BRINGING ORDER TO LLMS: THE LIE BRACKET IS ABSENT BUT TRAINABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

When language models compute transformations, do they know the rules? We ask this for $\mathfrak{so}(3)$ by investigating two separate aspects: representation of *group elements*, and representation of the *Lie bracket* that composes them. We find that pretrained models have neither of the above. Across eight models in five families, no pretrained or instruction-tuned checkpoint shows a measurable bracket, and none represents the composed group element beyond token-identity and last-token cues. LoRA fine-tuning using non-abelian composition installs element decoding well beyond the last-token bound in all eight models, and a partially faithful bracket in seven of eight. The installed structure sits at depths that differ by family (40–83%). Below a threshold on the number of generators, an order-blind predictor suffices. Above it, all eight models beat a baseline that sees only the bag of generator tokens, six of them clearly, but none beats, beyond seed noise, a baseline that knows the signed generator pair and not its order. During fine-tuning, element information emerges before bracket structure, and bracket fidelity does not predict behavioral margin across models. A second finding is methodological. The three statistics that would translate representation into behavior each return a forced positive: cross-model decoder agreement by shared supervision, alignment-based transfer by per-model fidelity under $SO(3)$-equivariance, and the causal order-swap correlation by a Baker–Campbell–Hausdorff identity that reproduces the effect ($r = 0.96$) with no model in the loop. Together, these findings separate three questions: representing a group's elements, representing its algebra, and using it.

1 INTRODUCTION

When a language model is trained to compose transformations, such as rotations, group elements, or sequential operations, does it learn to represent the *elements* being composed, or the *algebra* that composes them? The two are easy to distinguish in principle, but standard probing methodology (Belinkov, 2022; Hewitt & Liang, 2019) does not separate them. A model can carry a perfect linear representation of every group element in a sequence while carrying no representation of the operation that governs how those elements combine. This is the representational analog of predicting orbits without recovering the force laws (Vafa et al., 2025), and has been discussed in the recent literature on emergent structured representations (Li et al., 2022; Park et al., 2024; Nanda et al., 2023a; Chughtai et al., 2023; Engels et al., 2025).

We study how LLMs represent $\mathfrak{so}(3)$, a canonical three-dimensional non-Abelian Lie algebra, whose bracket is the cross product. We construct a synthetic composition task and ask four questions in sequence (Figure 1): (a) is the bracket present in pretrained models (*absent*); (b) can fine-tuning install it (*trainable*); (c) where does the installed structure live (*localized*); and (d) does behavior then use it (*not detectably*)? We discuss each of these in turn.

Across eight models in five families (sizes ranging 1.5B–13B), no pretrained or instruction-tuned checkpoint carries a measurable bracket structure (its significant cells stay within what chance alone would produce). Nor do these models represent the group *element*: linear decoding of the composed rotation looks strong, but the composition is largely fixed by the multiset of generators, and once that is controlled for, what remains is at the level the last token alone supplies. Pretraining supplies token identity and last-token position, but neither the element nor its algebra. Some families do carry

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

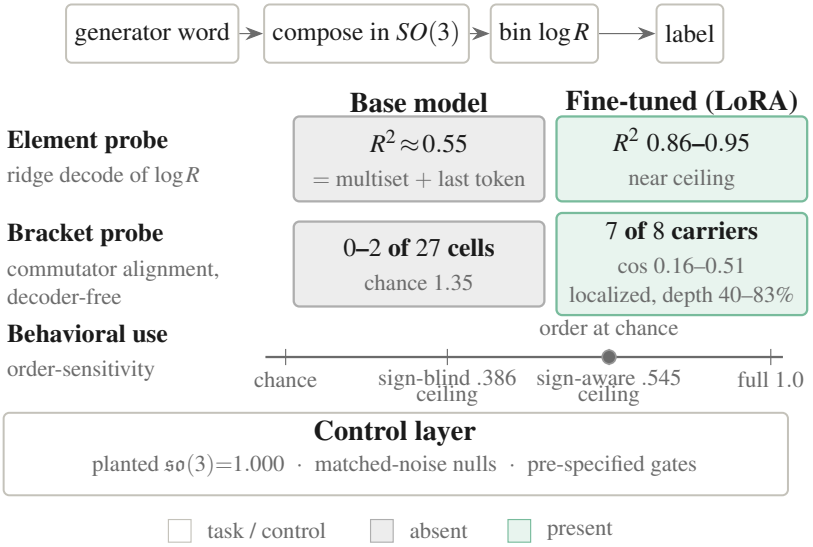

Figure 1: **What we measure, and what fine-tuning changes.** A generator word is composed in $SO(3)$ and its $\log R$ binned to a label (top). Two residual-stream probes—*element* (ridge decode) and decoder-free *bracket* ($\text{bracket}_{\cos}$)—are colored by verdict (gray absent, teal present): both absent in base models, both present after fine-tuning. Yet order-use stays absent—commutator accuracy clears the sign-blind ceiling (0.386) but sits at the sign-aware one (0.545, dot). All readouts run over a control layer (planted $\mathfrak{so}(3)$=1.000, matched-noise nulls, pre-specified gates).

causally used *abelian* group structure (Engels et al., 2025; Kantamneni & Tegmark, 2025), but not the non-commutative aspects.

LoRA fine-tuning (Hu et al., 2021) using non-abelian composition changes the outcome. Fine-tuned models decode the composed element far beyond the last-token bound, three of them beyond first-and-last-token cues, and carry a partially faithful bracket in seven of eight models. The eighth, Llama-3-8B, decodes the composed element as well as two Qwen carriers do, yet carries no interpretable bracket. Task structure predicts a *threshold* in the number of generators: below it the brackets lie mostly on existing generator axes, at only one or two magnitudes, and no model beats an order-blind baseline.

The installed structure concentrates in layer depth. In the Qwen family, the rising profile survives depth-matched null distributions that control for residual-stream norm and effective-rank growth (Heimersheim & Turner, 2023). Where it concentrates is a property of the model family, not of transformers: Qwen models peak late, OLMo-2 models at mid-depth, with peak location varying far less within a family than across families.

Whether any model uses the installed bracket is a separate concern. We find no detectable bracket-specific behavioral advantage. At $k = 12$, all eight models beat the ceiling of a predictor that is blind to both the generator sign and the order, six of them clearly. However, a predictor that knows which signed pair was composed, without its order, reaches 0.545. OLMo-2-7B is the strongest carrier, but it does not clear even the sign-blind ceiling at $k = 10$ across five seeds. Across the eight models, bracket strength and behavioral margin are uncorrelated (Spearman $\rho = 0.38$, $p = 0.35$, $n = 8$).

The harder finding is methodological. Three natural statistics for the questions above each return a confident positive that does not hold under matched null distributions (Figure 3D), and each positive is *forced*. Cross-model decoder agreement follows from supervision toward a common coordinate frame (Ding et al., 2021; Bansal et al., 2021). An alignment-based transfer score follows from each model's own fidelity combined with the $SO(3)$-equivariance of the cross product. The causal order-swap correlation is a Baker–Campbell–Hausdorff identity that holds with no model in the loop.

2 Related Work

Sequence models trained on structured domains develop internal representations of that structure: board states (Li et al., 2022; Nanda et al., 2023b), linearly encoded features (Park et al., 2024;

Gurnee & Tegmark, 2024), Fourier circuits for modular arithmetic (Nanda et al., 2023a), and group-theoretic circuits recurring across seeds (Chughtai et al., 2023; Stander et al., 2023). Pretrained LLMs carry multi-dimensional structure that is causally used (Engels et al., 2025; Kantamneni & Tegmark, 2025), and recent work recovers relational bindings (Lee et al., 2026) and the mechanism by which transformers compose group elements (Marchetti et al., 2026). These works establish that elements, bindings, and composition mechanisms are represented. We test the operation table itself, second-order structure with analytic ground truth, and find it absent in pretrained LLMs even where causally-used abelian structure is present. Concurrent work makes a complementary point in the *abelian* setting: Llama-3.1-8B represents cyclic concepts (months, weekdays, hours) with circular geometry, yet computes over them with a generic base-10 addition circuit rather than modular arithmetic in the concept's own period (Feucht et al., 2026): representational structure need not be the structure a model computes with. Our non-commutative case study shows that the relevant structure is not merely bypassed but *absent* in pretrained models, and must be installed.

The closest prior work fine-tunes language models on group composition and probes the running product (Li et al., 2025a; Kvinge et al., 2026), asking whether the composed *element* is tracked. We ask a different question: whether the *algebra* that composes those elements is represented, and find that the two come apart: fine-tuned Llama-3-8B represents the composed element within the range of the other models yet carries no interpretable bracket (Section 4.2). Whether fine-tuning creates new mechanisms or merely re-weights existing ones is an open question (Prakash et al., 2024; Chhabra et al., 2025; Jain et al., 2024; Wang et al., 2025). We supply a case with ground truth on the "creates" side: the bracket is absent from all eight base models and appears in seven of them after a small LoRA fine-tune.

A growing body of work asks which interpretability statistics are *entailed* rather than measured (Huh et al., 2024; Kornblith et al., 2019): similarity indices score high on random networks unless de-confounded (Cui et al., 2022), verdicts invert under a change of measure (Soni et al., 2024; Ding et al., 2021), probes and circuit discovery return plausible explanations on randomly initialized networks (Méloux et al., 2025), and depth profiles are far from flat under level-matched placebos (Braun, 2026). We add three provable instances: a cross-model agreement score that is set by each model's own fidelity, under $SO(3)$-equivariance, up to a computable factor; an intervention whose signature correlation is a matrix identity; and a frame-removal ablation that no rotation can instantiate. Appendix N covers the full literature, including symmetry discovery and Lie-structure methods (Moskalev et al., 2022; Forestano et al., 2023; Gruver et al., 2022; Dehdarirad et al., 2026).

3 METHODS

3.1 TASK, ALGEBRA, AND MODELS

We study the Lie algebra $\mathfrak{so}(3)$ of 3D rotations, a canonical three-dimensional non-Abelian Lie algebra where any order-sensitivity is attributable to the bracket. Furthermore, its automorphism group is $SO(3)$ acting by rotation, which is what makes several naïve cross-model statistics provably uninformative (Section 5). In the standard basis $\{e_1, e_2, e_3\}$, its bracket is

$$[e_i, e_j] \;=\; \varepsilon_{ijk}\, e_k, \qquad \text{equivalently} \qquad [a, b] \;=\; a \times b \quad \text{for } a, b \in \mathbb{R}^3, \tag{1}$$

the cross product: composing two rotations in different orders differs by an amount governed entirely by the cross product of their axes.

The model reads a word $w = g_{i_1} g_{i_2} \cdots$ over k named generators, each a rotation by a fixed angle θ about its own fixed axis ($\pm$ pairs included), and must name the rotation they compose. Its answer is a discretized label: the composed rotation vector $\log(R) \in \mathbb{R}^3$, with each Cartesian component (x, y, z) binned on a fixed grid, so the label is the triple of per-axis bin indices (Figure 2). Unless stated otherwise, all experiments use $k = 12$, $\theta = 0.6$, seed 0. An *abelian control* task replaces the generators with commuting ones, so every bracket is zero. It keeps the non-abelian task's k, label format and training recipe, and disqualifies metrics that score "represents its own task" rather than "represents the bracket" (Section 3.2).

To ask whether an LLM layer represents a *group element*, we fit a linear (ridge) decoder from the residual stream to the analytic target $\log(R) \in \mathfrak{so}(3)$ and evaluate on held-out prefixes. At $\theta = 0.6$ and lengths 2–4 the composed rotation is $\approx$92% determined by the multiset of generators, so a high

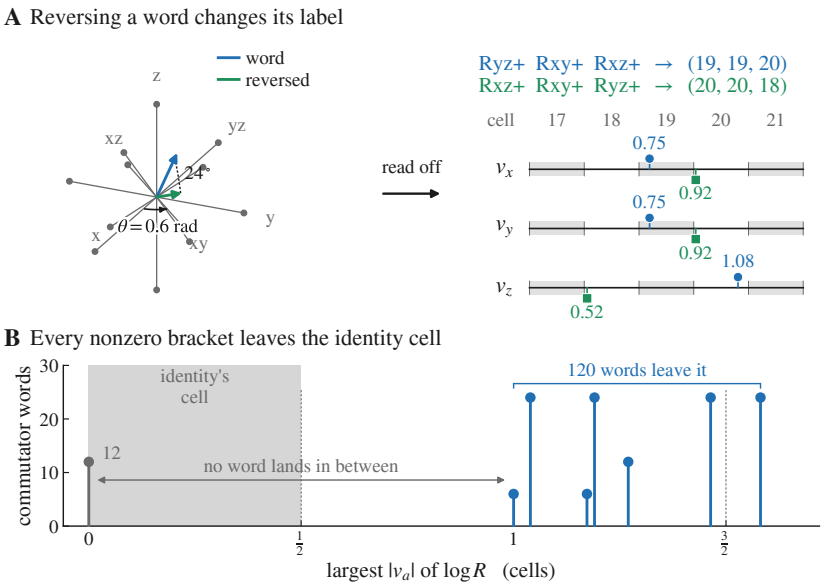

Figure 2: **The composition task and its labels. (A)** Each generator rotates by $\theta = 0.6$ rad about one of six axes (x, y, z and the diagonals xy, xz, yz; $\pm$ pairs give $k = 12$). The word `Ryz+ Rxy+ Rxz+` (blue) and its reverse (green) compose to rotation vectors 24° apart. Each component of $\log R$ is binned into 31 cells of 0.203 rad spanning $[-\pi, \pi]$ (cells 17–21 shown). The two orders land in different cells on every axis, one cell apart on v_x and v_y and two on v_z, and so receive different labels, $(19, 19, 20)$ vs. $(20, 20, 18)$. **(B)** The 132 commutator words by the largest absolute component of $\log R$, in cell units; dotted lines mark cell edges. The 12 words built from a $\pm$ pair have zero bracket and stay in the identity's cell. The other 120 all leave it, none between $\frac{1}{2}$ and 1, so every nonzero bracket changes the label.

R^2 score shows only that token identities are linearly readable. Our element test therefore decodes the order-dependent residual $Y - \mathbb{E}[Y \mid \text{multiset}]$ and reads it against (i) what the multiset plus the last token supplies (0.545), (ii) what the multiset plus the first and last tokens supplies (0.882), (iii) a stream encoding the multiset together with the full composed element (0.989; Appendix C.8). To ask whether a layer represents the *algebra*, we estimate a generator matrix M_i for each named generator from the model's own state updates, fit one linear map Φ from the analytic rotation vectors v_i to the M_i, and test whether the realized commutators $[M_i, M_j]$ match $\Phi(v_i \times v_j)$ on pairs Φ never saw as brackets (Eq. 2).

Take three generators X, Y, Z (rotations by θ about x, y, z) and the word XY. *Element:* after each prefix the model holds a specific rotation R: $\log R = (\theta, 0, 0)$ after X, and $\log\big(\exp(\theta e_y)\exp(\theta e_x)\big)$ after XY. The element test asks only whether a decoder recovers $\log(R)$ from the hidden state on held-out prefixes, i.e., whether the model knows *which* rotation it currently holds. *Bracket:* $\mathfrak{so}(3)$ dictates $[e_x, e_y] = e_x \times e_y = e_z$, so X-then-Y must differ from Y-then-X by a rotation about z; the bracket test asks whether the model's own generators obey this, $[M_X, M_Y] \propto \Phi(e_z)$. A model can pass the first and fail the second; it can represent the elements without the rule that composes them.

Words of length 2–4 are partitioned into a *train* set and a *held-out* set (disjoint words, same lengths). The *commutator* split is the $k(k-1)$ words $g_i g_j g_i^{-1} g_j^{-1}$ over ordered pairs of distinct signed generators (Figure 2B), and is the split that isolates order-sensitivity. Generator matrices M_i are estimated from the model's own transitions. For each generator i, one ridge-regularised affine regression $Z_{p+i} \approx A_i Z_p + b_i$ (penalty 10^{-2}, intercept included) is pooled over every non-empty prefix p and its extension $p{+}i$ across 2000 words. Z_p is the residual-stream state at the prompt's readout token (=>) after prefix p, projected onto the top-d principal directions of that layer's readout-token states, and $M_i = \mathrm{Re}\log A_i$, the real part of the principal logarithm (when A_i has negative real eigenvalues no real logarithm exists, and the real part is the logarithm with those directions reflected). Generator names are plain text, tokenized by each model's tokenizer into 2–4 BPE pieces with default special-token handling.

Table 1: The abelian control disqualifies three of four candidate metrics (probe instrument, Qwen2.5-1.5B). The abelian arm has every bracket exactly zero, yet at deep layers its decode R^2 and homomorphism ratio match the non-abelian fine-tune. Only commutator alignment is bracket-specific: it is undefined on the abelian arm and separates fine-tuned from base (0.882 vs. 0.517 at layer 25). Base-arm values and update-fit R^2 are in Appendix C.3.

Layer	decode R^2 (FT / abelian)	hom. ratio ↓ (FT / abelian)	commutator align. (FT / base)
8	0.985 / 0.643	0.079 / 0.631	0.859 / 0.605
14	0.982 / 0.702	0.088 / 0.547	0.864 / 0.526
19	0.986 / **0.937**	0.079 / **0.163**	0.855 / 0.505
25	0.983 / **0.947**	0.072 / **0.115**	**0.882** / 0.517

For each $(k, \theta, \text{length})$ we define two order-blind ceilings on the commutator split. The *sign-blind* ceiling is the best accuracy of a predictor that sees only the unordered axis pair, which for a commutator word is exactly its bag of tokens (0.386 at k=12). The *sign-aware* ceiling is the best accuracy of one that also sees the signed pair $\{g_i, g_j\}$ but not its order (0.545). Because every variant of a commutator word contains the same four tokens, knowing the signed pair already requires positional information (which two tokens come first). What the sign-aware predictor lacks is the order *within* the pair, which to leading order is exactly the sign of the bracket. Beating it is therefore our behavioral test of bracket use. Computation in Appendix B.

We experiment with eight models across five families, seven instruction-tuned checkpoints and one pretrained checkpoint (Llama-3-8B): Qwen2.5 (1.5B, 3B, 7B), Llama-3-8B, Mistral-7B-v0.3, OLMo-2 (7B, 13B), and Granite-3.1-8B. Our fine-tuning is LoRA (Hu et al., 2021) (r=32, α=64, 7 target modules, 12 epochs) on the k=12 non-abelian task. Base and fine-tuned arms of every model pass through the identical measurement code path. Full details are provided in Appendix A (Table 4).

3.2 Instruments and Controls

At a given layer and residual subspace (sub-dimension $d \in \{3, 8, 16\}$ of a PCA basis), we fit a linear map $\Phi : \mathfrak{so}(3) \to \mathbb{R}^{d \times d}$ from algebra coordinates to generator matrices M_i estimated from the model's own state updates, and score

$$\text{bracket}_{\cos} \;=\; \cos\big(\,[M_i, M_j],\; \Phi(v_i \times v_j)\,\big) \tag{2}$$

averaged over non-degenerate generator pairs (60 of 66 at k=12; the 6 pairs with $v_i \times v_j = 0$ are excluded). We ask whether the *realized* commutator of the model's generators agrees with the algebra. The criterion is internal consistency with the planted algebra's structure constants. Significance is computed by a 500-fold permutation test per (layer, d) cell; a 27-cell scan has 1.35 significant cells expected by chance.

A model is a (bracket) *carrier* if (a) at least 8 of its 27 cells are significant (1.35 expected by chance; ≤ 3 counts as absent, 4–7 as ambiguous) and (b) its best sub-dimension-8 cell is significant. Llama-3-8B fails both: 6 of 27 cells reach significance (ambiguous by count) and its sub-dimension-8 row is empty at all nine layers. Per-cell p-values are 500-fold permutation tests whose floor is 0.002. Evidence rests on cell counts (8–26 of 27 for carriers, 0–2 for every base arm).

Three controls run through the identical code path on every measurement: `exact` (a planted analytic $\mathfrak{so}(3)$ embedding; the readout must return 1.000), `random` (matched random matrices; must return ≈ 0), and `shuffled` (permuted pair correspondence; must return ≈ 0). Across all screening runs, the planted control returned 1.000 in every cell, and both negatives sat at their nulls (Appendix C). A verdict from a run whose control failed is *void*, and not reinterpreted.

The abelian arm (every bracket exactly zero) nonetheless scores as a genuine structure on three of four candidate metrics: decode R^2, homomorphism ratio and update-fit R^2 (Table 1). Only commutator alignment is bracket-specific, so we report the other three as fit diagnostics only.

To evaluate any cross-model statistic, a permutation null is not enough. We build a matched-noise null by planting *two independent* analytic $\mathfrak{so}(3)$ embeddings, injecting independent noise calibrated (by bisection) so each arm reproduces its measured within-model $\text{bracket}_{\cos}$, and running the identical estimation pipeline. Appendix E presents the construction and its own sanity gate: at zero noise, the simulator returns 1.000.

Table 2: Screening across eight models. "Cells" = significant cells at $p < 0.05$ out of 27 (chance 1.35), decoder-free readout, k=12. Best cell is at sub-dimension 8. Base arms are the instruction-tuned checkpoints (pretrained for Llama-3-8B) without task fine-tuning; the true pretrained checkpoints of four families are in Appendix K (Table 18) and are likewise absent; fine-tuned arms are LoRA on non-abelian composition. Peak depth is the best cell's fractional depth. Full per-layer profiles in Appendix K.

Model	Family	Base cells	FT cells	FT best bracket$_{\cos}$	Peak depth
Qwen2.5-1.5B	Qwen	2/27	22/27	+0.487 (p=.002)	82%
Qwen2.5-3B	Qwen	0/27	17/27	+0.223 (p=.002)	83%
Qwen2.5-7B	Qwen	1/27	19/27	+0.264 (p=.002)	82%
Mistral-7B-v0.3	Mistral	0/27	8/27	+0.160 (p=.002)	72%
OLMo-2-7B	OLMo-2	1/27	**26/27**	+**0.514** (p=.002)	50%
OLMo-2-13B	OLMo-2	2/27	19/27	+0.391 (p=.002)	40%
Granite-3.1-8B	Granite	2/27	17/27	+0.394 (p=.002)	50%
Llama-3-8B	Llama	2/27	6/27	+0.041 (p=.122)	—

4 RESULTS

4.1 BASE MODELS CARRY NEITHER THE ELEMENT NOR THE BRACKET

Before any task fine-tuning, we screen every model for the two structures in turn: first the bracket, then the group element, and find that neither is present. No base model carries the bracket (Table 2, "Base cells"): 0–2 significant cells against 1.35 expected, across five families and 1.5B–13B. The base arms are the instruction-tuned checkpoints (pretrained for Llama-3-8B) without task fine-tuning; re-measuring the true pretrained checkpoints of four families on the same grid returns the same verdict (1–2 cells each, Appendix K). The group element is absent too, beyond what the input already supplies. Raw decoding looks strong (R^2 0.86–0.95 in the base arms; Table 9) but is fixed by the multiset. The order-dependent residual reaches R^2 0.36–0.57 in every base arm, no more than 0.03 above the last-token bound (0.545).

4.2 FINE-TUNING INSTALLS BOTH THE ELEMENT AND THE BRACKET

When we fine-tune the base models using LoRA on non-abelian composition, both structures appear. The order-dependent residual probe reaches R^2 0.86–0.95 in all eight models, far above the last-token bound (0.545) at which every base arm sits. Against the stricter first-and-last-token bound (0.882), five of eight exceed it, three clearly (OLMo-2-7B 0.947, OLMo-2-13B 0.933, Granite 0.931), and the other three sit within 0.03 below it (Table 9). We claim element representation beyond first-and-last-token cues only for the three models that clearly exceed it. Seven of eight models also carry the bracket (Table 2, "FT"), on the identical readout and code path that returned the base-arm nulls. The eighth, Llama-3-8B, decodes the element at $R^2 = 0.864$, level with Qwen2.5-1.5B (0.870) and Qwen2.5-3B (0.859), which both carry the bracket, yet its sub-dimension-8 row is empty: element decoding and algebra representation come apart.

Carrier scores (0.160–0.514; Table 2) are read against a ceiling of 1.0: a mechanism that implements the group action $e^{\theta X_i}$ exactly scores 1.000 through the full readout path (Appendix C.7). Because the readout fits one linear operator per generator, a mechanism that composes a whole word at once can score below 1.0 even when exact. The strongest carrier, OLMo-2-7B, comes from a family on which the task was never developed.

An installed bracket matters behaviorally only when the task is order-sensitive, and that depends on the task size. For $k = 6$, every pairwise bracket lands on an existing generator axis, and through $k = 8$ the number of bracket directions and magnitudes remains small enough (Table 6, Appendix) that we predicted, before running the sweep, no margin over the sign-blind ceiling until k=10. This changes for $k \geq 10$ (Appendix D). We find that no behavioral margin over either ceiling appears before k=8 $\rightarrow$ 10: at k=8, Qwen2.5-1.5B is at -0.196 and OLMo-2-7B at -0.167, both having solved the general (held-out) task above their own ceilings (+0.237, +0.248).

The representation of the algebra, moreover, appears before the task demands it. In both families, the signature is present at k=8, one step *before* the threshold: Qwen's bracket-specific probe gap opens at k=6 $\rightarrow$ 8 (+0.305 at k=8), and OLMo-2-7B at k=8 carries the strongest sub-dimension-8 bracket

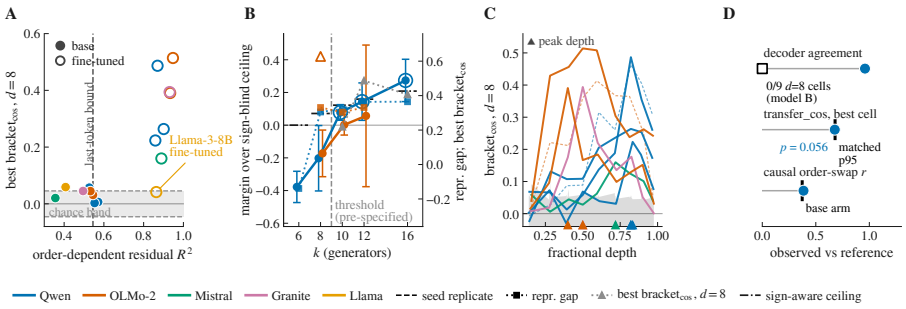

Figure 3: **(A) Absent, and trainable.** Best-cell bracket fidelity vs. order-dependent residual decode R^2 for the eight base arms: every residual value sits no more than 0.03 above the last-token bound (0.545, dashed) and every bracket value sits in the chance band (the 95th percentile of the maximum over nine d=8 null cells, averaged over the eight base arms: 0.075). Fine-tuned arms (hollow) move to residual R^2 0.86–0.95, and seven of eight leave the band. **(B) No detectable order use.** Behavioral commutator margin over each task size's sign-blind ceiling (95% t-intervals over seeds; rings: 95% interval above zero), two families: no gain before k=10 in either. Qwen2.5-1.5B rises above it, and at k=16 also clears the sign-aware ceiling (dash-dotted), though not significantly at three seeds; OLMo-2-7B stays at the sign-blind ceiling through five seeds. Right axis: the probe's commutator-alignment gap (squares) and best bracket$_{cos}$ at d=8 (triangles). **(C) Localized.** Depth profiles of bracket$_{cos}$ for all seven carriers (colors = families; dashed = seed replicates) over the depth-matched null band (grey): peak depth is seed-stable and family-dependent; profile shape is not. **(D) Entailed vs. measured.** Three statistics that appear to bridge representation and use, each against a reference that factors in its entailment: cross-model decoder agreement (0.959) against the decoder-free verdict on the same pair (where one model carries no bracket); the alignment-transfer best cell against the matched-noise null ($p = 0.056$; tick at the cell's own null $p_{95} = 0.68$); and the causal order-swap correlation on the carrier ($r = 0.39$) against the same protocol on a base model that cannot do the task ($r = 0.37$).

we measure anywhere (bracket$_{cos} = 0.628$, 25/27 cells—above its own 0.514 at k=12). Verdict direction and magnitude replicate across seed, $k \in \{10, 16\}$, and $\theta = 0.45$ (Appendix J, Table 16).

4.3 The bracket is layer-localized

The structure that fine-tuning installs is concentrated in depth. We build a per-layer surrogate null that matches each layer's generator-norm spectrum, ensemble anisotropy and common affine component (Appendix G, Table 13); it is centred at zero at every depth by construction (every surrogate family is sign-symmetric), and the rising-profile statistic clears it at the floor of that null ($p = 0.002$) for all three Qwen carriers. That null carries its own planted positive control, which recovered the planted layer at the floor of its null in all 126 of 126 layer-model-corruption combinations.

Where the structure sits is associated with the family (Table 2, "Peak depth" and Table 19): Qwen at 82–83%, OLMo-2 at mid-depth (40–59% across two sizes and two seeds; the table lists the primary seed), Mistral between, and Granite at OLMo-2's depth. That is a spread of 43 points across families against at most 19 within one.

4.4 An installed bracket is not detectably used

Within one model. OLMo-2-7B carries a strong bracket at every task size we train (best-cell bracket$_{cos} = 0.63$ at k=8, 0.56–0.59 at k=10 (seeds 0 and 3; seed 4 void, Table 10), 0.41–0.51 at k=12; 23–27 of 27 cells significant) and solves the general (held-out) task above its order-blind ceiling at every size (+0.25 at k=8, +0.11 at k=10; +0.03 at k=12, n.s.). Yet we find no detectable bracket-specific behavior: its commutator accuracy sits at the sign-blind ceiling at k=10 across five seeds (+0.007, 95% CI $[-0.055, +0.068]$) and is indistinguishable from it at k=12. Qwen2.5-1.5B, whose bracket at k=10 is less than half as strong (+0.219), is the one model that moves: its commutator accuracy rises with k (three-seed means +0.07, +0.14, +0.27 at k=10/12/16, against each size's sign-blind ceiling). Against the sign-aware ceiling, the one that isolates order use (Appendix B), Qwen's k=12 gain is +0.03 in the screening run and -0.020 ± 0.110 over the three-seed sweep

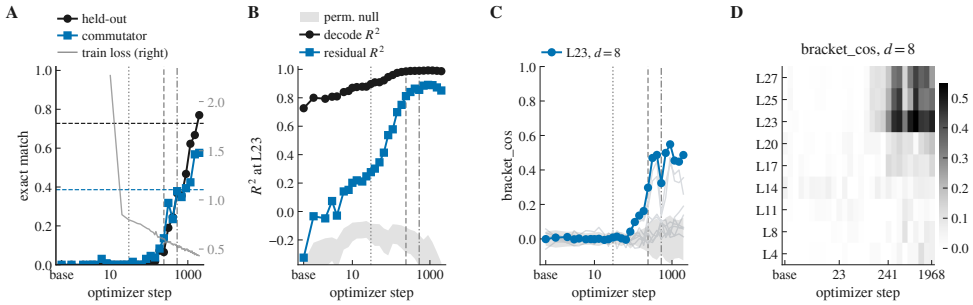

Figure 4: **Training dynamics** (Qwen2.5-1.5B, seed 0; x = optimizer step, log; base at left edge). **(A)** Held-out and commutator exact match with order-blind ceilings (dashed) and training loss (grey, right axis). **(B)** Decode and order-dependent residual R^2 at layer 23 (highest final residual R^2; probe refit per checkpoint), residual permutation null as a band. **(C)** $\text{bracket}_{\cos}$ at the peak cell (L23, d=8) with its null band; other layers faint. Vertical rules: t_{50} of element (dotted), algebra (dashed), behavior (dash-dot). **(D)** $\text{bracket}_{\cos}$ at d=8 by layer and step.

(Appendix D). At k=10 the strongest carrier does not beat even the sign-blind ceiling, and the one model that does, Qwen2.5-1.5B, does so by knowing which signed pair was composed, not in which order.

Across models. At k=12 (screening run, seed 0), bracket strength and behavioral margin show no detectable relation (Spearman $\rho = +0.38$, $p = 0.35$, n=8). All eight clear the sign-blind ceiling (0.386), six by 0.08–0.19, Llama-3-8B among them (+0.076) with no interpretable bracket; Granite clears it by five answers and Mistral-7B by two (53/132) despite a significant bracket. Against the sign-aware ceiling (0.545) only Qwen2.5-1.5B (+0.03) and Qwen2.5-7B (+0.008) are positive and the other six sit at or below it (Table 21). Within the Qwen family the two co-vary (1.5B > 7B > 3B). Behaviorally, then, these models exploit which signed generators occur, not the order the bracket encodes. The rank correlation is unaffected by this choice. We see that at k=12 the sign-aware ceiling is the same for every model, so margins against either ceiling have the same ranks and the same ρ.

4.5 The element appears smoothly before the algebra

To see *when* each structure appears, we re-run the k=12 fine-tune with the screening recipe and save the adapter at 27 checkpoints (steps 1–4, then a geometric grid of ratio 1.3 to 1,968 steps; the base model is step 0) and take the same three readouts at each: order-dependent residual R^2 (element, refit at every checkpoint on the nine-layer grid and read at the layer where the final checkpoint's residual R^2 is highest), best-cell $\text{bracket}_{\cos}$ (algebra), and commutator exact match (behavior). Each curve is normalized between its base and plateau; t_{50} is the half-rise step, and a logistic-fit width in $\log_{10}$ steps calls a rise abrupt (< 0.25) or gradual (> 0.5). We run three seeds each for Qwen2.5-1.5B and OLMo-2-7B (Figure 4; Appendix M, Table 23).

Two findings hold in all six runs. *No rise is abrupt* (0/18; every $10 \to 90\%$ rise spans $\geq 2.8\times$ in steps, with 3–17 checkpoints inside it). And *the element precedes the algebra* in every run (median t_{50} 30 vs. 241 for Qwen, 110 vs. 241 for OLMo). The algebra in turn precedes commutator accuracy in all three OLMo-2-7B seeds and two of three Qwen seeds (the third ties; two onsets are a tie when their t_{50} lie within one grid ratio, $1.3\times$, of each other). Because commutator accuracy never exceeds the sign-aware ceiling beyond seed noise (Section 4.4), this third curve tracks the task being learned, not the bracket being used. On the abelian control, commutator accuracy reaches its ceiling of 1.0, while the bracket readout is undefined at every checkpoint, as the construction requires.

5 Methodological Findings

Three statistics that appear to carry a single-model result across models or into behavior each collapse under a null that factors in what per-model facts already entail (Table 3).

Decoder agreement. Qwen2.5-1.5B and Llama-3-8B, fine-tuned to the same external $\mathfrak{so}(3)$ coordinates, agree when decoded at 0.959—but both decoders were trained to that common frame, so the agreement follows from the shared targets, not from shared representations. The decoder-free readout

Table 3: Three standard statistics, their observed values, and the nulls. Each clears a naïve null and fails the null that factors in its entailment. Full constructions in Appendices E–H.

Statistic	Observed	Against matched null	Entailment source
Cross-model decoder agreement	0.959	decoder-mediated	shared supervision
Alignment transfer, best cell	$+0.677$	$p = 0.056$; $0/17$ cells clear	fidelity $\times$ equivariance
Causal order-swap correlation	$r = +0.385$	base arm $r = +0.373$	BCH identity (r=0.96, no model)

breaks the tie: Qwen carries a strong bracket, Llama no interpretable one (Hewitt & Liang, 2019; Ding et al., 2021).

Transfer. Because $\mathrm{Aut}(\mathfrak{so}(3)) = SO(3)$ and the cross product is equivariant,

$$Q(a \times b) = (Qa) \times (Qb), \tag{3}$$

any two faithful copies of $\mathfrak{so}(3)$ are alignable by construction: under independent per-model errors α, β the expected transfer cosine is $\cos\alpha \cos\beta > 0$, so each model being *separately* accurate forces the statistic positive with no shared structure (Appendix E.2). The least-squares projection inflates it further ($1.2\times$ at d=3 to $4.7\times$ at d=16; Appendix E.2). The best cell ($+0.677$) looks decisive against a permutation null but sits at $p = 0.056$ against the matched-noise null, $0/17$ cells clearing.

Causal order-swap. Injecting a generator pair in both orders and correlating the behavioral difference with the bracket norm looks causal, but it is a Baker–Campbell–Hausdorff identity: to second order

$$\left\| \log(e^{\alpha M_i} e^{\alpha M_j}) - \log(e^{\alpha M_j} e^{\alpha M_i}) \right\| = \alpha^2 \, \|[M_i, M_j]\| + O(\alpha^3), \tag{4}$$

so the displacement tracks the bracket norm whether or not the model uses it. The identity gives $r = 0.961$ with no model in the loop; the behavioral version gives $r = +0.385$ on the carrier and $+0.373$ on a base model with *zero* task ability (Appendix H, Table 14); matching perturbation magnitude does not repair the test (Table 15).

Shared deviations. Whether two models deviate from ideal $\mathfrak{so}(3)$ in correlated ways stays confounded: the strongest cross-model pair (Qwen2.5-1.5B $\times$ OLMo-2-7B) reaches 29% and 44% of the two same-model ceilings on the mean (Appendix F) but fails the sign test ($p = 0.19$; Table 12), and does so under a shared frame. The ablation that would separate shared supervision from shared structure is not instantiable: 0 of the 5,000 in-band rotations (from 14,646 draws) preserve the sign-blind ceiling within tolerance (Appendix I), because that ceiling is held up by label collisions along the frame's own axes.

6 Discussion

Representing a group's elements, representing the algebra that composes them, and using that algebra are three separate properties. In $\mathfrak{so}(3)$, where ground truth is analytic, pretraining supplies neither structure, a small fine-tune installs both at family-dependent depths, and representation and use come apart within a single model, so a positive answer to one question does not imply the others. The synthetic setting makes these separations provable; whether *naturally* acquired structure is algebraic remains out of reach for a decoder-free readout. Methodologically, each statistic that appears to carry a single-model result across models or into behavior is forced positive, by shared supervision, by per-model fidelity under a symmetry, or by an algebraic identity.

We regard the negative results here as important to the field of tuning LLMs and using them to understand complex patterns and behaviors. We convert a vague question ("do models represent the structure behind a pattern?") into sharper, testable ones. The results also warn practitioners who fine-tune models to recover structure that a probe reading out cleanly, or a model beating a baseline, need not mean the structure is used. Our two-ceiling decomposition and matched-noise nulls are reusable instruments for this: the sign-aware ceiling isolates genuine structure-use from cheaper cues in any composition task, and the entailment audit transfers to any interventional claim, well beyond $\mathfrak{so}(3)$. Because the Lie algebra is demonstrably representable and our negative holds only at the sizes and task complexities tested, there is reason to expect that "use" is achievable under stronger pressure (larger k, order-forcing training, or scale).

AI USE STATEMENT

In this work, we used generative AI tools for coding assistance (drafting and refactoring simulator, extraction, and analysis scripts, and running experiments through coding agents), for discussing experimental design and interpreting results, for writing and editing assistance, and for review-style critical feedback on drafts. All model outputs were reviewed by the authors, who retain full responsibility for the scientific content, experimental design, empirical results, theoretical claims, and any errors.

ETHICS STATEMENT

This work studies the internal structure of open-weight language models on a synthetic, fully specified mathematical task ($\mathfrak{so}(3)$ composition); it involves no human subjects, no personal or sensitive data, and no deployment of models in consequential settings. All models are publicly released checkpoints used under their respective licences, and all data are generated analytically, so no privacy or consent concerns arise. The methodological contribution should enable safer practice: our central caution is that several standard interpretability statistics return confident positives that are entailed by construction rather than measured, and our matched-noise nulls and pre-specified verdicts are meant to make interpretability claims about model internals harder to overstate, a benefit for the reliability of evaluations on which higher-stakes decisions may eventually rest.

REPRODUCIBILITY STATEMENT

Every number in this paper traces to a persisted result file, except the few pre-fix or run-time values that the text marks as not persisted. The results figures load every plotted value from those files and fail rather than substitute; the task figure recomputes its grid and fails if its ceilings differ from the printed ones. Verdict tables, instruments, and model provenance are documented in the released repository (URL withheld for review; anonymized archive supplied as supplementary material). The matched-noise null simulators and planted controls are CPU-only: the planted controls re-run in minutes, and the full CPU tier, including the cross-model nulls, in a few hours (`docs/REPRODUCTION.md`). Verdict sentences for every branch of the pre-specified round (its branches and a seed top-up), including outcomes that did not occur, are reproduced verbatim in `docs/PRESPEC.md` (the transition-level plant in `results/prespec_round1/task4b_prespec.md`), and were written before the corresponding numbers existed; we say "pre-specified" rather than "pre-registered" because the record is the authors' own version history, for which no third-party timestamp exists.

REFERENCES

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

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

Johannes Bertram, Luciano Dyballa, T Anderson Keller, Savik Kinger, and Steven W Zucker. Decoding alignment without encoding alignment: A critique of similarity analysis in neuroscience. *arXiv preprint arXiv:2605.05907*, 2026.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

Florian Braun. Excess separability: Nuisance-controlled residual-stream probing for benchmark contamination detection. *arXiv preprint arXiv:2608.12652*, 2026.

Zitong Andrew Chen, Junaid Hasan, Akhil Srinivasan, Hemkesh Bandi, and Jarod Alper. Multiplication beyond groups: Stratified fourier mechanisms in transformer circuits. *arXiv preprint arXiv:2607.07066*, 2026.

Vishnu Kabir Chhabra, Ding Zhu, and Mohammad Mahdi Khalili. Neuroplasticity and corruption in model mechanisms: A case study of indirect object identification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3099–3122, 2025.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2302.03025.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, pp. 2990–2999. PMLR, 2016.

Tianyu Cui, Yogesh Kumar, Pekka Marttinen, and Samuel Kaski. Deconfounded representation similarity for comparison of neural networks. *Advances in Neural Information Processing Systems*, 35:19138–19151, 2022.

Tahereh Dehdarirad, Michael Felsberg, Gabriel Eilertsen, and Ziliang Xiong. Commutator-induced uncertainty in vaes. *arXiv preprint arXiv:2605.23449*, 2026.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. *Advances in neural information processing systems*, 34:1556–1568, 2021.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

Sheridan Feucht, Tal Haklay, Usha Bhalla, Daniel Wurgaft, Can Rager, Raphaël Sarfati, Jack Merullo, Thomas McGrath, Owen Lewis, Ekdeep Singh Lubana, et al. Arithmetic in the wild: Llama uses base-10 addition to reason about cyclic concepts. *arXiv preprint arXiv:2605.01148*, 2026.

Roy T Forestano, Konstantin T Matchev, Katia Matcheva, Alexander Roman, Eyup B Unlu, and Sarunas Verner. Identifying the group-theoretic structure of machine-learned symmetries. *Physics Letters B*, 847:138306, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

Nate Gruver, Marc Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. *arXiv preprint arXiv:2210.02984*, 2022.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024.

Stefan Heimersheim and Alex Turner. Residual stream norms grow exponentially over the forward pass. AI Alignment Forum, 2023. URL `https://www.alignmentforum.org/posts/8mizBCm3dyc432nK8`.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.07987.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, volume 2024, pp. 9944–10015, 2024.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*, 2025.

Hamza Keurti, Hsiao-Ru Pan, Michel Besserve, Benjamin F Grewe, and Bernhard Schölkopf. Homomorphism autoencoder–learning group structured representations from observed transitions. In *International Conference on Machine Learning*, pp. 16190–16215. PMLR, 2023.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning (ICML)*, 2019.

Henry Kvinge, Andrew Aguilar, Nayda Farnsworth, Grace O'Brien, Robert Jasper, Sarah Scullen, and Helen Jenne. Can neural networks learn small algebraic worlds? an investigation into the group-theoretic structures learned by narrow models trained to predict group operations. *arXiv preprint arXiv:2601.21150*, 2026.

Matthew L Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

Andrew Lee, Fernanda Viégas, and Martin Wattenberg. Tensor product representation probes reveal shared structure across linear directions. *arXiv preprint arXiv:2605.09967*, 2026.

Belinda Z Li, Zifan Carl Guo, and Jacob Andreas. (how) do language models track state? *arXiv preprint arXiv:2503.02854*, 2025a.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Zeyu Michael Li, Hung Anh Vu, Damilola Awofisayo, and Emily Wenger. Causes and consequences of representational similarity in machine learning models. *arXiv preprint arXiv:2505.13899*, 2025b.

Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

Siyi Lyu, Quan Liu, and Feng Yan. On the intrinsic limits of transformer image embeddings in non-solvable spatial reasoning. *arXiv preprint arXiv:2601.03048*, 2026.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Giovanni Luca Marchetti, Daniel Kunin, Adele Myers, Francisco Acosta, and Nina Miolane. Sequential group composition: A window into the mechanics of deep learning. *arXiv preprint arXiv:2602.03655*, 2026.

Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. The dead salmons of ai interpretability. *arXiv preprint arXiv:2512.18792*, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. *arXiv preprint arXiv:2404.08819*, 2024.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. *Advances in Neural Information Processing Systems*, 38:106423–106474, 2026.

Takeru Miyato, Masanori Koyama, and Kenji Fukumizu. Unsupervised learning of equivariant structure from sequences. *Advances in Neural Information Processing Systems*, 35:768–781, 2022.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. Liegg: Studying learned lie group generators. *Advances in Neural Information Processing Systems*, 35:25212–25223, 2022.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023a.

Neel Nanda, Andrew Lee, and Martin 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, 2023b.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.03658.

Nikhil Prakash, Tamar Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, volume 2024, pp. 8057–8082, 2024.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *Advances in Neural Information Processing Systems*, 38: 174627–174662, 2026.

Ansh Soni, Sudhanshu Srivastava, Konrad Kording, and Meenakshi Khosla. Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, pp. 2024–08, 2024.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. *arXiv preprint arXiv:2312.06581*, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024.

Muhammad Usama and Dong Eui Chang. Convergence without understanding: When language models agree on representations but disagree on reasoning. *arXiv preprint arXiv:2605.23315*, 2026.

Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

Keyon Vafa, Peter G Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models. *arXiv preprint arXiv:2507.06952*, 2025.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of llms via circuit analysis. *arXiv preprint arXiv:2502.11812*, 2025.

Wilson Wu, Louis Jaburi, Jason Gross, et al. Towards a unified and verified understanding of group-operation networks. In *International Conference on Learning Representations*, volume 2025, pp. 61683–61714, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

APPENDIX

A Models, Training, and Cost **16**

B The order-blind ceilings: definition and computation **17**

C Instruments: Definitions and Controls **18**

C.1 The decoder-free bracket readout 18

C.2 Per-run controls 19

C.3 The abelian control disqualifies non-specific metrics 19

C.4 A retired instrument: the closure/Jacobi battery 19

C.5 The Order-Blind Ceiling Moves With Both k And Length 20

C.6 The steering channel and its validity condition 20

C.7 The group-action reference class scores 1.000 20

C.8 Element decoding beyond the multiset 20

D The k-Transition Statement **21**

E Matched-Noise Nulls and the Transfer Entailment **22**

E.1 The transfer estimator 22

E.2 Why a positive transfer score is entailed 23

E.3 The shared-frame pre-test 24

F Shared Deviations: Ceilings and Pairings **24**

G Depth-Matched Null **25**

H The Causal Arm and Its Null Simulations **26**

I The Frame-Removal Impossibility **27**

J Robustness Replicates **28**

K Full Result Tables **29**

K.1 Screening, all arms 29

K.2 Depth profiles at d=8, all carriers 30

K.3 Fit diagnostics and base-model decode 31

K.4 Behavior does not track representation 31

K.5 Results not in the main text 32

L Generator Names and Model Scale **32**

M Training Dynamics **34**

N Extended Related Work **34**

A MODELS, TRAINING, AND COST

Roster. Seven instruction-tuned checkpoints and one pretrained (Llama-3-8B), all on the Hugging Face Hub and ungated except Llama-3-8B (Meta Llama 3 Community License), all with the seven LoRA target modules present under their standard names (`q/k/v/o_proj`, `gate/up/down_proj`), so the identical training script runs on every one unmodified.

Table 4: Model roster. Every model passes through the same training and measurement code path.

Model	Family	Blocks	Role
Qwen2.5-1.5B-Instruct	Qwen	28	reference carrier; k-sweep, probe, robustness, ceilings
Qwen2.5-3B-Instruct	Qwen	36	fills the 1.5B→7B gap
Qwen2.5-7B-Instruct	Qwen	28	screening; steering (Appendix K.5)
Mistral-7B-Instruct-v0.3	Mistral	32	screening
OLMo-2-1124-7B-Instruct	OLMo-2	32	screening; family-diversity entry; ceiling B
OLMo-2-1124-13B-Instruct	OLMo-2	40	second OLMo-2 size (within-family peak-depth test)
granite-3.1-8b-instruct	Granite	40	fifth family (Apache-2.0)
Llama-3-8B	Llama	32	screening; non-carrier

Two models were considered and not run, for reasons that are part of the protocol. **Phi-3.5-mini-instruct** (MIT, ungated) fuses attention and MLP projections into `qkv_proj`/`gate_up_proj`, so five of the seven LoRA target modules do not exist in it; it would receive a 2-module adapter against every other model's 7-module one, confounding "not a carrier" with "trained differently." It was rejected on the identical-pipeline rule. **Gemma-2-9b-it** is licence-gated and was not obtained.

Training. LoRA (Hu et al., 2021) with r=32, α=64, dropout 0.05, seven target modules, 12 epochs, on the k=12, θ=0.6 non-abelian composition task, seed 0 unless stated. The abelian control arm uses the same recipe on commuting generators (rotations of different angles about one common axis). The Llama-3-8B adapter was produced by exactly this pipeline (the cross-model script imports and calls the training routine unchanged), so its non-carrier verdict is a screening result, not a differently obtained one.

Measurement grid. Nine layers at matched depth fractions (14/29/39/50/61/71/82/89/96% of blocks), three residual sub-dimensions ($d \in \{3, 8, 16\}$), 27 cells per arm; 500-fold permutation test per cell; the 6 of 66 generator pairs with $v_i \times v_j = 0$ excluded from every cell. The 28-, 32-, 36- and 40-block models are compared at matched depth, not matched index.

Cost. All 7B/8B/13B training ran on one `71gb` MIG slice (a 7B LoRA fine-tune does not fit a `35gb` slice). The 3B and 1.5B fit the smaller slice. Fidelity scoring and all null simulations are CPU-only.

Table 5: GPU cost per screening row (Llama-3-8B was trained inside the cross-model pair and not timed separately; Qwen2.5-1.5B was only re-dumped). A full row including depth-null inputs costs 20–31 GPU-minutes; the frame-removal branch (Appendix I) stopped at its CPU gate and spent no GPU time.

Model	train	dump (FT)	dump (base)	fidelity+stats	total
Qwen2.5-7B	608 s	292 s	242 s	101 s	21 min
Mistral-7B-v0.3	650 s	229 s	184 s	124 s	20 min
Qwen2.5-3B	803 s	387 s	247 s	190 s	27 min
OLMo-2-7B	740 s	264 s	240 s	191 s	24 min
OLMo-2-7B, seed 1	737 s	329 s	—	32 s	18 min
OLMo-2-13B	1172 s	256 s	241 s	214 s	31 min
Granite-3.1-8B	828 s	264 s	193 s	184 s	24 min
Qwen2.5-1.5B (re-dump only)	—	198 s	196 s	96 s	8.2 min

Table 6: Bracket geometry by task size, computed from the generator set alone. At k=6 every commutator direction is one of six the model already has, at the one magnitude it has seen. From $k \geq 10$, a substantial fraction of commutator words has a bracket that lies off every generator axis, and a sign-blind bag can admit up to eight distinct labels. Full statement in Appendix D.

	k=6	k=8	k=10	k=12	k=16
bracket lands on an existing generator axis	100%	83%	70%	60%	68%
distinct bracket directions	6	8	12	18	24
distinct bracket magnitudes	1	2	3	3	3
max. distinct labels per sign-blind bag	2	6	8	8	8
zero-bracket pairs	20%	14%	11%	9%	7%

Perplexity of fine-tuned carriers. Token-level perplexity on the WikiText-2 (raw, v1) test set, sliding window of 1024 tokens with stride 512 so every token but the first is scored exactly once, NLL in fp32 from fp16 logits, adapters loaded (not merged), base tokenizer without special tokens; 2.9–3.0×10^5 tokens scored per model.

model	base	non-abelian LoRA	abelian LoRA	ratio (FT / base)
Qwen2.5-1.5B-Instruct	9.212	9.931	9.862	1.078 (abelian 1.071)
OLMo-2-7B-Instruct	8.091	8.617	—	1.065

B THE ORDER-BLIND CEILINGS: DEFINITION AND COMPUTATION

The *commutator* split is the $k(k-1)$ words $w_{ij} = g_i g_j g_i^{-1} g_j^{-1}$ over ordered pairs of distinct signed generators (Figure 2B). It isolates order-sensitivity because all eight sign-and-order variants of an axis pair contain the same four tokens. Each generator is a rotation by θ about a fixed axis, with both signs; the label of a word is $\log(R) \in \mathbb{R}^3$ with each Cartesian component binned into 31 cells of 0.203 rad over $[-\pi, \pi]$ (Section 3.1). Because w_{ij} realizes the bracket, its net rotation is small ($\approx \theta^2 \|[X_i, X_j]\|$), so the 132 words occupy only 31 distinct labels; the 12 words built from the $\pm$ pairs all land in the identity cell ($g_j = g_i^{-1}$, zero bracket).

An *order-blind* predictor sees the generators of a word but not their order, and must return one label per equivalence class of words it cannot distinguish. We report the best achievable accuracy (the majority label per class, summed over words) at two resolutions:

- **sign-blind** (blind to sign *and* order): the class is the unordered axis pair, equivalently the bag of the word's four tokens, so all sign/order variants of a pair collapse into one bag (21 bags);
- **sign-aware** (sees the signed pair, blind to its order): the class is $\{g_i, g_j\}$, so only the two orders of a fixed signed pair collapse (66 bags). Since the token bag is the same for all eight variants, this predictor must read which two tokens come first; it is order-blind only within the pair.

Sign-blind (0.386). The 21 axis-pair bags fall into four types by how the binning collapses their eight (or two) sign/order variants:

bag type	example	#bags	words/bag	distinct labels	majority×#bags
same axis	$\{x,x\}$	6	2	1	$2 \times 6 = 12$
two-label	$\{x,y\}$	6	8	2	$4 \times 6 = 24$
six-label	$\{x,xy\}$	6	8	6	$2 \times 6 = 12$
eight-label	$\{xy,xz\}$	3	8	8	$1 \times 3 = 3$
				total	$51/132 = 0.386$

The label count per bag follows the angle between the two axes: for a same-axis pair the commutator is the identity; for perpendicular pairs (two coordinate axes, or a coordinate and the diagonal orthogonal

to it) the eight variants collapse onto two cells; for a coordinate inside a diagonal (45°) they spread over six; and for two diagonals (60°) all eight stay distinct.

Sign-aware (0.545). Grouping instead by signed pair $\{g_i, g_j\}$ gives 66 bags of two words (the two orders). For the 6 same-axis bags both orders are the identity (majority $2/2$); for the 60 distinct-axis bags the two orders differ by the nonzero bracket and land in different cells (majority $1/2$). Hence $6 \cdot 2 + 60 \cdot 1 = 72$, i.e. $72/132 = 0.545$. Restricted to the 120 distinct-axis (nonzero-bracket) words, the sign-blind ceiling is $39/120 = 0.325$ and the sign-aware ceiling is exactly $60/120 = 0.500$: on the words where order is the whole question, an order-blind predictor is a coin flip.

The two ceilings differ in whether the predictor knows which signed pair was composed. Beating the sign-blind ceiling (0.386) shows a model reads position-dependent information from the word. Beating the sign-aware ceiling (0.545) is the test of bracket use, because the one thing that predictor lacks (the order within the pair) flips the sign of the bracket: $[g_j, g_i] = -[g_i, g_j]$. Both are computed by the same enumeration on each configuration's generator set and its analytic labels; both are reported across task size in Appendix C.5. The computation is deterministic given the generator set and the binning, and reproduces the 31 distinct labels and the 0.386 ceiling stated above exactly.

C Instruments: Definitions and Controls

C.1 The decoder-free bracket readout

If a residual subspace carries a faithful linear image of $\mathfrak{so}(3)$, there is a linear map $\Phi : \mathfrak{so}(3) \to \mathfrak{gl}(d)$ with $M_i = \Phi(v_i)$ and $[M_i, M_j] = \Phi(v_i \times v_j)$ for the fitted generator matrices M_i and their analytic rotation vectors v_i. Φ is fit from the 12 (v_i, M_i) pairs alone—four unknowns per matrix entry (three linear coefficients plus the common affine term C of (i) below) against twelve equations—and the bracket identity is then tested on pairs Φ never saw as brackets. The headline is $\text{bracket}_{\cos}$ (Eq. 2), the cosine between the realized commutator and its prediction. It replaces an earlier closure/Jacobi battery that failed its positive control (Appendix C.4).

Three design choices matter. **(i) An affine term is fit and removed.** The raw fitted generators share a large common component C—a generic per-step residual update, not part of $\mathfrak{so}(3)$. Fitting Φ through the origin fits far worse, and C contaminates the bracket via $[C+\Phi(a), C+\Phi(b)] = [\Phi(a), \Phi(b)] + [C, \Phi(b)] - [C, \Phi(a)]$. On the original four-layer Qwen2.5-1.5B fine-tune (layer 25, d=8; the dumps behind Table 8's caption), fitting C explicitly gives Φ-fit $R^2 = 0.91$ against the zero model, 0.82 against the mean (which removes C's own share), and 0.39 through the origin. **(ii) Degenerate pairs are excluded and counted**: where $v_i \times v_j = 0$ the cosine is noise-vs-noise; 6 of 66 pairs on the non-abelian arm, all 66 on the abelian arm. **(iii) Significance is a permutation test**: the null is the `shuffled` control run 500 times—how often does a wrong $v_i \leftrightarrow M_i$ pairing reach the observed agreement. The PCA basis is fit on centred states while uncentred states are projected, so the state mean lands in the common affine component C; this is one reason C must be fit and removed before the bracket is scored.

Noise ladder. The exact planted embedding is progressively corrupted with noise of relative norm ε per generator, over sixteen random embedding similarities ($\theta = 0.6$, $d = 8$):

ε	mean ± sd	[min, max]	single draw (original table)
0	1.000 ± 0.000	—	1.000
0.25	0.448 ± 0.248	$[-0.010, 0.830]$	0.698
0.5	0.259 ± 0.176	$[-0.004, 0.542]$	0.436
1.0	0.123 ± 0.085	$[0.012, 0.276]$	0.227
2.0	0.016 ± 0.047	$[-0.057, 0.106]$	0.055
4.0	0.008 ± 0.017	$[-0.019, 0.035]$	0.035

The ladder is far more dispersed than a single draw suggests: at $\varepsilon = 0.25$ the standard deviation equals the mean and the range spans 0 to 0.83, and the originally published single draw sits near the top of that distribution. The reference carrier's 0.487 interpolates to $\varepsilon \approx 0.23$ on the mean ladder, but is consistent with ε anywhere from ≈ 0.1 to ≈ 0.6 depending on the embedding basis. We therefore do not use ε as a calibration in the main text; the reportable quantity is the fraction of the planted ceiling

(0.49 for the reference carrier, 0.51 for the strongest), and the base arm of the original four-layer dump (-0.010 at layer 25, d=8) sits below the $\varepsilon = 4.0$ mean (0.008), i.e. at the noise floor. A caveat on magnitude: the fitting procedure injects spurious non-commutativity of ≈ 0.3 (the abelian arm's fitted generators do not commute: normalized $\|[M_i, M_j]\| = 0.278$ at layer 25, d=8 of the four-layer dump, 0.24–0.43 across its four layers); this floor is isotropic, dilutes the cosine toward zero without biasing its direction, and is priced in by the permutation null. Bracket *magnitude* alone is therefore never quoted as evidence.

C.2 Per-run controls

Table 7: Instrument controls, every screening run. `random` draws matrices at the arm's own generator scale; because the cosine is scale-invariant, its score depends only on (seed, k, d)—it checks the code path, not the model. `shuffled` is one draw from each cell's own permutation null (the same permutation for every cell of a given d, so a run's 27 values are correlated, not independent draws) and is read against that cell's null sd: on the widest run (OLMo-2-7B), $z = -0.29 \pm 0.97$ with $0/27$ cells beyond $|z| > 3$.

Run	`exact` (all 27 cells)	`random` (d=3/8/16)	`shuffled`, mean ± sd (27 cells)
Qwen2.5-1.5B	1.000000	$-0.066/+0.006/+0.004$	-0.011 ± 0.033
Qwen2.5-3B	1.000000	$-0.066/+0.006/+0.004$	-0.004 ± 0.023
Qwen2.5-7B	1.000000	$-0.066/+0.006/+0.004$	$+0.003 \pm 0.025$
Mistral-7B-v0.3	1.000000	$-0.066/+0.006/+0.004$	-0.005 ± 0.029
OLMo-2-7B	1.000000	$-0.066/+0.006/+0.004$	-0.015 ± 0.049
OLMo-2-13B	1.000000	$-0.066/+0.006/+0.004$	$+0.001 \pm 0.046$
Granite-3.1-8B	1.000000	$-0.066/+0.006/+0.004$	-0.016 ± 0.032
Llama-3-8B	1.000000	$-0.066/+0.006/+0.004$	-0.002 ± 0.031

C.3 The abelian control disqualifies non-specific metrics

The probe instrument (ridge decoder of the running group element; 3000 words, prefix-disjoint evaluation, layers $8/14/19/25$ of Qwen2.5-1.5B) reports four metrics (decode R^2, homomorphism ratio, commutator alignment, and a representation-level order-swap statistic; the main text's fourth candidate, update-fit R^2, comes from the generator dumps). Only commutator alignment isolates the bracket; the first two are matched or beaten by the abelian adapter, whose true algebra is one-dimensional with every bracket exactly zero:

Table 8: Probe metrics by layer; full version of Table 1, with the base column restored. At layer 25 the abelian arm reaches decode $R^2 = 0.947$ and homomorphism ratio 0.115, matching the non-abelian fine-tune; both metrics measure "represents its own task's group element and composes multiplicatively," which a commuting group does equally well. `update_fit_`R^2 (from the generator dumps, not the probe) is likewise highest on the abelian arm at layers 8/14 ($0.957/0.973$ vs. $0.919/0.905$), and a representation-level order-swap statistic fails to separate arms at all (base scores higher at three of four layers). Only commutator alignment (0.882 vs. 0.517) is bracket-specific.

Layer	decode R^2 (FT / base / abelian)	hom. ratio ↓ (FT / base / abelian)	commutator alignment (FT / base)
8	0.985 / 0.941 / 0.643	0.079 / 0.174 / 0.631	0.859 / 0.605
14	0.982 / 0.889 / 0.702	0.088 / 0.218 / 0.547	0.864 / 0.526
19	0.986 / 0.818 / 0.937	0.079 / 0.272 / 0.163	0.855 / 0.505
25	0.983 / 0.741 / 0.947	0.072 / 0.301 / 0.115	0.882 / 0.517

C.4 A retired instrument: the closure/Jacobi battery

An earlier instrument asked whether learned generators close into an algebra, fitting structure constants freely. Seven real configurations agreed (closure residual ≈ 1.0, $\text{jacobi}_{\text{rel}}$ 0.21–0.36, unchanged across a 16× sub-dimension range and across Qwen/Llama). None of it supports a conclusion, because the battery fails its positive control: run on the k=12 fine-tune whose generators

are recovered to 0.002–0.003 rad, it gives no separation from the base arm (which wins at layer 8), and the *abelian* arm scores best (jacobi$_{\text{rel}}$ 0.002–0.051 vs. 0.036–0.104 for the fine-tune), while analytic $\mathfrak{so}(3)$ generators through the identical code return a closure residual of 0 and jacobi$_{\text{rel}} \approx 3 \times 10^{-16}$. Two root causes: free structure constants are massively non-unique with 12 generators spanning a rank-3 space (condition numbers 53–350 on the abelian arm), and "is the algebra closed" is satisfied trivially by a commuting group. The replacement (Appendix C.1) asks "is the bracket the right bracket," which only a known target algebra makes answerable.

C.5 The Order-Blind Ceiling Moves With Both k And Length

For each $(k, \theta, \text{length})$ the ceiling is the best accuracy of a predictor that does not see the order of generators in a word (for the sign-aware row, only the order within the pair). It moves steeply with both k and length, so margins are never pooled:

k	6	8	10	12	14	16
commutator, sign-blind	0.600	0.500	0.433	0.386	0.352	0.325
commutator, sign-aware	0.600	0.571	0.556	0.545	0.538	0.533
`longer` ceiling	0.481	0.760	0.905	0.962	—	0.993

The `longer` (length-generalization) split is not reported as a bracket result: at k=12 its ceiling is 0.962—a bar a bag-of-generators model clears 96% of the time—and the order-dependent fraction of the task falls to 0.5% at length 8. The ceiling-free quantity is the per-step survival rate from a log-linear fit $\text{acc} = C\,p^{L-1}$ (free intercept in each range): 0.844 inside the training range (lengths 2–4) and 0.381 beyond it (5–8, log-linear $R^2 = 0.994$). Geometric decay is evidence *for* compositional behavior (a length-bounded lookup would break, not trace a line); the extrapolation penalty is a halving of per-step reliability.

C.6 The steering channel and its validity condition

Steering-vector experiments (Appendix K.5) are reported only where a single-generator efficacy check shows the channel is live and the injection is in the sub-unit regime ($\alpha < 1$). On Qwen2.5-7B-Instruct, efficacy rises monotonically $+0.077 \to +0.489$ over $\alpha = 0.1 \to 0.8$ (recorded at the time, not persisted to a result file); a null bracket correlation on a live channel is reportable, a null on a dead one is not (this rule voids the abelian arm of the causal experiment, Appendix H).

C.7 The group-action reference class scores 1.000

The `exact` control plants algebra elements $M_i = \Phi(v_i)$. A natural worry is that a mechanism faithfully implementing the group action instead updates by $e^{\theta X_i}$, whose generator $e^{\theta X_i} - I$ has commutators with higher-order Baker–Campbell–Hausdorff terms that a linear Φ does not predict; fed directly into the Φ fit, such generators cap at $0.975/0.947/0.908/0.770$ for $\theta = 0.3/0.45/0.6/1.0$ ($d = 8$, mean over 16 similarities). The production readout does not read generators that way: it estimates the transition operator A_i by ridge regression on state pairs and sets $M_i = \operatorname{Re}\log A_i$, and the principal logarithm returns θX_i exactly for $\theta \leq 1$. Planting $A_i = e^{\theta X_i}$ at the transition level and running the full path (synthetic state pairs $\to$ ridge fit, $R^2 = 1.000000$ over 330 transitions $\to$ logarithm $\to$ affine removal) returns 1.0000 ± 0.0000 at every θ and every similarity; forcing the affine term to zero changes nothing. There is one ceiling, 1.0, for both reference classes, and the 0.908 figure is reported only as what a readout without the logarithm would give. Under matched noise the two plants coincide, as they must once the logarithm turns a noisy operator into a noisy generator (0.373 ± 0.298 vs. 0.448 ± 0.248 at $\varepsilon = 0.25$); noise applied to the state pairs rather than the operator is averaged down by the ridge fit (0.756 ± 0.330), so the state-level ladder is the more optimistic calibration and neither is used in the main text.

C.8 Element decoding beyond the multiset

Why raw decode R^2 is uninformative. At $\theta = 0.6$ and lengths 2–4 the composed rotation is largely determined by the multiset of generators: the R^2 of a predictor that outputs the mean target over all orderings of the same multiset is 0.918 (rotation vector) and 0.903 (matrix entries) on the probe's

evaluation prefixes, $0.925/0.909$ on the held-out split, and $0.000/-0.160$ on the commutator split, where the multiset is uninformative by construction. Raw base-arm decode R^2 of 0.86–0.95 (Table 9; 0.79–0.97 under the word-level split of Table 20) therefore sits at the additive ceiling and shows only that token identities are linearly readable.

Residual probe. Target $Y - \mathbb{E}[Y \mid \text{multiset}]$, the multiset mean computed analytically over all orderings; ridge probe as in Appendix C.3 (3000 words, prefix-disjoint evaluation), best of four layers, permutation p (100 permutations, floor 0.0099). Gates on synthetic residual streams: one linearly encoding the multiset and the true composed element must score near 1 (observed 0.999); one encoding only the multiset must score near 0 (-0.134); one encoding the composed element alone, without the multiset, scores -0.015, since the target subtracts a multiset mean that stream cannot supply. A first construction that standardized train and evaluation rows separately scored 0.943 on the positive gate; it is a construction error, is kept on record, and was rebuilt as a single stream before any arm was scored. Reference bounds, computed post hoc and not part of the decision rule: a representation knowing the multiset and the *last* token can reach 0.545, and one knowing first and last 0.882 (both analytic); a synthetic stream encoding the multiset and the full composed element scores 0.989 under the same probe (the composer ceiling). A Hewitt–Liang control task (a random additive function of the same tokens) is decoded at 0.86–0.99 by every arm and is reported for completeness; being additive, it is easier than the residual and does not discriminate.

Table 9: Order-dependent residual decoding, all arms ($p = 0.01$ throughout). Every base arm sits at or below the last-token bound (0.545) except two that exceed it by 0.01–0.03; every fine-tuned arm sits above the first-plus-last bound (0.882) or within 0.03 of it. The pre-specified decision rule (residual $R^2 \geq 0.10$ at $p < 0.05$) fires for all sixteen arms and is reported as written; the analytic bounds are the post-hoc reading. The abelian adapter has an identically zero residual and is not scored.

model	arm	best layer	residual R^2	raw R^2	commutator-split residual R^2
Qwen2.5-1.5B	FT	19	0.870	0.986	0.838
Qwen2.5-3B	FT	25	0.859	0.983	0.843
Qwen2.5-7B	FT	8	0.901	0.988	0.848
Mistral-7B-v0.3	FT	22	0.889	0.982	0.852
OLMo-2-7B	FT	9	0.947	0.995	0.925
OLMo-2-13B	FT	20	0.933	0.992	0.912
Granite-3.1-8B	FT	20	0.931	0.992	0.914
Llama-3-8B	FT	16	**0.864**	0.986	**0.827**
Qwen2.5-1.5B	base	8	0.527	0.941	0.486
Qwen2.5-3B	base	10	0.571	0.949	0.493
Qwen2.5-7B	base	8	0.553	0.946	0.428
Mistral-7B-v0.3	base	16	0.356	0.860	0.337
OLMo-2-7B	base	9	0.545	0.947	0.499
OLMo-2-13B	base	12	0.532	0.947	0.489
Granite-3.1-8B	base	12	0.495	0.919	0.387
Llama-3-8B	base	9	0.407	0.890	0.368

D THE k-TRANSITION STATEMENT

Task structure, computed from the generator set alone. At k=6 the generators are $\pm x, \pm y, \pm z$; every bracket is proportional to the third axis, itself a generator direction. From k=12 the six axes include the three diagonals xy, xz, yz, which are not mutually orthogonal. Table 6 gives the geometry by task size.

Statement. For $k \leq 8$, the brackets lie mostly on existing generator axes (all of them at k=6, 83% at k=8) at only one or two magnitudes, and a token bag admits at most six labels. From $k \geq 10$ a constant fraction (30–40%) of commutator words has a bracket off every generator axis, and a bag can admit up to eight labels (three of the 21 bags at k=12). We predicted, before the sweep was run, that a positive margin over the sign-blind ceiling would first appear at k=8 $\rightarrow$ 10, and the training script refuses $k \leq 8$ without an explicit override. This is a prediction about learned behavior, not a

necessary condition: at k=8 a sign-aware predictor already exceeds the sign-blind ceiling (0.571 vs. 0.500; Appendix C.5), and the two ceilings coincide only at k=6 (0.600).

The sweep. Three seeds per k for Qwen2.5-1.5B; for OLMo-2-7B, three at k=8, five at k=10 and two at k=12 (the pre-specified k=12 top-up was stopped by its config-identity gate); epochs fixed at 12, every margin against its own ceiling:

Table 10: The k-sweep in two families. Commutator margins are against each size's sign-blind ceiling; held-out margins against that split's own order-blind ceiling. Bold commutator margins have a 95% t-interval above zero ($p = .016$, .006 at $k = 12, 16$); at k=10 the margin is +0.067 (one-sided $p = .030$, interval $[-0.006, +0.140]$). OLMo-2 95% t-intervals: commutator $[-0.055, +0.068]$ at k=10 and $[-0.376, +0.490]$ at k=12; held-out $[+0.079, +0.147]$ at k=10. The held-out column is the "learned the general task" check the commutator column requires: it is at or above ceiling from k=8 in both families (Qwen's k=16 value, −0.005, is within noise of its ceiling; k=6 in Qwen is below its held-out ceiling, a word-space exhaustion, and is not run for OLMo-2). Probe gap is the bracket-specific alignment, FT minus base, at the deepest probed layer. The prediction (no commutator gain before k=10) holds in both families; a positive margin from k=10 (significant from k=12) appears in Qwen and not in OLMo-2 (five seeds, per-seed −0.067 to +0.044). OLMo-2's $\text{bracket}_{\cos}$ at k=10 is given for seeds 0 and 3; seed 4 (+0.420, 25/27) is reported as void under a pre-specified absolute-value gate on the `shuffled` control, applied to the added adapters, that we subsequently found mis-calibrated (it flags ordinary null draws at d=3, where seed 0's existing adapter would fail it too; under the z-criterion of Table 7 seed 4 passes), and is not un-voided. Qwen's $\text{bracket}_{\cos}$ at $k \leq 8$ was not dumped; the probe gap carries the k=8 representational claim there.

			Commutator		Held-out			
Model	k	Seeds	Margin	Ceiling	Margin	Ceiling	Probe gap	$\text{bracket}_{\cos}$ best d=8
Qwen2.5-1.5B	6	3	−0.378	0.600	−0.112	0.507	−0.139	—
	8	3	−0.196	0.500	+0.237	0.465	+0.305	—
	10	3	+0.067	0.433	+0.109	0.596	+0.323	+0.219
	12	3	**+0.139**	0.386	+0.019	0.674	+0.364	+0.487
	16	3	**+0.268**	0.325	−0.005	0.767	+0.364	+0.407
OLMo-2-7B	8	3	−0.167	0.500	+0.248	0.465	+0.331	**+0.628**
	10	5	+0.007	0.433	+0.113	0.600	+0.304	+0.590 / +0.559
	12	2	+0.057	0.386	+0.026	0.673	+0.331	+0.514 / +0.413

The transition is specific, not generic. The held-out split's largest jump is at k=6 → 8 (margin $-0.112 \rightarrow +0.237$) and its margin *declines* thereafter ($+0.109, +0.019, -0.005$) as its ceiling climbs; a capacity or data story predicts that pattern, not the commutator one. Training-set size is ruled out on shape: data grows most at k=6 → 8 (+238%, 1012 → 3425 words), where the commutator margin does not move, and only +38% at k=8 → 10, where it does. A token-matched control (k=12 retrained on 3174 words, fewer than k=8's 3425) gives commutator margin +0.118 at five seeds ($p = 0.0067$), excluding training-set size as the driver.

The reference run. At k=12 (the three seeds of the sweep, 95% t-intervals): commutator 0.525 ± 0.110, margin $+0.139 \pm 0.110$ over the 0.386 sign-blind ceiling ($p = 0.016$; per seed 0.576, 0.492, 0.508); held-out 0.693 ± 0.038, margin $+0.019 \pm 0.041$ ($p = 0.096$); the abelian control (seed 0) sits at its ceilings (0.998 against 1.000 held-out, 1.000 against 1.000 commutator) and never exceeds them. The held-out leg is the fragile one—its margin is intrinsically tiny and does not clear significance at three seeds—and is never quoted without its seed count.

E Matched-Noise Nulls and the Transfer Entailment

E.1 The transfer estimator

Two models A, B are trained on the same algebra with generator sets related by a rotation Q_0 and axis permutation π, $v_B[\pi(i)] = Q_0 v_A[i]$, where π permutes the axes and both signed generators of

an axis move together. Per model and cell, the fitted generators give Φ and an algebra-coordinate map $\gamma(K) = \arg\min_w \|K - \Phi(w)\|_F$; γ is invariant to the PC basis (both K and each $\Phi(e_a)$ transform alike under an orthogonal change of basis), which is what makes two models of different widths comparable with no basis alignment.

Folds close circularity by partition. k=12 gives 6 axes, $\binom{12}{2} = 66$ pairs, 6 degenerate, 60 scored. One fold per unordered axis pair: fit Φ and the alignment on the 8 generators of the other 4 axes, score the 4 pairs whose *both* generators were held out. $15 \times 4 = 60$ pairs, each scored exactly once, and no generator entering a fit contributes to any pair that fold scores. Every fit set spans $\mathfrak{so}(3)$: the only coplanar triples among x, y, z, xy, xz, yz are $\{x, y, xy\}$, $\{x, z, xz\}$, $\{y, z, yz\}$, and adding any fourth axis leaves the plane. The alignment $\hat{R}$ is orthogonal Procrustes on 8 paired γ-coordinates constrained to $\det = +1$ (a reflection has $(Ra) \times (Rb) = -R(a \times b)$ and inverts brackets rather than carrying them).

Positive controls, provable rather than hoped for. Planting two exact $\mathfrak{so}(3)$ embeddings on different generator sets in independent random bases gives $c^A_{ij} = v_i \times v_j$ and $c^B = Q_0(v_i \times v_j)$ by Eq. 3, so the score is 1 by identity:

construction	required	observed
planted exact, $d = 3/8/16$, independent bases, $Q_0 = 118.5°$	1.000	1.000000
same algebra, inequivalent representation (spin-2$\oplus$ spin-1 vs. spin-1)	1.000	1.000000
pair partition	60 / 6	exact, every run

The second row is a scope limit made measurable: the readout compares the 3-dimensional algebra shadow and is blind to which *representation* of $\mathfrak{so}(3)$ a model carries. The natural first estimator—a $\mathfrak{gl}(d)$ intertwiner S with $S\Phi_A(v) = \Phi_B(Q_0 v)S$—was rejected because its control degenerates: both planted arms are spin-1 $\oplus\,(d-3)$ trivials, so by Schur's lemma the intertwiner space has dimension $1 + (d-3)^2$ (26 at d=8, 170 at d=16) and the minimum-residual solution may sit entirely on the trivial block, where the transported bracket is zero and the cosine undefined.

E.2 Why a positive transfer score is entailed

Φ_A is fit so that $\gamma_A(X_i) \approx v_i$; hence $\hat{R} \approx Q_0$ before any measurement. If each model separately carries the bracket, both realized bracket fields sit near $v_i \times v_j$ (in their own frames) and agree for that reason alone. Under independent per-model angular errors α, β with uniform azimuth, $\mathbb{E}[\cos] = \cos\alpha \cos\beta$. The least-squares γ moreover discards error components outside the Φ image, so the realized null exceeds even that product, increasingly with d:

sub-dimension	3	8	16
null mean $\div$ product rule	1.16	3.14	4.73

Matched-noise construction. Plant two exact $\mathfrak{so}(3)$ embeddings in independent random bases ($Q_0 = 118.5°$ with an axis permutation for the planted controls, the simulator's default $Q_0 = 152.9°$ for the product-rule ratios above, and $Q_0 = I$ with the identity correspondence for Table 11 and the pre-test, matching the shared-frame data), inject independent per-generator noise, calibrated by bisection so each arm reproduces a target within-model $\text{bracket}_{\cos}$, and run the identical fold/Procrustes/γ path; 500 draws for the operating-point table below and the ratio table, 300 per cell in the shared-frame pre-test and Appendix F (plus 300 correspondence permutations per cell in the pre-test). Sanity gates: at zero noise the simulator returns 1.000000 at $d = 3/8/16$, reproduces the planted-control requirements, and scores 60 pairs with 6 degenerate. At the measured operating points (d=8):

Table 11: The matched-noise null for the transfer statistic, at d=8 with model A fixed at Qwen2.5-1.5B L23. Two models sharing nothing score 0.19–0.30 on average with p_{95} up to 0.75 (500 draws; an independent persisted 500-draw run of the first row gives mean 0.192, sd 0.194, p_{95} 0.562). The correspondence-permutation null the design originally specified, evaluated with 300 permutations on one simulated matched-noise draw per cell of the Qwen2.5-1.5B × Qwen2.5-7B pre-test, centres at +0.004 ($p_{95} \approx 0.10$–0.22) and would have been cleared by independent approximation alone.

Model B	$\cos_A$	$\cos_B$	Null mean	sd	p_{95}	Product rule	Ratio
Mistral-7B L26	0.487	0.126	0.192	0.193	0.541	0.061	3.14×
Mistral-7B L23	0.487	0.160	0.223	0.214	0.609	0.078	2.87×
Qwen2.5-7B L23	0.487	0.264	0.295	0.257	0.751	0.128	2.30×

The null is insensitive to noise structure: with low-rank (rank 1–2) or anisotropic (4:1, 10:1) errors calibrated to the same per-arm $\text{bracket}_{\cos}$, null means move by at most 0.01 and the p of the Qwen2.5-7B L23 cell (third row) stays in 0.11–0.14 under the 500-draw nulls (0.150 under the 300-draw null of the pre-test below).

E.3 The shared-frame pre-test

All carriers are fine-tuned on the same generator set, so the estimator runs on real pairs with $Q_0 = I$—both models supervised toward identical coordinates, the most favourable condition and hence an upper bound on any frame-removed result.

pairing	interpretable cells	cells clearing matched null	min p
Qwen2.5-1.5B × Qwen2.5-7B	17 / 27	0 / 17	0.0565
Qwen2.5-1.5B × Mistral-7B	8 / 27	0 / 8	0.123

A cell is interpretable only when both models independently pass their own $\text{bracket}_{\cos}$ test there. The best cell of the stronger pairing (L25/d=8) returns +0.677 at $p = 0.0565$; the peak-fidelity cell (L23/d=8) returns +0.632 at $p = 0.150$. Transfer and frame-fixed variants agree to within 0.03 in every cell (0.001 at the best cell), so the fitted alignment contributes nothing over the known frame—the entailment, measured. The statistic is withdrawn, not demoted.

F Shared Deviations: Ceilings and Pairings

Definition. Same folds, same γ. For each pair, $\delta_{ij} = c_{ij} - a_{|w_{ij}|} w_{ij}$, where $w_{ij} = v_i \times v_j$ and one scale a is fit per distinct bracket magnitude (three at k=12) *within each model alone*; $\text{dev}_{\cos}$ is the cosine between the two models' deviation fields (60 pairs × 3 components), with B mapped into A's frame by $Q_0^\top$. The per-magnitude scale is a nuisance step: both models shrink brackets as a function of magnitude because they share the task, and leaving that in would score task-sharing as model-sharing. Fit within each model separately, it can create no cross-model agreement.

Two-sided control. Both arms exact $\mathfrak{so}(3)$ on the same perturbed generator set (independent bases) must score 1; the same with deviations drawn separately must score 0. Observed: shared $+1.0000 \pm 0.0000$, independent -0.0195 ± 0.1698, at arm fidelities of 0.87–0.88. The control is already $Q_0 \neq I$ (152.9° with a non-identity permutation, so some pairs take the antisymmetric sign-flip path). The bug this control caught: the shared arm first returned $\approx +0.6$ (pre-fix run, not persisted), because where the correspondence maps A's pair (i, j) onto B's stored pair in the opposite order, both B's realized bracket and its analytic w are negated; looking the pair up without the sign flip dragged a perfectly shared field down. Fixed, the control returns 1.0000 exactly; the threshold was not moved.

Table 12: Shared-deviation statistic. Ceilings are the same model at two seeds, s0 and s1 (LoRA init, batch order, and the train/held-out partition all reseed): the highest any genuine pair could reach, ≈ 0.15–0.23, not 1.0. Cross-model pairings are shared-frame throughout.

Pairing	Interp. cells	Clearing	Binomial p	Mean $\text{dev}_{\cos}$	Positive	Sign test
Ceiling A: Qwen2.5-1.5B s0 × s1	16/27	6	0.0001	+0.2292	15/16	0.0005
Ceiling B: OLMo-2-7B s0 × s1	24/27	9	$< 10^{-4}$	+0.1499	20/24	0.0015
Qwen2.5-1.5B × OLMo-2-7B	21/27	5	0.0032	+0.0657	14/21	0.189
Qwen2.5-1.5B × Qwen2.5-7B	17/27	3	0.0503	+0.0120	9/17	1.000

The sign split discriminates. A real shared-deviation effect makes nearly every cell positive (ceilings $15/16$, $20/24$). Neither cross-model pairing does. The Qwen×Qwen pairing's three high cells (68%, 88%, 185% of ceiling at their own cells) sit in a field otherwise indistinguishable from coin flips—excursions, not a weak version of the ceiling's pattern. The strongest pairing clears the matched-noise null in $5/21$ cells (binomial $p = 0.0032$) but fails the sign test ($14/21$, $p = 0.19$); pairing each model's own peak cell rather than matched depth reaches $+0.412$ at Qwen L23 × OLMo L16 (one cell, post-hoc, 142% of the OLMo-side per-cell ceiling of $+0.290$). Percentages above 100% are reported as flagged and read as "the deviation field is task-driven rather than run-driven," never as excess agreement: a replicate ceiling bounds how reproducibly one model re-derives its own deviations across seeds, not how much two models at fixed adapters can agree.

Per-cell ceilings. The ceiling varies more than threefold across cells (Ceiling A: L25/$d3$ $+0.736$, L27/$d8$ $+0.511$, L25/$d8$ $+0.460$, L23/$d3$ $+0.423$, L23/$d8$ $+0.417$, L20/$d8$ $+0.239$, L23/$d16$ $+0.121$), so every ratio is quoted against the ceiling at the same cell, with the mean quoted only alongside.

Caveat. Cells are not independent (sub-dimensions are nested and adjacent layers share subspace), so a binomial count over cells overstates the effective number of tests; the counts above are reported with this caveat regardless of outcome.

G Depth-Matched Null

Why the per-cell permutation null cannot test a profile. It holds the fitted set $\{M_i\}$ fixed and permutes only the $v_i \leftrightarrow M_i$ correspondence, so every depth-varying geometric property is preserved in the null by construction: it factors depth into each cell's significance and says nothing about the profile across depth.

Surrogate. At each layer, replace the fitted generators with random matrices matched to that layer's generator norm spectrum, ensemble anisotropy and common affine component (surrogate S1, whose null every R and p below quotes; a second family S2 matches the activation covariance instead and is stored beside it; a third, S3, is isotropic and matches only each generator's norm $\|M_i - C\|$, an unmatched baseline), with the algebra destroyed; recompute the profile statistic; 500 draws per layer, *seeded per layer*. The statistic R is the rising profile: mean $\text{bracket}_{\cos}$ over the three deepest layers minus mean over the three shallowest, tested one-sided (rising) against the surrogate's joint null.

Positive control. Plant an exact $\mathfrak{so}(3)$ at one layer of an otherwise surrogate stack, at each of nine layers in turn, at two corruption levels: the profile must spike at that layer and only there. Recovered $9/9$ at the floor of the null ($p = 1/4001$: the mean of 40 planted stacks against 4000 bootstrap means of 40 surrogate draws) in all seven carriers at both levels ($126/126$). The surrogate profile is centred at zero at every depth, as it must be (negating a surrogate negates its $\text{bracket}_{\cos}$, so every family is symmetric about zero; 189 cells: seven carriers × nine layers × three surrogate families; null mean within ± 0.005), and under it the peak lands in the final third in 26–46% of draws (mean 33%)—the uniform expectation for 3 of 9 layers.

Calibration. The false-positive rate on unplanted layers is 0.014–0.097 against 0.050 expected across the fourteen model×corruption combinations. The per-run criterion coded into the null ("at

most one of eight unplanted layers below $p < 0.05$"; `h2_depth_null.py`) is violated by chance 5.7% of the time per run ($1 - 0.95^8 - 8(0.05)(0.95^7)$), hence somewhere in nine runs $\sim 40\%$ of the time even for a perfect construction; it failed in 5 of 14 combinations and was not relaxed—a mis-calibrated criterion is reported as coded, with the calibration beside it.

Two construction defects caught by the controls. (i) A single RNG stream reused across layers gives every layer the same Gaussian draws, coupling the per-layer nulls and shrinking the profile statistic's variance—an anti-conservative null; detected by the positive control's false-positive rate (about three times the expected count on one model, in a pre-fix run that was not persisted), fixed to per-layer seeding; the shipped post-fix rates are 1–7 of 72 across the fourteen combinations. (ii) Rescaling the generators of the planted $\mathfrak{so}(3)$ individually to match an observed norm spectrum destroys the bracket relation it exists to carry ($[c_iM_i, c_jM_j] = c_ic_j[M_i, M_j]$); detected because the plant scored below zero at one layer in a pre-fix run; fixed to a single common scale for the plant, per-generator matching for the surrogate only.

Table 13: Depth-matched null, n=7. The rising test is one-sided by design and returns ≈ 1 for any falling profile; the mirror tail is computed post-hoc, labelled as such, and never converts a failed rising test into a passed one. Mistral's base arm reaches $R = +0.038$ ($p = 0.044$), 73% of its fine-tuned value.

carrier	family	blocks	peak	depth	R	p (rising)
Qwen2.5-1.5B	Qwen	28	L23	82%	+0.318	0.002
Qwen2.5-3B	Qwen	36	L30	83%	+0.146	0.002
Qwen2.5-7B	Qwen	28	L23	82%	+0.158	0.002
Mistral-7B-v0.3	Mistral	32	L23	72%	+0.053	0.056
Granite-3.1-8B	Granite	40	L20	50%	+0.009	0.379
OLMo-2-7B	OLMo-2	32	L16	50%	−0.061	1.000 (mirror 0.002, post-hoc)
OLMo-2-13B	OLMo-2	40	L16	40%	−0.063	0.986 (mirror 0.016, post-hoc)

Seed replicate: profile shape is not a model property at small $|R|$. Recomputed on a second training seed where one exists: Qwen2.5-1.5B holds ($+0.318 \to +0.356$, $z = +13.6 \to +14.6$, same peak layer); OLMo-2-7B flips ($-0.061 \to +0.071$, $z = -3.07 \to +3.03$), nominally significant in both directions from a seed change alone, with peak depth moving only 50% → 59%. The claim that OLMo-2 profiles fall is withdrawn. What survives is peak location (Qwen 82% at both seeds; OLMo mid-depth at both) and the magnitude contrast ($|R| \approx 0.32$–0.36 for Qwen, 5× OLMo-2's ≈ 0.06–0.07). Rule adopted: quote R's sign only with a seed replicate behind it, and treat $|R| \lesssim 0.1$ as unresolved regardless of z. OLMo-2-13B's R rests on a single seed and is read the same way.

Family consistency. Peak depth spans 40–83% across four families and varies by at most 19 points within one (OLMo-2: 40%, 50%, 59% across two sizes and two seeds). The within-family prediction for which the second OLMo-2 size was added (it peaks mid-depth) was confirmed. Both OLMo-2 models peak at absolute layer 16; at n=2 this is not distinguished from the depth-fraction reading and is not claimed—exactly as layer 23 was not claimed at n=3. Why families differ is not established: architecture, corpus, and instruction-tuning recipe are confounded across these families, and "family" is the best available predictor, not a mechanism. A practical corollary: the mid-depth band (40–50%) is the strongest region for three of seven carriers and weak for the three Qwen models—probe depth must be chosen per model, never by a shared default.

H The Causal Arm and Its Null Simulations

Protocol. Steer with the *fitted* generators at the reference carrier's strongest cell (Qwen2.5-1.5B, layer 23, d=8; refitted for this arm on 700 random words, a label-free fit whose prefixes may include the evaluated words), injecting αM_i then αM_j and the reverse at every token position, prompt and generated tokens alike, and score the composition task itself (which, unlike a judged trait, has a defined order-blind ceiling). Statistic: correlation of $\|[M_i, M_j]\|$ with the behavioral order-swap over all 66 generator pairs (the six same-axis pairs included, since their fitted brackets are never exactly zero; the identity and the simulation below use the 60 pairs with non-zero analytic bracket),

permutation p. Negative arms: the base model (no adapter), which cannot perform the task, and the abelian adapter, the matched negative whose true algebra has every bracket zero. Prediction, as stated in the experiment script (`h1_causal_bracket.py`): base null, carrier positive.

Table 14: The causal arm. Carrier and base are positive at comparable magnitude and p, while the base scores 0.000 exact-match. The abelian arm's channel is dead and its null is unreportable under the validity rule of Appendix C.6. Ceiling: in-sample order-blind majority on the 96 evaluated words (48 of length 2, 48 of length 3); 84 of their 90 bags hold a single word, so it is near 1 by construction.

arm	channel	efficacy α 0.1→0.8	unsteered exact	ceiling	corr	perm p
carrier (k=12 adapter)	live	$0.117 \to 0.629$	0.875	0.958	$+0.385$	0.0021
base (no adapter)	live	$0.082 \to 0.607$	**0.000**	0.958	$+0.373$	0.0020
abelian adapter	dead	$0.000 \to 0.003$	1.000	1.000	-0.082	0.51

The cause is an identity. By Eq. 4 the activation-level order-swap is bracket-proportional by construction. Measured on the dumped L23/d8 generators (the screening fit of the cell; the causal arm refits the same cell on its own words) with no model in the loop (operator-level swap $\|E_iE_j - E_jE_i\|_F$ at the calibrated $\alpha = 0.8$, deterministic): $r = +0.961$ (carrier generators), $+0.888$ (base generators), permutation $p = 10^{-4}$; the state-level form, averaged over 256 random states, gives $+0.963 \pm 0.006$ and $+0.868 \pm 0.020$ over 20 seeds, and both forms tend to 1 as $\alpha \to 0$, the BCH limit. Any behavioral readout monotone in perturbation size inherits it. No causal verdict is reported; H1 remains correlational.

The proposed repair, simulated before adoption. Hold the activation-level perturbation matched across pairs and vary only direction. Simulated on CPU with planted generators of known bracket, states drawn from the layer's measured PC covariance, and behavioral readouts that never compute a bracket, plus one bracket-using readout kept as a power control (Pearson r against $\|[M_i, M_j]\|$, 60 pairs, carrier's fitted generators, $\alpha = 0.2$, 5 seeds):

Table 15: Null simulation of the repaired causal design. Matching α per pair in advance (the implementable form) does nothing. Matching the realized displacement nulls only magnitude-monotone readouts; any direction-sensitive readout—which an LLM behavioral readout is—retains $r \approx 0.58$ with no bracket used anywhere. Matching the angle as well drives the power control to $+0.069$ (five-seed mean; not significant in any seed, permutation p from 0.15 to 0.79): the angle is the channel the bracket acts through, and removing it leaves no experiment.

Readout (no bracket used)	Unmatched	Matched α (a priori)	Matched displacement	Matched displ. & angle
Magnitude-monotone (identity)	$+0.627 \pm 0.018$	$+0.629 \pm 0.010$	-0.015 ± 0.142	-0.015 ± 0.142
Magnitude-monotone (sqrt)	$+0.567 \pm 0.047$	$+0.563 \pm 0.031$	-0.015 ± 0.142	-0.015 ± 0.142
Linear functional of displ.	$+0.568 \pm 0.047$	$+0.617 \pm 0.035$	$+\mathbf{0.582 \pm 0.061}$	-0.006 ± 0.103
Saturating direction-sensitive	$+0.524 \pm 0.118$	$+0.527 \pm 0.105$	$+\mathbf{0.586 \pm 0.084}$	$+0.095 \pm 0.217$
Bracket-using (power control)	$+0.728 \pm 0.012$	$+0.702 \pm 0.013$	$+0.894 \pm 0.012$	$+0.069 \pm 0.085$

An open puzzle. The original steering experiment on Qwen2.5-7B-Instruct without task fine-tuning (contrastive trait generators, calibrated $\alpha = 0.8$, live channel, Appendix C.6) found corr $= -0.206$ ($p = 0.30$) at the calibrated α and a negative correlation at every α tested ($-0.372/-0.064/-0.107/-0.206$), with an α-scaling exponent of $+1.43$ (BCH predicts ≈ 2). Given that the activation channel is bracket-proportional at $\approx +0.9$ by construction, a model reading its steered activations monotonically would have shown a positive correlation for free; that it did not is stranger than it first appeared, and we leave it open.

I THE FRAME-REMOVAL IMPOSSIBILITY

Design. Set 2 is set 1 re-framed: a rotation $Q_0 \in SO(3)$ (angle constrained to 60–120° so the frames are neither near-identical nor near-antipodal; a draw is also rejected when any set-2 generator axis lies

within 20° of a set-1 axis), an axis permutation π, and a disjoint token-name pool sharing no substring with set 1. Direction-level task quantities are exactly preserved because a rotation commutes with the cross product:

quantity	set 1	set 2	
bracket lands on an existing generator axis	60%	60%	exact
distinct bracket directions / magnitudes	18 / 3	18 / 3	exact
zero-bracket pairs, mapped 1:1 under π	6 of 66	6 of 66	exact

The gate. A CPU acceptance gate, run before any training, requires set 2 to match set 1 on every label-derived statistic within tolerance, the binding one being the sign-blind ceiling on the commutator split (tolerance 0.010).

draws	in-band (60–120°)	passing	best $\lvert\Delta\text{ceiling}\rvert$	tolerance
14,646	5,000	**0**	0.0379	0.010

Mean Δ(ceiling) $= -0.108 \pm 0.019$, independent of rotation angle (-0.092 at 10–20°, -0.107 at 170–180°); the angle band is not what breaks it.

Mechanism, measured. `bin_vec` bins Cartesian components against a fixed axis-aligned grid, so nothing downstream of it is rotation-equivariant. Set 1's axes x, y, z, xy, xz, yz are exactly the symmetry directions of that grid; their commutator labels collide inside each order-blind bag, and the collisions hold the ceiling up:

	distinct labels	per bag (mean)	sign-blind ceiling
set 1	31	3.71	0.3864
set 2, best in-band draw	39	4.19	0.3485
set 2, typical	37–39	4.95–5.05	0.265–0.288

Removing the shared frame requires supervising the two models toward different label coordinates; matching difficulty requires the label statistics to be unchanged. These are the same requirement with opposite signs. The obvious repair—bin set 2 in the rotated frame—restores every statistic exactly and makes set 2 a pure relabeling, which is the thing the ablation exists to avoid. The two repairs that remain are decisions, not resamples: (i) re-derive the tolerances for a decoder-free-only protocol—the order-blind ceiling is a *behavioral* baseline, and the frame-free readout has no behavioral arm; or (ii) replace the labeling map with one whose difficulty is frame-independent by construction.

J Robustness Replicates

Decoder-free bracket$_{\cos}$ on Qwen2.5-1.5B, one variable changed per row (plus the OLMo-2-7B seed-1 replicate that supplies ceiling B), all else at the reference operating point; 500 permutations; controls in every run (`exact` $= 1.000$, `random`/`shuffled` within ± 0.15). Pre-specified success criterion: same verdict direction, comparable magnitude.

Table 16: All replicates hold ($p = 0.002$ at every best cell). *Margin and p are three-seed means from the k-sweep (Table 10); other columns are seed 0. †Run to estimate Ceiling B. Degenerate-pair partitions are exact in every run (40+5 at k=10, 60+6 at k=12, 112+8 at k=16). Seed 1 is the tightest replicate. k=10 is the weakest at 45% of the reference magnitude—the same direction the behavioral sweep shows (+0.067 vs. +0.139), two independent readouts agreeing that k=10 sits just past the transition. The peak layer moves to L27 at k=10 and k=16 (at k=16, L23's +0.397 is a statistical tie with L27's +0.407): peak presence is more stable than peak location.

Variant	Sig. cells /27	Best d=8 cell	Peak	Behavioral margin	Own sign-blind ceiling	Margin p
k=12, seed 0 (ref.)	22	+0.487	L23	+0.139	0.386	0.016*
k=12, seed 1	18	+0.463	L23	—	—	—
k=10, seed 0	14	+0.219	L27	+0.067	0.433	0.03*
k=16, seed 0	17	+0.407	L27	+0.268	0.325	0.0063*
θ=0.45, seed 0	22	+0.349	L23	+0.144	0.477	—
OLMo-2-7B, seed 1†	25	+0.413	L19	+0.023	0.386	—

K Full Result Tables

K.1 Screening, all arms

Table 17: Full screening statistics (27 cells = 9 layers $\times$ $d \in \{3, 8, 16\}$; chance 1.35 significant cells). All FT best cells are at $p = 0.002$ except Llama-3-8B, whose sub-dimension-8 row is empty at all nine layers. The Llama-3-8B base arm was first measured on a 4-layer grid (1/12) and re-dumped to the full grid (2/27, Table 18); its fine-tuned arm was re-dumped to the full grid and the 18 shared cells reproduce to 3×10^{-16}. Significant cells by sub-dimension ($d3/d8/d16$): Qwen1.5B 8/7/7, Qwen3B 7/6/4, Qwen7B 4/8/7, Mistral 1/5/2, OLMo-7B 8/9/9, OLMo-13B 7/7/5, Granite 4/6/7, Llama 4/0/2.

model	arm	cells	sig. (p<.05)	mean	min	max	best d=8 cell
Qwen2.5-1.5B	FT	27	22	+0.178	−0.029	+0.487	+0.487, L23 (82%)
Qwen2.5-1.5B	base	27	2	+0.004	−0.059	+0.057	+0.057 (p=.034), L14
Qwen2.5-3B	FT	27	17	+0.075	−0.033	+0.381	+0.223, L30 (83%)
Qwen2.5-3B	base	27	0	−0.015	−0.087	+0.007	+0.007 (p=.381), L35
Qwen2.5-7B	FT	27	19	+0.097	−0.023	+0.264	+0.264, L23 (82%)
Qwen2.5-7B	base	27	1	−0.013	−0.099	+0.057	+0.003 (p=.449), L14
Mistral-7B-v0.3	FT	27	8	+0.035	−0.080	+0.160	+0.160, L23 (72%)
Mistral-7B-v0.3	base	27	0	+0.003	−0.080	+0.046	+0.020 (p=.289), L16
OLMo-2-7B	FT	27	26	+0.287	−0.032	+0.572	+0.514, L16 (50%)
OLMo-2-7B	base	27	1	−0.005	−0.056	+0.086	+0.031 (p=.188), L13
OLMo-2-13B	FT	27	19	+0.137	−0.014	+0.532	+0.391, L16 (40%)
OLMo-2-13B	base	27	2	+0.007	−0.035	+0.082	+0.045 (p=.018), L16
Granite-3.1-8B	FT	27	17	+0.101	−0.027	+0.410	+0.394, L20 (50%)
Granite-3.1-8B	base	27	2	+0.007	−0.028	+0.052	+0.045 (p=.100), L36
Llama-3-8B	FT	27	6	+0.038	−0.044	+0.277	+0.041 (p=.122), L31
Llama-3-8B	base	27	2	+0.002	−0.043	+0.061	+0.059 (p=.068), L5

Table 18: True pretrained checkpoints, same 27-cell grid and three controls (all passing). The bracket is absent before and after instruction tuning, and the order-dependent residual is no more than 0.03 above the last-token bound in both. †Llama-3-8B has no Instruct arm: its screening arm is this same pretrained checkpoint, first measured on an earlier 12-cell grid (1/12), which the 27-cell verdict replaces. Its base grid starts at layer 5 (the fine-tuned grid at layer 4); the other eight layers match.

Pretrained checkpoint	Sig. cells /27	Best d=8 cell	Residual R^2 (raw)	Instruct arm sig. /27
Qwen2.5-1.5B	1	+0.051 (p=.042), L14	0.520 (0.941)	2
Qwen2.5-7B	1	+0.048 (p=.154), L4	0.566 (0.949)	1
OLMo-2-1124-7B	2	+0.081 (p=.042), L9	0.515 (0.929)	1
Granite-3.1-8B	1	+0.028 (p=.206), L36	0.489 (0.919)	2
Llama-3-8B	2	+0.059 (p=.068), L5	0.407 (0.890)	1/12†

K.2 Depth profiles at d=8, all carriers

Table 19: Per-layer bracket$_{\cos}$ at sub-dimension 8 for all seven carriers (block count in parentheses; the 28-block Qwen2.5-7B and 32-block OLMo-2-7B share the depth rows of their size-mates above them; Granite shares OLMo-2-13B's).

Qwen2.5-1.5B (28)	L4	L8	L11	L14	L17	L20	L23	L25	L27
depth	14%	29%	39%	50%	61%	71%	82%	89%	96%
bracket$_{\cos}$	+.068	+.066	+.013	+.058	+.137	+.188	+.**487**	+.355	+.261
p	.006	.012	.263	.056	.002	.002	.002	.002	.002
Qwen2.5-3B (36)	L5	L10	L14	L18	L22	L26	L30	L32	L35
depth	14%	28%	39%	50%	61%	72%	83%	89%	97%
bracket$_{\cos}$	+.003	+.073	−.030	+.069	+.039	+.031	+.**223**	+.183	+.077
p	.463	.006	.916	.016	.034	.060	.002	.002	.002
Qwen2.5-7B (28)	L4	L8	L11	L14	L17	L20	L23	L25	L27
bracket$_{\cos}$	−.023	+.131	+.107	+.109	+.205	+.235	+.**264**	+.253	+.173
p	.892	.002	.002	.002	.002	.002	.002	.002	.002
Mistral-7B-v0.3 (32)	L5	L9	L13	L16	L19	L23	L26	L29	L31
depth	16%	28%	41%	50%	59%	72%	81%	91%	97%
bracket$_{\cos}$	+.058	+.006	+.045	+.028	+.062	+.**160**	+.126	+.101	+.040
p	.036	.391	.082	.072	.008	.002	.002	.002	.052
OLMo-2-7B (32)	L5	L9	L13	L16	L19	L23	L26	L29	L31
bracket$_{\cos}$	+.040	+.429	+.456	+.**514**	+.508	+.304	+.238	+.263	+.243
p	.006	.002	.002	.002	.002	.002	.002	.002	.002
OLMo-2-13B (40)	L6	L11	L16	L20	L24	L29	L33	L36	L39
depth	15%	28%	40%	50%	60%	72%	82%	90%	98%
bracket$_{\cos}$	+.016	+.096	+.**391**	+.167	+.184	+.099	+.130	+.152	+.031
p	.146	.002	.002	.002	.002	.002	.002	.002	.174
Granite-3.1-8B (40)	L6	L11	L16	L20	L24	L29	L33	L36	L39
bracket$_{\cos}$	+.008	+.018	+.175	+.**394**	+.220	+.135	+.177	+.052	+.001
p	.325	.146	.002	.002	.002	.002	.002	.038	.481

K.3 Fit diagnostics and base-model decode

Table 20: Fit-quality diagnostics at each arm's best sub-dimension-8 cell. Decode R^2 here comes from a 70/30 word-level split that does not exclude shared prefixes, unlike the prefix-disjoint probe of Section 3.1. Φ-fit R^2 is measured against the zero model, so it includes the common component C, which carries 0.17–0.77 of $\|M\|^2$ across the rows; against the mean the values are 0.15–0.81. These are reported to show the fits are healthy, never as evidence: the abelian control scores well on Φ-fit and decode R^2 with no bracket. Base decode R^2 spans 0.786 (Qwen2.5-3B) to 0.972 (OLMo-2-7B); Llama-3-8B decodes at 0.989 after fine-tuning and carries no bracket. Predicted magnitude and relative error track the verdict (carriers 0.14–0.38 and 1.05–3.0; non-carriers ≈ 0.07–0.14 and 4.0–9.9) because they derive from the same fit as the cosine—corroborating, not independent. OLMo-2-7B's relative error of 1.052 is the lowest recorded: for the first time the error is roughly the size of the predicted bracket rather than larger.

model	arm	Φ-fit R^2	predicted mag.	bracket rel. err.	decode R^2
Qwen2.5-1.5B	FT	0.910	0.270	1.444	0.995
Qwen2.5-1.5B	base	0.802	0.086	5.436	0.941
Qwen2.5-3B	FT	0.915	0.240	2.037	0.994
Qwen2.5-3B	base	0.572	0.072	7.983	0.786
Qwen2.5-7B	FT	0.783	0.192	1.979	0.994
Qwen2.5-7B	base	0.784	0.107	5.593	0.936
Mistral-7B-v0.3	FT	0.794	0.144	2.996	0.992
Mistral-7B-v0.3	base	0.670	0.086	7.258	0.929
OLMo-2-7B	FT	0.957	0.378	1.052	0.997
OLMo-2-7B	base	0.846	0.129	4.641	**0.972**
OLMo-2-13B	FT	0.916	0.199	1.794	0.996
OLMo-2-13B	base	0.743	0.138	5.102	0.963
Granite-3.1-8B	FT	0.723	0.182	1.765	0.997
Granite-3.1-8B	base	0.857	0.109	4.018	0.935
Llama-3-8B	FT	0.515	0.071	9.911	**0.989**
Llama-3-8B	base	—	—	—	0.917

K.4 Behavior does not track representation

Table 21: Behavioral commutator margin vs. representational bracket at k=12 (screening run, seed 0), eight models, against the sign-blind ceiling ($0.386 = 51/132$) and the sign-aware ceiling ($0.545 = 72/132$; Appendix B). Mistral sits two answers above the sign-blind ceiling (53/132) while carrying a significant bracket; Llama-3-8B has a positive margin and no interpretable bracket; OLMo-2-7B has the strongest bracket and a middling margin; Granite sits five answers above the sign-blind ceiling with the third-strongest bracket. Qwen2.5-1.5B (76/132) and Qwen2.5-7B (73/132) are above the sign-aware ceiling (72/132), by four and by one answer; the other six are at or below it. Spearman rank correlation between sign-blind margin and $\text{bracket}_{\cos}$ across the eight: +0.381, $p = 0.352$ (asymptotic; exact permutation $p = 0.360$). Within the Qwen family alone the two co-vary (1.5B > 7B > 3B on both), a further reason to read orderings as family-specific.

model	commutator acc.	margin vs. sign-blind	margin vs. sign-aware	best d=8 $\text{bracket}_{\cos}$
Qwen2.5-1.5B	0.576	+0.189	+0.030	+0.487
Qwen2.5-7B	0.553	+0.167	+0.008	+0.264
Qwen2.5-3B	0.492	+0.106	−0.053	+0.223
OLMo-2-13B	0.508	+0.121	−0.038	+0.391
OLMo-2-7B	0.477	+0.091	−0.068	+**0.514**
Llama-3-8B	0.462	+0.076	−0.083	+0.041 (n.s.)
Granite-3.1-8B	0.424	+0.038	−0.121	+0.394
Mistral-7B-v0.3	0.402	+0.015	−0.144	+0.160

K.5 Results not in the main text

The steering-vector and Procrustes scripts behind this section are retired; their result files are released as the record.

Steering-vector experiments on base models. Before the fine-tuning results, we tested whether contrastive steering generators in a base model (Qwen2.5-7B-Instruct) are organized by a bracket: injecting two generators in both orders and correlating the behavioral difference with $\|[A,C]\|$. Null on a live channel (corr $= -0.206$, $p = 0.303$ at calibrated $\alpha = 0.8$; $-0.372/-0.064/-0.107/-0.206$ at $\alpha = 0.1/0.2/0.4/0.8$; layer scan over layers 4–24, every permutation $p \geq 0.134$). A sequential-additive control correlates at -0.222, so the group protocol is not distinguishable from a two-phase additive one. This is the behavioral face of the Absent pillar: frozen steering vectors in a pretrained model are not a Lie algebra, and the bracket one computes from them is arithmetic the experimenter performed.

Procrustes transfer of steering generators. Between Qwen2.5-1.5B and Llama-3-8B, aligned/naive/random direction transfer reaches 0.821–0.895 / -0.008–0.143 / -0.01–0.04 across layers 4–24 (peak at layer 16), with bracket transfer at roughly half that (0.244–0.452, vs. ≈ 0.02 naive). On concepts excluded from the fit, cosine to the target model's own generator is 0.802 ± 0.064 (naive 0.120, random -0.011), closing the circularity objection. The behavioral round-trip is real but not specific: at a calibrated working point the aligned port matches the target model's own behavioral effect (ratio 1.01) but so does the naive port, and scoring every concept shows the aligned port has zero concept specificity (-0.007 vs. the own-generator ceiling's $+0.047$, paired $p = 0.010$, top-1 $0/8$) despite cosine 0.749. A direction interpolated from random toward the target's own generator already recovers near-full specificity by cosine ≈ 0.72 ($+0.043$ vs. the own-generator $+0.047$), below the aligned port's 0.749, so the Procrustes residual is structured, not noise. Direction transfers; behavior does not; and both results are untouched by the entailment analysis of Section 5, which concerns fine-tuned pairs. We note them here because they are the first appearance of a pattern the paper then finds everywhere: alignment statistics are not behavioral equivalence.

L Generator Names and Model Scale

Every base arm in the main text names its generators with the strings of Section 3.1 (`Rx+`, `Rxy-`, ...; 2–4 BPE pieces), which mean nothing to a pretrained model, and the Qwen2.5 base arms stop at 7B. The absence of the bracket in base models (Section 4.1) could therefore reflect names the model does not recognise, or a scale too small to carry the algebra. We test both on base checkpoints, with no training of any kind.

Design. *Semantic* names state each generator in words, from `rotate +0.6 rad about x` to `rotate -0.6 rad about the y+z diagonal`; the abelian control's commuting generators, all about z, become `rotate +0.6 rad about z` to `rotate -3.6 rad about z`. Names are joined by "`; `" instead of a space. Nothing else changes: rotation vectors, generator indices, the word sample, the readout at => and the labels are those of the *token* names, the paper's default. A four-generator prompt grows from 14 to 47 tokens in the Qwen2.5-32B tokenizer:

token: `Compose: Rxz+ Rx+ Rxy+ Rxz- =>`

semantic:
```
Compose: rotate +0.6 rad about the x+z diagonal;
    rotate +0.6 rad about x; rotate +0.6 rad about the x+y
diagonal;
    rotate -0.6 rad about the x+z diagonal =>
```

Each arm runs the identical pipeline: the base-arm generator dump (2000 words) on nine proportional-depth layers $\times$ $d \in \{3, 8, 16\}$, $\text{bracket}_{\cos}$ with the three controls and a 500-fold permutation test per cell, and the residual probe of Appendix C.8 (3000 words, best of four layers); each arm run here has its abelian control. Semantic names run on Qwen2.5-Instruct at 1.5B, 3B, 7B and 32B, on OLMo-2-7B-Instruct and on the pretrained Qwen2.5-1.5B. Token names run on Qwen2.5-32B-Instruct; its 1.5B–7B counterparts are the published base arms (Table 17). Qwen2.5-32B has 64 blocks. Rounding the 28-block grid in proportion, which reproduces the grids of the 32-, 36- and 40-block models, gives

layers $9, 18, 25, 32, 39, 46, 53, 57, 62$ (14–97% depth); the residual probe uses layers $19, 32, 44, 57$. A memory probe fixed the dtype by a pre-specified rule (float16, the dtype of every published arm, if the model fits with every hidden state finite; bfloat16 otherwise). In float16 the model fits one `71gb` MIG slice with no module offloaded (61.0 GiB of weights, peak 62.6 GiB reserved of 69.8 GiB), and every hidden state at the nine layers is finite on 31 prompts per naming.

Pre-specified rules. Gates, verdict rules and outcome sentences were written before any number of the corresponding run existed (`docs/PRESPEC.md`, sections "Semantic generator names" and "Qwen2.5 scale line to 32B"). An arm is void if, in any cell, `exact` departs from 1 by 10^{-6} or more, or $|$`random`$|$ or $|$`shuffled`$|$ exceeds 0.15; configuration, pair counts (60 scored and 6 degenerate; abelian 0 and 66), permutation-null sanity and finiteness are gated too. The verdict applies the carrier definition of Section 3.2 as coded for Table 18: *absent* if at most 3 of the 27 cells are significant ($p < 0.05$; 1.35 expected by chance), *carrier* if at least 8 are and the best d=8 cell has $p < 0.05$, *ambiguous* otherwise. The first of the two pre-specifications also required no significant d=8 cell for *absent*; the second reports that stricter rule beside the verdict. The residual probe counts as decoding the element beyond the input only above 0.575, the last-token bound (0.545) plus 0.03. As a sensitivity check that never changes a verdict, Φ-fit and update-fit R^2 at the best d=8 cell are compared with the band of the eight published base arms at theirs (Φ-fit 0.572–0.862, update-fit 0.671–0.941; Table 20, with Llama-3-8B from its 27-cell re-dump).

Results. The bracket is absent at every scale under both namings (Table 22). Under token names Qwen2.5-32B has 0 of 27 significant cells (best d=8 cell $+0.036$, p=.102), after 2, 0 and 1 at 1.5B, 3B and 7B. Under semantic names every Qwen2.5 size has 0 of 27, with best d=8 cells between $+0.035$ and $+0.068$ ($p \geq .078$); OLMo-2-7B has 2 (both at d=16) and the pretrained Qwen2.5-1.5B 0. All six semantic arms and the 32B token arm are absent under the stricter rule too; the published Qwen2.5-1.5B token arm, with two significant d=8 cells (L14, L20), is absent under the coded rule only. No arm is void: `exact` returned 1.000000 in every cell of all ten non-abelian arms, $|$`random`$| \leq 0.066$ throughout, and $|$`shuffled`$| \leq 0.090$ on the token arms and ≤ 0.146 on the semantic arms, the largest on Qwen2.5-32B and within the 0.15 bound. (Control $|z|$ against each cell's own permutation null, reported but not gated, peaks at 4.19, `random` in one 32B semantic cell.) The seven abelian arms (six semantic, the 32B token arm) have $\text{bracket}_{\cos}$ undefined in all 27 cells, as the construction requires, and a mean bracket magnitude (0.323–0.381) comparable to that of their non-abelian arms (0.283–0.384): the fitting floor of Appendix C.1, which is why magnitude is never quoted as evidence. Their residual target is identically zero and is not scored.

Residual decode. Under token names the order-dependent residual sits at the last-token bound at every scale (0.527, 0.571, 0.553 and 0.554 from 1.5B to 32B; none above 0.575). Under semantic names it lies far below the bound: 0.098–0.152 on the Qwen2.5 line, 0.191 on OLMo-2-7B and 0.119 on the pretrained 1.5B, with raw decode lower as well (0.674–0.825 across the six semantic arms, against 0.941–0.949 under token names). The pre-specified rule separates only above from not above the threshold, and no arm is above it; the distance below the bound was not pre-specified, and we do not interpret it. One possibility, untested, is that with names spanning several tokens less of the last generator's identity is linearly readable at the readout token.

Fit sensitivity. Every semantic arm's Φ-fit R^2 at its best d=8 cell (0.278–0.467) lies below the published base-arm band, whose minimum is 0.572, while update-fit R^2 at the same cells (0.684–0.839) stays inside its band. The 32B token arm is inside both (0.800 and 0.754). Under semantic names the transition regressions fit within the published range, but the fitted generators are further from a linear image of $\mathfrak{so}(3)$ than in any published base arm. An absent verdict there rests on a weaker fit: the semantic-name null is weaker evidence than the token-name null.

What this licenses. The base-arm absence of Section 4.1 is not an artifact of meaningless names: with each rotation named in words, no base checkpoint of either family tested carries the bracket. Nor is it an artifact of scale within Qwen2.5 up to 32B. It does not license a scale claim outside that one family, and the semantic-name null, which rests on weaker fits, is less informative than the token-name null. One natural-language naming was tested, and no fine-tuned arm was rerun under it.

Table 22: Base checkpoints under token and semantic generator names, no training. Same 27-cell grid, three controls (all passing) and residual probe as Tables 17 and 9. The first eight rows are the Qwen2.5-Instruct scale line; its 1.5B, 3B and 7B token rows are the published base arms. Φ-fit R^2 is taken at the best d=8 cell (published base-arm band 0.572–0.862; every semantic row lies below it). Residual R^2 is the best of four layers, layer given (last-token bound 0.545, threshold 0.575). Verdict by the coded rule (absent: at most 3 of 27 cells significant). OLMo-2-7B is the Instruct checkpoint; †the true pretrained checkpoint, whose token-name row is in Table 18.

model	names	sig. cells /27	best d=8 cell	Φ-fit R^2	residual R^2 (raw)	verdict
Qwen2.5-1.5B	token	2	+0.057 (p=.034), L14	0.802	0.527 (0.941), L8	absent
Qwen2.5-1.5B	semantic	0	+0.039 (p=.150), L17	0.457	0.100 (0.680), L14	absent
Qwen2.5-3B	token	0	+0.007 (p=.381), L35	0.572	0.571 (0.949), L10	absent
Qwen2.5-3B	semantic	0	+0.035 (p=.202), L14	0.404	0.138 (0.674), L18	absent
Qwen2.5-7B	token	1	+0.003 (p=.449), L14	0.784	0.553 (0.946), L8	absent
Qwen2.5-7B	semantic	0	+0.047 (p=.134), L17	0.409	0.098 (0.680), L14	absent
Qwen2.5-32B	token	0	+0.036 (p=.102), L32	0.800	0.554 (0.948), L19	absent
Qwen2.5-32B	semantic	0	+0.068 (p=.078), L32	0.467	0.152 (0.825), L19	absent
OLMo-2-7B	semantic	2	+0.073 (p=.132), L26	0.278	0.191 (0.683), L16	absent
Qwen2.5-1.5B†	semantic	0	+0.040 (p=.184), L14	0.443	0.119 (0.682), L14	absent

M Training Dynamics

Table 23: Training dynamics, three seeds per model: median [min–max] over seeds of the step completing 10/50/90% of each rise, and the logistic-fit width in $\log_{10}$ steps. Per-seed verdicts: G = gradual (> 0.5), I = intermediate, A = abrupt (< 0.25). Element = order-dependent residual R^2 (probe refit at every checkpoint on 2000 words, read at the layer with the highest final residual R^2: Qwen L23/L23/L17, OLMo L9/L16/L13); algebra = best-cell $\text{bracket}_{\cos}$ (Qwen L23 d=8/8/3, OLMo L26/L26/L19 d=3); behavior = commutator exact match. Element precedes algebra at every layer for OLMo-2-7B; for Qwen2.5-1.5B from L14 deeper, with two seeds tied at L4–L11. Qwen seed 2's layer is a near-tie (L17 0.845 vs. L23 0.844; t_{50} 84 vs. 38) and does not move the median.

Quantity	t_{10}	t_{50}	t_{90}	Width	Verdicts
Qwen2.5-1.5B					
element	1 [1–3]	30 [18–84]	241 [186–314]	2.75 [2.22–2.84]	3G
algebra	110 [84–186]	241 [241–314]	689 [530–689]	0.54 [0.49–0.86]	2G 1I
behavior	186 [143–186]	408 [408–530]	1514	0.96 [0.83–1.12]	3G
OLMo-2-7B					
element	65 [65–84]	110 [110–143]	241 [186–314]	0.49 [0.49–0.54]	1G 2I
algebra	110 [110–143]	241 [186–314]	530 [314–689]	0.65 [0.37–0.83]	2G 1I
behavior	110 [110–143]	530 [408–689]	1514 [896–1514]	0.89 [0.86–1.01]	3G

N Extended Related Work

Structured representations in language models. Sequence models trained on structured domains develop internal representations of that structure: board states (Li et al., 2022; Nanda et al., 2023b), linearly encoded features (Park et al., 2024; Gurnee & Tegmark, 2024), Fourier circuits for modular arithmetic (Nanda et al., 2023a), group-theoretic circuits recurring across seeds (Chughtai et al., 2023; Stander et al., 2023; Wu et al., 2025), and representation-theoretic mechanisms beyond groups (Chen et al., 2026). Pretrained LLMs moreover carry multi-dimensional structure that is causally used: circular features for modular day/month arithmetic (Engels et al., 2025) and helical number representations (Kantamneni & Tegmark, 2025). Closest to us in spirit, Lee et al. (2026) shows a bag of linear directions misses *relational* structure in Othello, recovering a tensor-product binding of square and color embeddings; and Marchetti et al. (2026) reverse-engineers the Fourier mechanism by which small transformers trained from scratch compose group elements. These works establish that elements, bindings, and composition mechanisms are represented. We test something else: the

operation table itself, which is second-order structure with analytic ground truth. In pretrained LLMs it is absent even where element structure, including causally used abelian structure, is present.

State tracking and group composition. Composing non-commutative elements is the canonical state-tracking problem. Constant-depth transformers cannot simulate non-solvable groups such as S_5 step by step (Liu et al., 2022; Merrill et al., 2024), a bound Lyu et al. (2026) carry to $SO(3)$ for frozen vision embeddings. What transformers learn instead are shallow shortcuts, and solvable structure, abelian structure in particular, admits the cheapest ones (Liu et al., 2022), which is why an abelian arm is the right disqualifier for a bracket metric. Nearest to our setup, Li et al. (2025a) fine-tune pretrained LMs on permutation composition and probe the running product, finding two distinct tracking mechanisms; Kvinge et al. (2026) ask whether narrow models trained on a group operation carry properties the task never demanded, such as subgroup membership. Those studies ask whether, and how, the *running element* is tracked. We ask whether the *algebra that composes elements* is represented, on a continuous group where the answer has analytic ground truth, and find that the two come apart: fine-tuned Llama-3-8B represents the composed element ($R^2 = 0.86$) and carries no interpretable bracket, and across the eight models bracket strength does not predict behavioral order-sensitivity ($\rho = 0.38$, Section 4.4).

Probing methodology: outputs versus laws, representation versus use. The probing literature warns that a successful probe need not indicate a functional representation (Belinkov, 2022; Ravichander et al., 2021), motivating control tasks (Hewitt & Liang, 2019), amnesic probing (Elazar et al., 2021), and falsifiability standards (Leavitt & Morcos, 2020). Our element/algebra distinction is the representational form of a behavioral one: a sequence model can predict a system's outputs near-perfectly while recovering no coherent world model (Vafa et al., 2024), and an orbit predictor with $R^2 > 0.999$ implies a different force law for every sample rather than Newton's (Vafa et al., 2025). Fine-tuned Llama-3-8B is that case in our setting. Recent LLM studies report the dissociation of Section 4.4 in natural domains: cyclic geometry present while the computation runs through base-10 addition (Feucht et al., 2026), and correctness information decodable and transferable across models yet causally inert (Usama & Chang, 2026). We extend this program from features to *algebraic relations*: our abelian control eliminates three of four candidate structure metrics, and every retained readout carries an analytic positive control.

What fine-tuning changes. Whether fine-tuning creates mechanisms or re-weights existing ones is actively contested. Entity-tracking circuits survive fine-tuning intact (Prakash et al., 2024), follow-up work frames gains as *circuit amplification*, meaning strengthened or replicated components rather than novel mechanisms (Chhabra et al., 2025), and Jain et al. (2024) reports modulation of pre-existing capabilities on procedural tasks. But circuit analysis on tasks where fine-tuning yields substantial gains finds significant edge changes and newly emerging circuits (Wang et al., 2025). In weight space, LoRA learns less than full fine-tuning (Biderman et al., 2024) yet adds "intruder" singular directions nearly orthogonal to the pretrained spectrum (Shuttleworth et al., 2026), and model-diffing recovers fine-tuning-specific latents only after correcting its own sparsity artifacts (Minder et al., 2026). Our Trainable pillar contributes a case with ground truth on the "creates" side: the bracket is measurably absent from eight base models and installed by a small LoRA fine-tune in seven of them across four families; in the two families we sweep over task size it appears one step before the size at which it can matter behaviorally. At least for structure the pretraining distribution never demanded, fine-tuning creates, not merely amplifies.

Cross-model similarity and transfer. Representation comparison is well studied: SVCCA (Raghu et al., 2017), CKA (Kornblith et al., 2019), stitching (Bansal et al., 2021), relative representations (Moschella et al., 2022), and convergence claims (Huh et al., 2024). Its pitfalls are increasingly well studied too. Similarity indices admit spuriously high values even for random networks unless deconfounded (Cui et al., 2022), conclusions can invert under a change of measure (Soni et al., 2024; Ding et al., 2021), high similarity does not imply functional equivalence (Klabunde et al., 2025), and decoding-level alignment need not reflect encoding-level alignment (Bertram et al., 2026). Shared data and shared objectives each raise cross-model similarity on their own (Li et al., 2025b), and convergence metrics recalibrated against scale-matched baselines largely lose their global signal (Gröger et al., 2026). More broadly, probes, sparse autoencoders, and circuit discovery all return plausible explanations on randomly initialised networks (Méloux et al., 2025), and probe-accuracy

depth profiles are far from flat under level-matched placebo baselines (Braun, 2026). Interpretability statistics need nulls that price in what is already entailed. Our entailment results give this critique a provable instance in the algebraic case: agreement statistics inflated by per-model fidelity under equivariance by a computable factor, and a frame-removal ablation shown not instantiable. Matched-noise nulls calibrated to each model's own fidelity are our proposed replacement for permutation nulls in this setting.

Interventions and causal claims. Activation steering (Turner et al., 2024; Zou et al., 2023; Li et al., 2023) and causal intervention methods (Meng et al., 2022; Geiger et al., 2021) ground representational claims in behavior, with growing attention to pitfalls: subspace patching can succeed by activating dormant pathways (Makelov et al., 2024), metric choice changes verdicts (Zhang & Nanda, 2024), and norm-matched orthogonal perturbations reproduce most of a steering vector's behavioral effect (Venkatesh & Kurapath, 2026). We add a cautionary instance of a different kind: an intervention whose signature correlation is a matrix identity, reproduced at full strength with no model in the loop. Simulating a proposed intervention's null before running it is the inexpensive discipline this failure suggests.

Symmetry and Lie structure in neural networks. Equivariant architectures build group structure in by construction (Cohen & Welling, 2016), and homomorphism-constrained autoencoders train the composition law into a latent space (Keurti et al., 2023). Symmetry-discovery methods recover Lie generators from data (Moskalev et al., 2022) and identify the algebra they span by the closure residual of their commutators (Forestano et al., 2023), which is the quantity $\text{bracket}_{\cos}$ scores, computed there on generators fit to be symmetries of a dataset rather than read from a model's own state updates. Nearer our readout, Gruver et al. (2022) measure emergent rather than built-in equivariance in pretrained vision models with Lie derivatives, and Miyato et al. (2022) fit linear transition operators to latent sequences and block-diagonalise them into irreducibles. Dehdarirad et al. (2026) use the finite Baker–Campbell–Hausdorff deviation of learned latent generators as a measure of non-commutativity. Section 5 is a caution for any such readout, ours included, on how readily a behavioral correlate of that deviation is entailed by the identity itself. We ask the converse of the constructive line for generic transformers: whether a model trained only on composition *behavior* acquires the algebra internally, and we provide the measurement apparatus to answer without circularity.