposed order in PC1–PC2 and count self-crossings. Across eight models, imposed RSA correlates with fewer crossings (Spearman $\rho =$

0.84, using the negative crossing count). This measures the appearance of the projection, not planarity of the full geometry.

- **Behavior.** Across the same eight models, imposed RSA correlates with one-step successor accuracy at Spearman $\rho = 0.83$. Both correlations use one point per model, with each measure averaged over three random orders.
- **Persistent homology.** We compare cyclic and hierarchical specifications using the same arbitrary tokens, token assignments, query templates, layer, and final pre-generation readout at $N{=}7$ and $N{=}15$. Normalized top-H_1 persistence is greater under the cyclic specification in all four model-by-size comparisons ($\Delta = +0.083$ to $+0.140$); all crossed-bootstrap 95% confidence intervals exclude zero, and exact paired tests give $p \leq 0.027$ (Table 10). The query-shuffle null independently permutes entity identities within each query template before averaging activations into centroids, breaking entity correspondence across templates. Only Qwen-27B at $N{=}15$ independently clears this null. Each persistence bar records the range of distance thresholds over which a loop exists; the longest bar is not robustly separated from the second-longest. We therefore treat persistent homology as a relative, confirmatory diagnostic rather than standalone certification of a single dominant loop (Lin & Kriegeskorte, 2024; Carlsson, 2009).

Table 10: **Matched persistent-homology audit.** ΔH_1 is cyclic minus hierarchical top-H_1 persistence, normalized by median off-diagonal distance. Confidence intervals resample scrambles and query templates together. p_{pair} is the exact two-sided paired sign-flip test; p_{abs} compares the cyclic condition with the query-shuffle null.

Model	N	ΔH_1	95% CI	p_{pair}	p_{abs}
Gemma-31B	7	$+0.083$	$[+0.028, +0.157]$	.027	.884
Gemma-31B	15	$+0.140$	$[+0.063, +0.226]$	.002	.331
Qwen-27B	7	$+0.128$	$[+0.037, +0.210]$	.018	.110
Qwen-27B	15	$+0.136$	$[+0.009, +0.189]$	.002	.009

The cross-model correlations connect RSA to projected appearance and task performance, while the matched persistent-homology comparison supports the cycle–hierarchy distinction. The correlations span model families and sizes and are not independent of the scale trend; they do not establish within-model prediction or a causal effect of geometry on accuracy. Residual representations are anisotropic (Ethayarajh, 2019); mean-centering removes their shared direction before the cone-preserving permutation test in Appendix H.

J TOPOLOGY RANGE: NOT JUST CYCLES

The formation-and-override result is not specific to 1-D cyclic concepts; it extends to 2-D grids, reported with the same full-dimensional RSA, permutation null, and imposed-vs-natural dominance used for the cyclic concepts.

An imposed 2-D layout overrides a native 2-D prior. On a $3{\times}3$ block of QWERTY keys (Q W E / A S D / Z X C) the model's native keyboard geometry is a weak 2-D grid, keyboard-distance RSA $+0.53$ at Qwen-27B, exceeding the alternative 1-D templates (reading-order $+0.20$, alphabetical $+0.31$; $p{<}.01$), and weak at Gemma-31B. (These are single-letter keys, so the signal is spatial, not a shared sub-token artifact.) Imposing a redefined layout (shuffled "right of"/"below" relations) overrides it: imposed-layout vs native-QWERTY RSA $+0.78$ vs $+0.09$ at Qwen-27B (dominance $+0.69$, 6/6 scrambles significant against the null) and $+0.42$ vs $+0.02$ at Gemma-31B (dominance $+0.40$, 5/6), the 2-D analog of the imposed-order dominance on cyclic concepts.

Formation on arbitrary tokens. Imposing a $3{\times}3$ grid on diverse everyday nouns reorganizes their geometry toward the specified layout: grid-distance RSA exceeds the 1-D line template (Gemma-31B $+0.43$, beating the line on 6/6 scrambles). Because noun-to-grid assignments are re-randomized across scrambles, the effect cannot be explained by a fixed pretrained grid among the nouns.

Scope. The model's native 2-D grids (keyboard, chess) are weaker and more model-dependent than its native cyclic rings, so the cyclic concepts carry the main results; the 2-D evidence shows the formation-and-override mechanism is not tied to a single topology.

K SAME WEEKDAYS UNDER CYCLE, LINE, AND HIERARCHY

We hold the seven weekday tokens, layer $(0.75\,n)$, and final pre-generation readout fixed while imposing a cycle, an open line, or a binary tree (Section 5.2 and Fig. 6). Each imposed condition averages 12 neutral/diverse queries over 10 randomized token-to-position assignments. The matched no-rule baseline instead averages six query phrasings at each of six natural-order offsets, yielding 36 final-token states per weekday.

The no-rule geometry has significant natural-cycle structure in both models: RSA $+0.57$ for Gemma-31B (95% crossed-bootstrap CI $[0.17, 0.78]$, exact $p{=}.014$) and $+0.60$ for Qwen-27B ($[0.19, 0.72]$, $p{=}.031$). Exact tests enumerate all 7! label permutations; the bootstrap resamples query phrasings and offsets. Under an imposed cycle, both models match the specified cyclic distances. Under a tree, both match depth distances and retain essentially no cyclic RSA ($-0.04/-0.01$).

For the line, we additionally report *effective dimension*, the participation ratio $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the entity-centroid PCA eigenvalues. It uses no entity ordering: an ideal line has value 1 and a planar ring has value 2, while noise or additional directions can raise either value. The imposed line contracts from 4.0 to 1.5 dimensions in Gemma but only from 3.8 to 2.7 in Qwen, so clean line formation is capability-dependent.

Table 11: The same weekdays under different relational specifications. Imposed rows are means over 10 randomized assignments; NAT is the matched no-rule baseline at the final pre-generation token. Target RSA uses cyclic distance for NAT/cycle, linear distance for line, and depth distance for tree. Effective dimension is the PCA-eigenvalue participation ratio.

	Gemma-31B		Qwen-27B	
specification	target RSA	eff-dim	target RSA	eff-dim
NAT (no rule)	$+0.57$	4.3	$+0.60$	3.2
cycle (ring)	$+0.72$	4.0	$+0.71$	3.8
line	$+0.60$	**1.5**	$+0.68$	2.7
tree (depth)	$+0.67$	1.8	$+0.68$	2.1