REASONING GEOMETRY OF LANGUAGE MODELS: HOW PROMPTING, TRAINING, AND SCALE SHAPE JUDGMENT TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model reasoning exhibits behaviorally anchored functional geometry, but prompting and post-training reshape it in fundamentally different ways. Controlled contrasts identify reasoning-associated ($\mathbf{u}$) and judgment-associated ($\mathbf{v}$) factors: reasoning is $\mathbf{u}$-weighted (S_R=0.96), judgment is $\mathbf{v}$-weighted (S_J=−0.36), yielding Γ=1.33. A 2×2 factorial reveals asymmetric interaction: reasoning remains $\mathbf{u}$-weighted across policy states, while judgment geometry depends on reasoning context. **Prompting and post-training act as distinct geometric operators:** prompting reweights engagement of this functional organization, whereas post-training reorients the representation frame in which it is expressed. Held-out Procrustes alignment removes 95% of the final-layer $\mathbf{v}$–$\mathbf{u}$ rotation gap, showing that most angular change reflects shared frame reorientation rather than loss of functional structure. Despite coordinate variation, relative functional organization persists across probes, training states, and tasks. Interventions along either factor are no more consequential than norm-matched random directions, separating behavioral readability from selective steerability. Endpoint shift predicts reflection success (AUROC 0.85). Together, these results identify relational trajectory geometry as a substrate for model observability, distinguishing functional change from representational reorientation across prompting and post-training.

1 INTRODUCTION

Open-weight models are increasingly modified through prompting, fine-tuning, preference optimization, and deployment-time adaptation. Yet representational change alone does not reveal *what kind of internal change occurred*: a model may change how it engages an existing functional organization, or the representation frame itself may change while functional relations persist. Distinguishing these possibilities is essential for interpreting internal change across prompting and post-training. We study this problem through two operational functions: *reasoning demand*, whether the model must construct a derivation, and *judgment policy*, the criterion for committing to versus withholding an answer given a fixed reasoning trace. Controlled manipulations reveal an asymmetric, interacting geometry: reasoning remains preferentially $\mathbf{u}$-weighted across policy states, whereas judgment geometry depends on reasoning context.

Prompting and post-training act as distinct geometric operators. Prompting primarily reweights engagement of this functional organization: CoT/ReAct-style scaffolds are more $\mathbf{u}$-weighted than few-shot prompting. Post-training instead reorients the representation frame in which the organization is expressed. This reveals a distinction that representational drift alone cannot make: prompting can change how a functional organization is engaged, while post-training can substantially move its coordinates without erasing the organization itself. Held-out Procrustes alignment removes 95% of the final-layer $\mathbf{v}$–$\mathbf{u}$ rotation gap, while the relative functional ordering survives across probe frames, supervision constructions, and tasks. Thus large representational change need not imply functional drift. Endpoint shift predicts reflection success (AUROC 0.85), further showing that this geometry provides a substrate for model observability.

We test these claims primarily on Qwen2.5 (1.5B–72B) and GSM8K, with additional task- and architecture-level evaluations (Yang et al., 2024; Cobbe et al., 2021).

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.

Contributions. (1) **Behaviorally anchored relational geometry:** reasoning and judgment form an asymmetric, interacting organization whose coordinate realization varies across probes, checkpoints, supervision objectives, and tasks, while their relative ordering is more reproducible (§5.4). (2) **Distinct geometric operators:** prompting reweights engagement of this organization, whereas post-training reorients its representation frame while preserving relative functional ordering (§5.6). (3) **Monitoring utility:** trajectory endpoint shift predicts reflection success (AUROC 0.85 vs. 0.55 for the best scalar baseline; Appendix v).

2 Related Work

Prompting and in-context learning. CoT (Wei et al., 2022), ToT (Yao et al., 2023a), self-consistency (Wang et al., 2023b), and ReAct (Yao et al., 2023b) alter reasoning through intermediate computation, search, aggregation, and tool interaction. In-context learning further shows strong effects of exemplar structure and format (Min et al., 2022). We ask how prompting changes engagement of behaviorally anchored functional organization across layers and positions.

Training-time geometry. Instruction tuning, preference optimization, and parameter-efficient adaptation alter behavior and parameters differently (Ouyang et al., 2022; Rafailov et al., 2023; Hu et al., 2021), motivating study of how post-training reorganizes internal representations. We compare post-training and prompting within a shared functional geometry, distinguishing functional engagement from representation-frame change.

Representation geometry and probing. Work on linear representation structure and probing asks how semantic and behavioral information is organized in hidden states (Park et al., 2024; Hernandez et al., 2024; Burns et al., 2023; Li et al., 2023; Gurnee & Tegmark, 2024). Related work studies layer-wise dynamics and representation evolution during reasoning (Su et al., 2024; Wang et al., 2023a), while circuit methods and sparse autoencoders seek localized or mechanistic decompositions of model computation (Conmy et al., 2023; Bricken et al., 2023). Activation steering and refusal or abstention directions show that behaviorally relevant structure can be recovered from low-dimensional representation space (Arditi et al., 2024; Zou et al., 2023a; Turner et al., 2023; Li et al., 2023). Steering work typically reads a direction's behavioral effect as evidence of its functional identity; we test the converse and find that behaviorally identified factors need not be steerable, so the two properties require separate validation. Rather than treating learned directions as fixed latent axes, we study relational geometry across trajectories, asking which functional relations persist as prompting and post-training reshape their coordinates.

Geometric reasoning frameworks (concurrent). Recent work studies reasoning through representation-space flows, layer-wise dynamics, failure localization, and progress or stability metrics (Li et al., 2025a; Zhou et al., 2026; Tacheny, 2025; Li et al., 2025b; Kirsanov et al., 2025; Jiang et al., 2026). These approaches characterize the geometry and dynamics of reasoning itself. We instead study how prompting and post-training transform behaviorally anchored functional geometry, and what persists across those transformations.

Prompting–training interaction. Few-shot behavior depends on demonstration structure (Min et al., 2022), CoT gains vary with scale (Wei et al., 2023), and instruction tuning changes in-context behavior (Longpre et al., 2023). We therefore ask whether prompting and post-training act on functional organization through distinct geometric operations.

3 Framework and Theoretical Foundations

3.1 Hidden-state trajectories

Consider a transformer with L layers processing T tokens. At each layer ℓ, hidden states $\mathbf{h}_1^{(\ell)}, \ldots, \mathbf{h}_T^{(\ell)} \in \mathbb{R}^H$ trace a trajectory through representation space. We write the layer-wise update as

$$\mathbf{h}_t^{(\ell+1)} = F_\theta^{(\ell)}\left(\mathbf{h}_t^{(\ell)},\ \{\mathbf{h}_s^{(\ell)}\}_{s\le t}\right), \tag{1}$$

where $F_\theta^{(\ell)}$ is the ℓ-th transformer block with parameters θ. Prompting changes context under fixed θ, whereas post-training changes θ itself. We ask whether these interventions induce distinct transformations of trajectory geometry.

3.2 Geometric diagnostics

We use three complementary diagnostics: *divergence* for paraphrase stability, *curvature* for local trajectory structure, and *projected displacement* for engagement of behaviorally anchored functional readouts.

Trajectory divergence. Given semantically equivalent inputs x, x' (e.g., paraphrases of the same question) under the same prompt strategy, let $e(x)$ denote the final token of the complete prompt. We measure endpoint divergence at layer ℓ as

$$D_{\cos}(x, x'; \ell) = 1 - \cos\Big(\mathbf{h}^{(\ell)}_{e(x)}(x),\ \mathbf{h}^{(\ell)}_{e(x')}(x')\Big). \tag{2}$$

Lower divergence indicates greater representational stability. Unless stated otherwise, divergence and endpoint shift use the prompt endpoint; suffix controls additionally use the question endpoint, which is exactly invariant by Proposition 3.2.

Discrete curvature. For a trajectory of L2-normalized states $\hat{\mathbf{h}}^{(\ell)}_1, \ldots, \hat{\mathbf{h}}^{(\ell)}_T$, we define the mean second-difference norm

$$\kappa^{(\ell)} = \frac{1}{T-2} \sum_{t=2}^{T-1} \left\| \hat{\mathbf{h}}^{(\ell)}_{t+1} - 2\hat{\mathbf{h}}^{(\ell)}_t + \hat{\mathbf{h}}^{(\ell)}_{t-1} \right\|. \tag{3}$$

This is a metric-specific measure of local second-order trajectory variation, not an intrinsic curvature or direct measure of reasoning complexity.

Projected displacement. Let $\phi_u^{(\ell)}$ and $\phi_v^{(\ell)}$ denote fixed projection heads introduced in §5.3 and behaviorally validated in §5.4. For brevity, $\mathbf{h}^{(c)}_{q,p}$ below denotes the prompt-endpoint state at the layer being analyzed. For question q, we use the mean Direct representation across its paraphrases as the reference,

$$\bar{\mathbf{u}}^{\text{dir}}_q = \tfrac{1}{P} \textstyle\sum_{p=1}^{P} \phi_u^{(\ell)}(\mathbf{h}^{(\text{dir})}_{q,p}), \qquad \bar{\mathbf{v}}^{\text{dir}}_q = \tfrac{1}{P} \textstyle\sum_{p=1}^{P} \phi_v^{(\ell)}(\mathbf{h}^{(\text{dir})}_{q,p}). \tag{4}$$

For condition c and paraphrase p, projected displacements are

$$\Delta\mathbf{u}^{(c)}_{q,p} = \phi_u^{(\ell)}(\mathbf{h}^{(c)}_{q,p}) - \bar{\mathbf{u}}^{\text{dir}}_q, \qquad \Delta\mathbf{v}^{(c)}_{q,p} = \phi_v^{(\ell)}(\mathbf{h}^{(c)}_{q,p}) - \bar{\mathbf{v}}^{\text{dir}}_q. \tag{5}$$

With N questions and P paraphrases per question, we report mean per-example displacement magnitudes,

$$D_u^{(c)} = \tfrac{1}{NP} \textstyle\sum_{q=1}^{N} \sum_{p=1}^{P} \|\Delta\mathbf{u}^{(c)}_{q,p}\|_2, \qquad D_v^{(c)} = \tfrac{1}{NP} \textstyle\sum_{q=1}^{N} \sum_{p=1}^{P} \|\Delta\mathbf{v}^{(c)}_{q,p}\|_2, \tag{6}$$

and use $D_u^{(c)}/D_v^{(c)}$ as an operational measure of relative factor engagement. Direct is slightly nonzero because individual paraphrases are measured relative to their per-question Direct mean.

Measurement convention. States are L2-normalized before measurement. Divergence and later directional comparisons use cosine geometry; curvature uses Euclidean second differences of normalized states.

Assumption 3.1 (Causal transformer)**.** At every layer, the hidden state at position t depends only on preceding-layer states at positions $s \le t$.

3.3 Causal position invariance

Proposition 3.2 (Suffix invariance)**.** *Under Assumption 3.1, let $x = (x_1, \ldots, x_T)$ and $x' = (x_1, \ldots, x_T, s_1, \ldots, s_K)$ differ only by an appended suffix. Then, for every prefix position $t \le T$ and every layer ℓ,*

$$\mathbf{h}^{(\ell)}_t(x) = \mathbf{h}^{(\ell)}_t(x').$$

Table 1: Mean prompt-endpoint divergence (cosine distance) across paraphrases by layer and prompting strategy. Higher values indicate lower representational stability.

Layer	Direct	CoT	Few-shot	ReAct-style
7	0.338	0.338	0.462 (+37%)	0.425 (+26%)
14	0.300	0.300	0.403 (+34%)	0.374 (+25%)
21	0.280	0.280	0.416 (+49%)	0.352 (+26%)
27	0.288	0.287	0.384 (+33%)	0.380 (+32%)

Intuition: Causal masking prevents appended tokens from affecting prefix states, making them exactly invariant. Accordingly, suffix CoT produces numerically tiny question-endpoint differences but much larger prompt-endpoint changes (Appendix h). Proof in Appendix a.

4 Experimental setup

Model. Primary experiments use Qwen2.5-1.5B-Instruct (Yang et al., 2024) (28 layers, 1536 dimensions), with cross-scale evaluation through 72B (N=100 per scale) and additional span-matched curvature control at 7B.

Data and prompts. We use GSM8K (Cobbe et al., 2021) with 3 rule-based paraphrases per question (fixed seed), verified equivalent by MiniLM (mean cosine 0.986, all > 0.9). Core divergence findings replicate with independently generated paraphrases (GPT-4o-mini; Appendix y). Prompt conditions are *Direct* (no scaffolding), *CoT* ("Let's think step by step" suffix), *Few-shot* (3 worked examples), and *ReAct-style* (Thought/Action/Observation template).

Extraction and statistics. Hidden states are extracted at layers 7, 14, 21, and 27 (quarter, half, three-quarter, and final depth) and L2-normalized before measurement. For paired comparisons, we average the three paraphrases within each question and use Wilcoxon signed-rank tests over N=100 question-level pairs. We report paired standardized mean differences $d_z = \bar{D}/s_D$; positive d_z indicates that Direct exceeds the comparison condition. For predefined multi-layer divergence and curvature families, we apply Benjamini–Hochberg FDR correction at α=0.05; other analyses use experiment-specific tests described in Appendix y.

5 Experiments

5.1 Trajectory divergence under paraphrase

Divergence across conditions. Prompting strategies differ sharply in paraphrase stability. Direct and CoT have nearly identical prompt-endpoint divergence ($\Delta < 0.001$; Table 1), whereas Few-shot and ReAct-style increase divergence by 33–49% and 25–32%, respectively. This contrast follows the causal structure of the prompt: an appended suffix cannot alter representations already computed for the question prefix (Proposition 3.2), whereas preceding context can affect subsequent states. At the matched-token question endpoint, Few-shot still increases divergence by 19–31% (Appendix y), showing that the effect reflects preceding context rather than endpoint-token identity.

5.2 Curvature profiles across prompt strategies

Curvature decreases with depth across all prompting conditions ($\sim$1.7 at L7 to $\sim$1.4 at L27; Fig. 1a).

Few-shot's early-layer difference ($d_z = +1.90$ at L14) vanishes under span matching, while full-trajectory late-layer differences emerge strongly from 7B through 72B. At 1.5B, a span-matched late-layer contrast vs. length-matched filler gives $d = +0.43$, p=0.012. At 7B, span matching preserves substantial differences (L20: $d = -1.47$; L27: $d = -0.56$), with span- and depth-dependent sign (Appendix y). A direction-only turning-angle metric recovers substantial depth-dependent prompting effects, including the early-to-late reversal for few-shot, while differing in the final-layer CoT effect (Appendix y). Thus prompting induces robust depth-dependent trajectory structure, while its exact layerwise profile depends on span and metric.

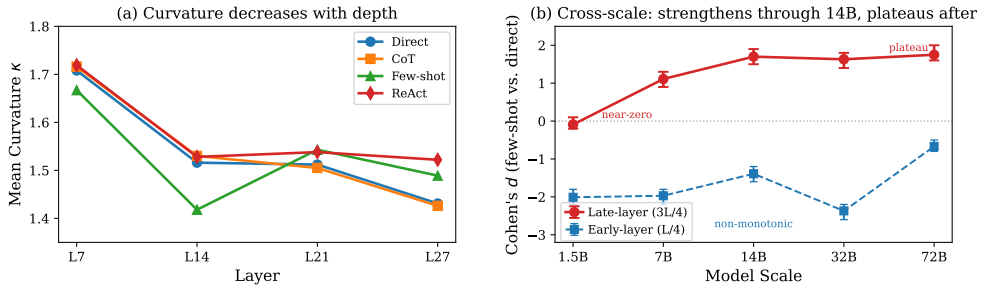


Figure 1: (a) Curvature decreases with depth. Few-shot's early-layer difference at 1.5B vanishes under span matching. (b) Full-trajectory late-layer differences emerge strongly from 7B through 72B; a 7B span-matched control preserves substantial mid-to-late differences, with span-dependent sign (§5.7).

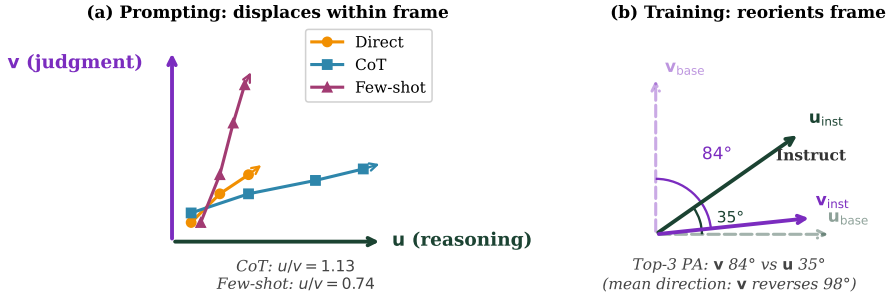


Figure 2: Reasoning- and judgment-associated readouts. (a) CoT/ReAct are relatively $\mathbf{u}$-weighted, while Few-shot has lower $\mathbf{u}/\mathbf{v}$ ratios. (b) At final depth, $\mathbf{v}$ rotates $96°$ and $\mathbf{u}$ $35°$; held-out Procrustes attributes most of the gap to shared reorientation (Appendix n).

Causal position and context tests. At the question endpoint, suffix CoT differs negligibly from Direct ($\Delta < 0.001$), consistent with Proposition 3.2, whereas placing the same cue before the question increases divergence by 7–14% at L7–L21. Thus the same cue has different representational effects depending on causal position. The proposition constrains only prefix states, not prompt-endpoint suffix effects such as the small effect at 14B (Table 4).

5.3 Projected displacement via projection heads

Setup. We train MLP projection heads on a separate identity-tracking pretext (200 episodes, $K{=}4$, entity-disjoint split), designed only to induce generic state-like and progression-like structure: $\mathbf{v}$ separates entity identity and $\mathbf{u}$ tracks within-entity temporal order. The pretext assigns no reasoning or judgment semantics; those labels are established only by the independent behavioral contrasts below. A shared $1536{\to}256{\to}256{\to}128$ MLP splits into 64-dimensional readouts. Training details are in Appendix e.

Prompting changes relative factor engagement: CoT and ReAct-style are more $\mathbf{u}$-weighted than Few-shot (Figure 2a). This ordering is depth- and probe-dependent (3/4 heads at L27), whereas the relative behavioral organization established below is more reproducible (Appendix t).

Prompt-level validation. The trained readouts transfer from the synthetic pretext to GSM8K: CoT exceeds Few-shot in relative $\mathbf{u}$ engagement ($d{=}1.10$), whereas random projections do not reliably recover the same ordering (Appendix t).

5.4 Behavioral semantics of the $\mathbf{u}/\mathbf{v}$ factors

The pretext supplies a coordinate frame, not semantic labels. We assign functional meaning only through independent behavioral contrasts, then test whether the resulting relation survives changes in probes, wording, supervision, and tasks.

Table 2: Projected displacement at layer 14 (prompt endpoint). The $\mathbf{u}/\mathbf{v}$ ratio summarizes relative projected displacement under the fixed probe; interpretation focuses on cross-condition ordering.

Condition	$\|\Delta\mathbf{u}\|$	$\|\Delta\mathbf{v}\|$	$\mathbf{u}/\mathbf{v}$ ratio
Direct	0.003 ± 0.002	0.004 ± 0.003	0.75
CoT	0.053 ± 0.007	0.047 ± 0.004	**1.13**
Few-shot	0.031 ± 0.004	0.042 ± 0.004	**0.74**
ReAct-style	0.079 ± 0.005	0.054 ± 0.003	**1.47**

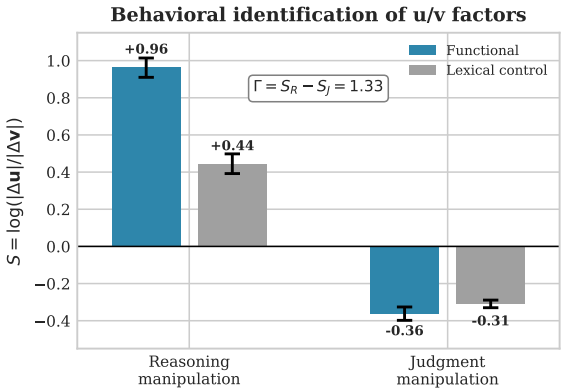


Figure 3: Behavioral anchoring of $\mathbf{u}/\mathbf{v}$ (base checkpoint). Reasoning demand is $\mathbf{u}$-weighted (S_R=0.96), while fixed-trace judgment is $\mathbf{v}$-weighted (S_J=−0.36), yielding Γ=1.33. Gray bars show matched lexical controls; error bars are 95% bootstrap CIs over N=200. The relative dissociation persists across checkpoints (Appendix o).

Independent behavioral identification. On N=200 held-out GSM8K problems using the Qwen2.5-1.5B base checkpoint, we vary *reasoning demand* by requiring a derivation (R+) versus supplying it (R−), while holding the problem and answer fixed. We vary *judgment policy* with a byte-identical supplied trace, contrasting permissive ("return unless explicit contradiction") and conservative ("commit only if fully supported") decision rules. For each question, we compute the paired R+−R− and J+−J− projected displacements $\Delta\mathbf{u}_i$, $\Delta\mathbf{v}_i$ and define

$$S_i = \log \frac{\|\Delta\mathbf{u}_i\| + \epsilon}{\|\Delta\mathbf{v}_i\| + \epsilon},$$

where ϵ=10^{-6} ensures numerical stability. We report the mean S over questions for each condition and define $\Gamma = S_R - S_J$. In the $\epsilon \to 0$ limit, fixed multiplicative rescaling of either factor adds a constant to every S_i and therefore cancels in Γ.

Reasoning is preferentially $\mathbf{u}$-weighted (S_R=0.96, 95% CI $[0.91, 1.01]$), while fixed-trace judgment is preferentially $\mathbf{v}$-weighted (S_J=−0.36, $[-0.40, -0.33]$), yielding Γ=1.33 ($[1.26, 1.39]$; Fig. 3). Paired functional-over-lexical effects are significant for both contrasts ($p < 10^{-4}$; Appendix ad); the judgment excess is smaller but reliably positive. We therefore use "reasoning" and "judgment" as shorthand for these operational contrasts, not as a general cognitive decomposition.

Robustness across probes, supervision, and tasks. Across all tested heads, Γ>0, although the full signed pattern S_R>0, S_J<0 holds only in a subset, indicating that relative functional organization is more reproducible than absolute signed orientation (Appendix t). The signed judgment contrast is not tied to a single policy wording: three independently phrased decision-rule variants all yield S_J<0 (range −0.38 to −0.23; Appendix ad). The relation also survives an independently constructed answer-choice/stage supervision objective (Γ=0.248, 95% CI $[0.18, 0.31]$; Appendix ad). Across tasks, Γ remains positive on CommonsenseQA (0.10, $[0.04, 0.17]$) and ProofWriter (0.068, $[0.035, 0.101]$), despite task-dependent signed judgment geometry (Appendix ad, ae). The ordering S_R>S_J thus reproduces on the two further tasks, but with small margins; only GSM8K yields the signed pattern. We do not claim that Γ is invariant to arbitrary reparameterizations of the probe output

Table 3: $\mathbf{u}/\mathbf{v}$ under post-training. $\mathbf{v}$ magnitude decreases at early depth under the fixed probe, while the large final-layer angular gap is mostly explained by shared reorientation (Appendix n).

Layer	Base $\mathbf{u}/\mathbf{v}$	Instruct $\mathbf{u}/\mathbf{v}$	$\mathbf{v}$ angle	$\mathbf{u}$ angle	Effect
7	0.70	2.02	56°	21°	$\mathbf{v}\downarrow$ (t=−28.3, p<0.001)
14	1.47	1.66	61°	37°	$\mathbf{v}$ stable
21	2.35	2.79	40°	44°	$\mathbf{v}\downarrow$
27	1.77	1.95	**96°**	**35°**	shared reorientation (App. n)

space; rather, we test whether the ordering it measures reproduces across independently trained and differently supervised probe constructions.

Asymmetric factorial structure. A 2×2 factorial (N=200) crosses reasoning demand with judgment policy. Reasoning remains $\mathbf{u}$-weighted for both policies (S_R=1.34 permissive, 0.18 conservative). Judgment changes with reasoning context: it is $\mathbf{v}$-weighted with supplied reasoning (S_J=−0.30) but $\mathbf{u}$-weighted under active derivation (S_J=0.21; Appendix ad). Active derivation also produces longer traces (702 vs. 72 characters), as expected under the active-derivation condition. Thus the functions are distinguishable but not globally independent: reasoning context changes the geometric expression of judgment.

5.5 The factors are readable but not selectively steerable

We test whether the behaviorally identified factors are causal handles. Two intervention methods agree that they are not. A linear-surrogate back-projection (2–20% of $\|\mathbf{h}\|$, L14/L27, N=50) changes generation at rates indistinguishable from norm-matched random directions (Appendix t). Exact local-Jacobian steering at the decision logit (L27, α=0.05, held-out N=200) yields $|\Delta_v|$=0.370 and $|\Delta_u|$=0.062; the ratio is large and precisely estimated, but $|\Delta_v|$ sits at the 45th percentile of norm-matched random directions, $|\Delta_u|$ at the 20th, and the v–u gap at the 57th percentile of gaps between random-direction pairs. The asymmetry is therefore within the ordinary spread of arbitrary directions, not a factor-specific property.

This separates two notions that are easy to conflate. The factors are behaviorally identifiable (§5.4) and informative as monitoring signals, yet perturbing along them is no more consequential than perturbing along a random direction of the same norm. Behavioral identifiability of a low-dimensional readout does not imply that the readout is a control axis.

Self-consistency and decision stability. Self-consistency reduces answer variance through output aggregation rather than reduced trajectory curvature; curvature predicts CoT, but not Few-shot, decision stability (Appendix j).

Post-reflection accept/abstain. Reflection (Madaan et al., 2023) reduces second-difference curvature across layers and raises final accuracy from 13% to 47.5%. Of 174 initially incorrect cases, 44.8% are corrected; endpoint shift between passes predicts correction success (AUROC 0.85; 0.84 on all 200 cases without conditioning on pass-1 outcome), outperforming predictive entropy, perplexity, response-length change, and curvature change (best AUROC 0.55). Because the shift requires both passes, it supports post-reflection accept/abstain rather than pre-reflection routing (Appendix k, q, v).

5.6 Base vs. instruction-tuned

Post-training reorients the representation frame. Applying the fixed Instruct-trained projection to base and instruct states reveals two effects. First, $\mathbf{v}$ magnitude decreases strongly at early depth ($\mathbf{u}/\mathbf{v}$: 0.70→2.02 at L7, t=−28.3, p<0.001), while $\mathbf{u}$ changes less. Second, the final-layer mean $\mathbf{v}$ direction rotates 96° versus 35° for $\mathbf{u}$ (Fig. 2b). Held-out orthogonal Procrustes on paired L27 states (5-fold, N=300) reduces this raw 61° gap to 2.9° (95%), showing that most angular change reflects shared frame reorientation rather than loss of functional organization (Appendix n). The remaining fixed-probe signature is therefore primarily factor-differentiated magnitude rather than a $\mathbf{v}$-specific rotation.

Table 4: Cross-scale results (N=100). Suffix CoT yields tiny question-endpoint (QE) differences but substantial prompt-endpoint (PE) changes; prefix effects are stable across scale. Curvature d compares Few-shot with Direct on the full trajectory (sign: Few-shot minus Direct); span matching was performed at 1.5B and 7B (Appendix y).

Metric	1.5B	7B	14B	32B	72B
QE dist	2×10^{-6}	1.2×10^{-4}	3.8×10^{-5}	3×10^{-5}	6.2×10^{-5}
PE dist	0.12	0.15	0.22	0.15	0.15
PE/QE	65k	1.2k	5.6k	5.0k	2.5k
Prefix (L/4, %)	+14.4	+13.7	+13.8	+14.9	+14.3
Curv. d, early	-2.01 $[-2.2, -1.8]$	-1.97 $[-2.2, -1.8]$	-1.39 $[-1.6, -1.2]$	-2.37 $[-2.6, -2.2]$	-0.67 $[-0.8, -0.5]$
Curv. d, late	-0.09 $[-0.2, 0.1]$	$+1.11$ $[0.9, 1.3]$	$+1.70$ $[1.5, 1.9]$	$+1.63$ $[1.4, 1.8]$	$+1.75$ $[1.6, 2.0]$

The behavioral dissociation nevertheless survives post-training. Repeating the contrasts on the instruct model with the same fixed projection and stimuli preserves the signed separation, with a larger measured Γ in that fixed Instruct-trained probe frame ($1.33\rightarrow1.59$, non-overlapping CIs; Appendix o). A direct-baseline control indicates that this increase is not explained by uniform magnitude rescaling. An independently trained Base probe changes absolute signed orientation but preserves $S_R > S_J$, showing that probe provenance affects coordinate realization more strongly than relative functional organization (Appendix p). Post-training thus substantially changes coordinate realization while preserving the relative functional organization.

Cross-family. A Mistral base–instruct comparison provides a second-family instance: both families show their strongest divergence change at intermediate depth and similar 3L/4 curvature decreases (Qwen: -6.0%; Mistral: -5.9%), with architecture-dependent magnitudes (Appendix m).

Prompting–training interaction. Prompting and post-training affect the validated geometry differently. CoT remains relatively $\mathbf{u}$-weighted across base and instruct models, while Few-shot maintains lower $\mathbf{u}/\mathbf{v}$ ratios (Appendix o). Together with the frame reorientation above, these results support distinct geometric roles: prompting changes engagement of the functional organization, whereas post-training changes the representation frame in which it is expressed. Direct compositional effects are analyzed in Table 7 and Appendix d.

5.7 Cross-scale validation

Scale-dependent signatures. Full-trajectory late-layer curvature differences emerge strongly from 7B through 72B. Span matching removes the early 1.5B difference, while at 7B substantial differences remain (L20: $d=-1.47$; L27: $d=-0.56$) with opposite sign to the full-trajectory effect. Thus prompt-sensitive curvature survives span control at larger scale, but its direction depends on span definition and depth.

Causal position across scale. Across 1.5B–72B, suffix CoT yields tiny QE differences but orders-of-magnitude larger PE changes (PE/QE 1.2×10^3–6.5×10^4), consistent with Proposition 3.2. A 14B forward-twice control gives a $\sim10^{-8}$ numerical floor, while prefix CoT increases divergence by 13.7–14.9% across scale.

Task and architecture breadth. On CommonsenseQA, Few-shot increases L7 divergence by $3.3\times$, with a different depth profile (Appendix b). Sparse-MoE OLMoE-1B-7B also shows near-invariant QE states (PE/QE 500–$7000\times$) and late-layer curvature differences ($d=+0.60$–$+0.88$; full-trajectory, not span-controlled; Appendix c).

6 Discussion

Distinct operators on functional geometry. Prompting and post-training reshape functional geometry in different ways. Prompting primarily changes how an existing organization is engaged, with CoT/ReAct-style scaffolds shifting trajectories toward the reasoning-associated factor relative to few-shot prompting. Post-training instead largely reorients the representation frame in which

this organization is expressed while preserving relative reasoning–judgment structure. The central distinction is therefore between *functional engagement* and *representation-frame change*, not simply between small and large representational shifts.

An interacting, relational organization. Reasoning and judgment are behaviorally distinguishable but not globally independent. Reasoning remains $\mathbf{u}$-weighted across policy states, whereas judgment geometry depends on whether reasoning is supplied or actively generated. Moreover, absolute coordinates vary across probe frames, checkpoints, supervision constructions, and tasks, while their relative organization is more reproducible. We therefore interpret $\mathbf{u}$ and $\mathbf{v}$ as behaviorally identified coordinates within a learned frame, not universal latent axes. The relative organization is also sample-efficient: on the primary GSM8K setting, repeated subsampling detects $\Gamma{>}0$ reliably with as few as 10 questions (Appendix u).

Three distinctions. Together, the experiments separate three notions that are often conflated: functional engagement from representation-frame change, coordinate realization from relational organization, and behavioral readability from selective steerability. A 96° apparent rotation is largely removed by held-out alignment; the $2{\times}2$ factorial reveals conditional interaction rather than independent factors; and two intervention methods find no selectivity over norm-matched random directions. Thus large representational change need not imply functional drift, distinguishability need not imply independence, and readability need not imply steerability.

The reproducible object. Across these analyses, the reproducible object is not a particular signed coordinate system but a relational functional organization: absolute coordinates vary across probes, supervision objectives, tasks, and post-training, while the reasoning–judgment ordering persists and can be recovered from small calibration sets. Post-training provides the strongest instance of this distinction: a large apparent factor-specific rotation is largely removed by a shared frame alignment that itself can be estimated from few paired examples.

Implications for observability. Endpoint shift predicts reflection success (AUROC 0.85 versus 0.55 for the best tested scalar baseline), showing that useful signal lies not simply in how much a trajectory moves but in how its representation changes relative to its prior state. This suggests monitoring adapted models through functionally anchored, alignment-aware trajectory statistics rather than raw directional movement alone.

Limitations and scope. The ordering $S_R{>}S_J$ survives independently trained probe frames and a distinct supervision objective but is only weakly positive on two further tasks, and the readouts capture only $\sim 4\%$ of displacement norm. The reflection predictor is established on GSM8K; frozen transfer to CommonsenseQA is inconclusive (AUROC 0.61 on only 19 initially incorrect cases; Appendix v), leaving task-general monitoring open.

Beyond reasoning and judgment. The present $\mathbf{u}/\mathbf{v}$ construction captures one pair of operational functions, not a unique decomposition of model computation. The judgment-associated factor is also nearly orthogonal to independently constructed refusal and abstention directions (Appendix af), suggesting that it does not simply recover those control directions. More broadly, the same framework could track interacting functional relations such as uncertainty, policy adherence, and safety boundaries while separating shared representational reorientation from function-specific drift.

7 Conclusion

Language-model reasoning has monitorable functional geometry. Reasoning demand and judgment policy produce behaviorally distinguishable but interacting signatures; prompting changes how this organization is engaged, whereas post-training largely reorients the representation frame in which it is expressed while preserving relative functional organization. Large representational movement therefore need not imply functional drift. Endpoint shift predicts reflection success (AUROC 0.85), and intervention tests show that the learned factors are readable but not selectively steerable relative to norm-matched random directions. Together, these results establish structured internal trajectories as a substrate for model observability, revealing how models reason and change in ways that outputs alone cannot capture.

REPRODUCIBILITY STATEMENT

All experiments use publicly available models (Qwen2.5-{1.5B, 7B, 14B, 32B, 72B}-Instruct, Mistral-7B variants, OLMoE-1B-7B; appendix pilots additionally use TinyLlama-1.1B-Chat and Gemma-2B-it) and publicly available datasets including GSM8K, CommonsenseQA, and ProofWriter. Paraphrases and experimental splits use fixed seeds. Code, trained projection heads, scripts, and cached trajectories sufficient to reproduce the reported analyses will be released upon acceptance.

AI USE STATEMENT

Generative AI tools (Claude and GPT) were used to provide feedback on research methodology and experimental design, assist with sanity-checking the interpretation of experimental results, check the theoretical framework and the proof of Proposition 3.2 for correctness and clarity, refine presentation and prose, generate simulated peer reviews, organize manuscript text, assist with code implementation, draft and edit portions of the appendix based on experiment logs, and generate an independent paraphrase set used in a robustness analysis (Appendix y). We did not use generative AI tools to develop the conceptual framework, generate original claims, propose hypotheses, translate, or clean datasets. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Kedia, Rajdeep Puri, and Balaji Patil. Explanations for CommonsenseQA: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 3050–3065, 2021.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. Anthropic.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

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

Evan Hernandez, Martin Wattenberg, Arnab Sen Sharma, Jacob Andreas, Tal Haklay, Yonatan Belinkov, Kevin Meng, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, 2024.

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.

Xinyan Jiang, Ninghao Liu, Di Wang, and Lijie Hu. Beyond scalars: Evaluating and understanding LLM reasoning via geometric progress and stability. In *International Conference on Machine Learning*, 2026.

Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1855–1888, 2025.

Bo Li, Guanzhi Deng, Ronghao Chen, Junrong Yue, Shuo Zhang, Qinghua Zhao, Linqi Song, and Lijie Wen. REMA: A unified reasoning manifold framework for interpreting large language model. *arXiv preprint arXiv:2509.22518*, 2025a.

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

Melody Zixuan Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Kumar Teru, Adam Santoro, Guillaume Lajoie, and Blake A. Richards. Tracing the representation geometry of language models from pretraining to post-training. *arXiv preprint arXiv:2509.23024*, 2025b. NeurIPS 2025.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. OLMoE: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*, 2024.

Nicolas Tacheny. Dynamics of agentic loops in large language models: A geometric theory of trajectories. *arXiv preprint arXiv:2512.10350*, 2025.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, 2021.

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

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru R. Zhang. The geometry of reasoning: Flowing logics in representation space. In *International Conference on Learning Representations*, 2026.

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

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

Note to reviewers: The appendices contain essential methodological details referenced throughout the main text, including: statistical corrections (App. y), evaluation protocols (App. f), intervention procedures (App. t), and reflection setup (App. k).

A PROOFS

Proof of Proposition 3.2. By induction on ℓ. At $\ell = 0$, the embedding layer maps each token independently: $\mathbf{h}_t^{(0)}(x) = \mathrm{Embed}(x_t) = \mathbf{h}_t^{(0)}(x')$ for $t \leq T$. For the inductive step, suppose $\mathbf{h}_s^{(\ell)}(x) = \mathbf{h}_s^{(\ell)}(x')$ for all $s \leq T$. By Assumption 3.1, $\mathbf{h}_t^{(\ell+1)}(x')$ depends only on $\{\mathbf{h}_s^{(\ell)}(x')\}_{s \leq t}$. Since $t \leq T$, the suffix positions $s > T$ are excluded by the causal mask. By the inductive hypothesis, all terms in the conditioning set are identical, so $\mathbf{h}_t^{(\ell+1)}(x') = \mathbf{h}_t^{(\ell+1)}(x)$. □

B COMMONSENSEQA TASK GENERALIZATION

To test whether geometric signatures generalize beyond mathematical reasoning, we evaluate on CommonsenseQA (N=100, 3 paraphrases per question).

Table 5: CommonsenseQA: divergence and curvature by layer and condition.

Metric	Condition	L7	L14	L21	L27
Divergence	Direct	0.138	0.130	0.119	0.162
	CoT	0.147	0.139	0.126	0.170
	Few-shot	**0.462**	**0.389**	**0.360**	**0.302**
Curvature	Direct	1.756	1.567	1.571	1.652
	CoT	1.769	1.574	1.560	1.551
	Few-shot	**1.745**	**1.542**	**1.514**	**1.292**

Prompt-sensitive geometry beyond GSM8K. Few-shot again substantially changes divergence on CommonsenseQA (3.3× at L7, p<0.001), and curvature is also strongly prompt-sensitive, although its depth profile and sign differ from the controlled GSM8K result. Thus geometric prompt effects are present beyond mathematical reasoning, but their form is task-dependent.

C MIXTURE-OF-EXPERTS ARCHITECTURE GENERALIZATION

To test whether geometric signatures depend on dense architecture, we evaluate OLMoE-1B-7B (Muennighoff et al., 2024), a sparse MoE model with 64 experts and top-8 routing per token (6.9B total, 1.3B active parameters, 16 layers).

Table 6: MoE architecture generalization (OLMoE-1B-7B, N=100). QE/PE position test and curvature effects at late depth fractions replicate on sparse MoE (not span-matched; as in dense models, full-trajectory effects may include length contributions). Layers aligned by depth fraction (L3/7/11/15 ≈ 25%/50%/75%/100%).

Metric	L3 (25%)	L7 (50%)	L11 (75%)	L15 (100%)
QE dist	8.5×10^{-5}	6.8×10^{-4}	5.8×10^{-4}	4.7×10^{-4}
PE dist	0.59	0.58	0.45	0.23
PE/QE	7.0k	0.9k	0.8k	0.5k
Prefix (%)	+20.4	+22.2	+22.8	+24.6
Curv. d (few-shot vs. direct)	−1.62	+0.60	+0.88	+0.65

Signatures replicate on sparse MoE. Despite fundamentally different computation (each token routes through 8 of 64 experts rather than a single FFN), geometric signatures appear: (1) QE/PE position test holds (QE dist 8.5×10^{-5}–6.8×10^{-4}, PE dist 0.23–0.59, ratio 500–7000×); (2) few-shot shows curvature differences at early layers ($d = -1.62$ at L3, p<0.001; potentially influenced by prompt length) and late layers ($d = +0.60$ to $+0.88$, all p<0.001). MoE QE distances are higher than dense models (10^{-4} vs 10^{-6}–10^{-4} for dense) but remain orders of magnitude below PE. This QE excess is concentrated in the mid layers (peak at L7, 6.8×10^{-4}; smallest at L3, 8.5×10^{-5}). The key finding is qualitative: *the near-invariant QE signature persists under sparse routing; late-layer curvature also differs, though this is full-trajectory (not span-controlled).*

D PROMPTING–TRAINING FACTOR ENGAGEMENT

Table 7 summarizes which factor each operator primarily engages.

Table 7: Prompting and training engage the factors differently. Prompting is reported as $\mathbf{u}/\mathbf{v}$ displacement ratio (L14); post-training as the $\mathbf{v}$ magnitude reduction under the fixed probe (L7) and behavioral dissociation change (Appendix o). Entries indicate the factor-specific signature of each operator.

Intervention	Relative signature	Measure
CoT prompting	relatively $\mathbf{u}$-weighted	$\mathbf{u}/\mathbf{v} = 1.13$
Few-shot prompting	lower $\mathbf{u}/\mathbf{v}$ than CoT/ReAct	$\mathbf{u}/\mathbf{v} = 0.74$
Post-training	$\mathbf{v}$ magnitude change + shared reorientation	$\mathbf{u}/\mathbf{v}$ ratio $0.70 \rightarrow 2.02$ (L7)

The $\mathbf{u}/\mathbf{v}$ geometry shows that prompt strategies retain different relative factor weightings across training states: CoT remains more $\mathbf{u}$-weighted than Few-shot, while post-training introduces factor-differentiated magnitude changes and shared reorientation (Appendix o).

E EXPERIMENTAL DETAILS

Paraphrase generation. For each GSM8K question, we generate 3 paraphrases using deterministic rule-based transformations: name substitution (e.g., "John" $\rightarrow$ "James"), structural reordering (moving clauses), and phrasing variation (e.g., "How many" $\rightarrow$ "What is the total number of"). Seed offsets $42 + i$ ensure reproducibility for question i. Semantic equivalence is validated via sentence-transformer embeddings (mean cosine similarity $= 0.986$, 100% above 0.9 threshold).

Prompt formatting. All prompts use Qwen's chat template with system message "You are a helpful assistant."

- **Direct:** User message contains only the question text.
- **CoT (suffix):** User message is the question text followed by "Let's think step by step."
- **CoT (prefix):** User message is "Let's think step by step." followed by the question text.
- **Few-shot:** Three worked examples are prepended as additional user/assistant turn pairs before the target question.
- **ReAct:** The question is wrapped in a "Thought: [question]\nAction: solve\nObservation:" template.

Trained projection heads. The projection heads used in Experiment 4 were pre-trained on a separate identity-tracking task using a composite loss:

$$\mathcal{L}_{\text{decomp}} = \mathcal{L}_{\text{same}} + \lambda_{\text{smooth}}\,\mathcal{L}_{\text{smooth}} + \lambda_{\text{barrier}}\,\mathcal{L}_{\text{barrier}} + \lambda_{\text{order}}\,\mathcal{L}_{\text{order}}, \tag{7}$$

where:

$$\mathcal{L}_{\text{same}} = \frac{1}{|\mathcal{P}^+|} \sum_{(i,j)\in\mathcal{P}^+} \|\phi_v(h_i) - \phi_v(h_j)\|^2, \tag{8}$$

$$\mathcal{L}_{\text{barrier}} = \frac{1}{|\mathcal{P}^-|} \sum_{(i,j)\in\mathcal{P}^-} \max(0, m - \|\phi_v(h_i) - \phi_v(h_j)\|), \tag{9}$$

$$\mathcal{L}_{\text{order}} = \frac{1}{|\mathcal{O}|} \sum_{t:t+2\le T} \max\big(0,\, \delta + d(\bar{u}_t, \bar{u}_{t+1}) - d(\bar{u}_t, \bar{u}_{t+2})\big), \tag{10}$$

$$\mathcal{L}_{\text{smooth}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\bar{u}_{t+1} - \bar{u}_t\|^2, \tag{11}$$

with $\mathcal{P}^+$ = same-entity pairs, $\mathcal{P}^-$ = cross-entity pairs (sampled uniformly), $\mathcal{O}$ = temporal triplets within each entity, margin m=1.0, and rank margin δ=0.1. The projection head is a shared three-layer

MLP ($1536 \to 256 \to 256 \to 128$, GELU activations) whose output splits into $\phi_u, \phi_v : \mathbb{R}^H \to \mathbb{R}^{64}$. In (10), $\bar{u}_t$ is the step-averaged $\mathbf{u}$-projection and $d(a, b) = 1 - \cos(a, b)$ is cosine distance. Each head is trained on identity-tracking episodes with $K{=}4$ entities per episode, using hidden states extracted from Qwen2.5-1.5B-Instruct. The same fixed projection heads are then applied unchanged to both base and instruct checkpoints for the post-training comparison. Training uses Adam (lr $= 10^{-3}$) with cosine annealing for 600 epochs, with unit weight on $\mathcal{L}_{\text{same}}$ and weights $\lambda_{\text{smooth}} = 0.5$, $\lambda_{\text{barrier}} = 0.5$, $\lambda_{\text{order}} = 0.2$. Heads are trained separately for each layer; we report diagnostic results from layers 7, 14, 21, and 27. The LLM parameters are entirely frozen during projection-head training; only the projection heads are optimized.

Decomposition training example. To illustrate the projection loss, consider a four-entity episode: "Entity A moves to the park. Entity B stays home. Entity A buys ice cream. Entity B reads a book." Hidden states are extracted at each token step. $\mathcal{L}_{\text{same}}$ pulls together the $\mathbf{v}$-projections of steps referencing Entity A ("moves to the park," "buys ice cream") and separately those referencing Entity B, so that each entity's states cluster in $\mathbf{v}$-space. $\mathcal{L}_{\text{barrier}}$ pushes Entity A's $\mathbf{v}$-cluster away from Entity B's. $\mathcal{L}_{\text{order}}$ enforces that Entity A's $\mathbf{u}$-projections respect the temporal ordering (step 1 before step 3). $\mathcal{L}_{\text{smooth}}$ regularizes consecutive $\mathbf{u}$-projections to vary gradually across steps. The result is a factored representation: $\mathbf{v}$ encodes *which* entity is active (entity identity), $\mathbf{u}$ encodes *where* in the episode the model is (temporal progress). The semantic interpretation of these factors as reasoning- and judgment-associated is established by independent behavioral validation (§5.4), not by the pretext task itself.

Irrelevant passages (Experiment 3). Noise passages are generated to be topically unrelated to math: excerpts about cooking, travel, and history. Each passage is 50–100 tokens. Passages are prepended in fixed order before the question, with no separator beyond whitespace.

Sequence-length controls. Few-shot prompts are 3–4$\times$ longer than direct prompts due to exemplars. Primary control: span-matched curvature is computed only over question-token spans, excluding exemplar tokens, ensuring length-matched comparisons. Divergence is measured at prompt-endpoint; the few-shot divergence increase (Table 1) also appears at question-endpoint (span-matched analysis, Appendix y). Paraphrases within each condition have similar lengths by construction, so within-condition divergence comparisons are not confounded by length.

F Evaluation protocol for accuracy measurements

All accuracy numbers reported in this paper follow a strict evaluation protocol to ensure reproducibility.

Model and inference. We use Qwen2.5-1.5B-Instruct (and 7B/14B/32B/72B variants for cross-scale) via HuggingFace Transformers. Decoding is **greedy** (temperature $= 0$, no sampling) for all accuracy measurements to ensure determinism. **Exception: self-consistency experiments** (Appendix j) require sampling; we use temperature $= 0.7$, top-$p = 0.95$, $n = 8$ samples per question. For SC, endpoint dispersion measures variability at the *completion endpoint* across the $n{=}8$ sampled generations. The chat template uses system prompt "You are a helpful assistant." with user/assistant roles.

Answer extraction. We extract the final numeric answer using a multi-stage regex:

1. If output contains "####", extract the number immediately following (GSM8K format).
2. Otherwise, search for patterns: "Final Answer: ⟨number⟩", "the answer is ⟨number⟩", or "= ⟨number⟩" at line end.
3. Fallback: extract the last number in the output.

Numbers are normalized (commas removed, converted to float). If no number is extracted, the answer is marked incorrect.

Data sampling. We sample 100 questions from GSM8K test set (1319 total) using seed $= 42$ with random shuffling. We filter to questions with ground-truth solution text < 300 characters to avoid unusually long solutions. The same primary 100-question sample is used for Experiments 1–4; experiments using larger or separate samples (C: N=200; rotation: N=300; cross-family: N=50) are noted explicitly.

Token budget experiments (Table 22). The "12% at 64 tokens → 80% at 512 tokens" scaling reflects a *truncation failure mode*: when `max_new_tokens` is insufficient for the model to complete its reasoning chain and output a final answer, accuracy drops catastrophically. At 128 tokens with CoT prompting, the model typically begins reasoning but is cut off before producing a parseable answer, yielding 0% under our extraction protocol. This is not a claim about model capability but about the interaction between generation budget and answer-extraction requirements. **Without the CoT suffix** (direct prompting), accuracy at 128 tokens is 12%: the model can output short answers when not prompted to reason extensively.

Baseline accuracy context. Published Qwen2.5 evaluations report substantially higher GSM8K accuracy under benchmark prompting and sufficient generation budgets (Yang et al., 2024). Our lower direct accuracy therefore reflects a different prompting and decoding regime rather than, by itself, an evaluation failure.

Few-shot exemplar selection. For few-shot prompting, we use 3 worked examples from the GSM8K training set (indices 0, 1, 2 after shuffling with seed $= 42$). These exemplars are: (1) a rate problem ("Natalia sold clips..."), (2) a subtraction problem ("Weng earns..."), (3) a multi-step arithmetic problem ("Betty is saving..."). The same 3 exemplars are used for all questions and all paraphrases to ensure consistency. Exemplars are formatted with full CoT reasoning chains from the training set annotations.

Ordering ablation. Table 8 shows $\mathbf{u}/\mathbf{v}$ displacement ratios under three exemplar orderings. The low $\mathbf{u}/\mathbf{v}$ ratio is consistent across all orderings ([0,1,2]: 0.75; [2,1,0]: 0.81; [1,2,0]: 0.80), with 96–98% of individual questions showing $\mathbf{u}/\mathbf{v} < 1$. Cohen's $d = 2.04$ ($p < 10^{-4}$, one-sample t-test against unit ratio) confirms few-shot's geometric effect is robust to exemplar sequence.

Table 8: Few-shot ordering ablation (N=100 questions, layer 14). All orderings preserve the low $\mathbf{u}/\mathbf{v}$-ratio pattern.

Order	$\|\Delta\mathbf{u}\|$	$\|\Delta\mathbf{v}\|$	$\mathbf{u}/\mathbf{v}$ ratio
[0,1,2]	0.031 ± 0.003	0.041 ± 0.004	0.75 ± 0.09
[2,1,0]	0.031 ± 0.003	0.038 ± 0.004	0.81 ± 0.11
[1,2,0]	0.032 ± 0.003	0.040 ± 0.004	0.80 ± 0.10

Reproducibility. All experiments use fixed random seeds.

G ADDITIONAL RESULTS

Curvature inheritance. Layer-to-layer curvature changes scale with the perturbation ratio $\eta_\ell = \max_t \|\Delta g_t^{(\ell)}\|/\|\Delta \mathbf{h}_t^{(\ell)}\|$: we observe $|\kappa^{(\ell+1)} - \kappa^{(\ell)}| \lesssim 4\eta_\ell$ across all layers (mean $|\Delta\kappa| = 0.03$, $\eta \approx 0.01$). This is an empirical regularity; a formal bound for L2-normalized trajectories would require controlling the normalization map (since $\widehat{\mathbf{a}+\mathbf{b}} \neq \hat{\mathbf{a}} + \hat{\mathbf{b}}$), which we do not attempt. Steps with near-zero displacement ($\|\Delta\mathbf{h}\| < 10^{-10}$) are excluded from curvature averages.

Magnitude–information dissociation. The $\mathbf{u}/\mathbf{v}$ projections capture only $\sim$4% of displacement norm ($\rho = 0.96$ unprojected), yet this small subspace separates prompting regimes and tracks the tested post-training contrast. The unprojected residual has comparable curvature ($\kappa_r/\kappa_{\text{full}} \approx 1.0$), while the $\mathbf{v}$-displacement fraction and early-layer curvature are only weakly correlated ($r = -0.12$, n.s.). Behaviorally discriminative signal is therefore present in a low-norm projected component, while most displacement and curvature-carrying motion remains outside it.

Full divergence table. Table 9 reports per-question divergence statistics for all 100 questions across conditions.

Table 9: Per-layer mean pairwise cosine distance (trajectory divergence) averaged across 100 questions and 3 paraphrases. Standard deviations in parentheses.

Condition	L7	L14	L21	L27
Direct	0.338 (0.124)	0.300 (0.106)	0.280 (0.102)	0.288 (0.095)
CoT	0.338 (0.124)	0.300 (0.106)	0.280 (0.101)	0.287 (0.094)
Few-shot	0.462 (0.103)	0.403 (0.088)	0.416 (0.092)	0.384 (0.086)
ReAct-style	0.425 (0.112)	0.374 (0.096)	0.352 (0.091)	0.380 (0.094)

Curvature raw means. Table 10 shows mean discrete curvature by layer and condition (full prompt trajectory).

Table 10: Mean discrete curvature κ by layer and condition (full prompt trajectory). Bold highlights the early-layer Few-shot differences discussed in the main text; statistical tests are reported in Table 11.

Layer	Direct	CoT	Few-shot	ReAct-style
7	1.698	1.705	**1.663**	1.711
14	1.519	1.526	**1.480**	1.536
21	1.472	1.478	1.525	1.490
27	1.406	1.389	1.455	1.490

Curvature statistical tests. Table 11 reports question-level paired statistics (Wilcoxon signed-rank), mean curvature difference $\bar{\Delta}\kappa$ (Direct minus comparison), and d_z for all layer $\times$ condition comparisons. Paraphrases are aggregated per question to avoid pseudo-replication. CoT shows large $|d_z|$ due to highly consistent small differences; the absolute effect ($|\bar{\Delta}\kappa| \leq 0.016$) is minimal.

Table 11: Curvature statistical tests vs. direct baseline (N=100 questions, 3 paraphrases aggregated per question, Wilcoxon signed-rank test).

Layer	Comparison	$\bar{\Delta}\kappa$	d_z	p
7	CoT	-0.007	-4.33	<0.001
7	Few-shot	$+0.035$	$+1.78$	<0.001
7	ReAct-style	-0.013	-1.43	<0.001
14	CoT	-0.007	-3.73	<0.001
14	Few-shot	$+0.039$	$+1.90$	<0.001
14	ReAct-style	-0.017	-1.80	<0.001
21	CoT	-0.006	-1.64	<0.001
21	Few-shot	-0.053	-1.31	<0.001
21	ReAct-style	-0.018	-0.75	<0.001
27	CoT	$+0.016$	$+1.50$	<0.001
27	Few-shot	-0.049	-0.67	<0.001
27	ReAct-style	-0.085	-1.75	<0.001

Experiment 4: Layer 27 displacement. At layer 27, $\mathbf{u}/\mathbf{v}$ ratios increase across all conditions:

Cross-scale geometric signatures. Figure 4 shows how divergence and curvature evolve across model scales (1.5B→7B→14B→32B→72B), with $\mathbf{u}/\mathbf{v}$ ratios at 1.5B, complementing Table 4.

Table 12: Projected displacement at layer 27.

Condition	$\|\Delta\mathbf{u}\|$	$\|\Delta\mathbf{v}\|$	$\mathbf{u}/\mathbf{v}$
Direct	0.010 ± 0.007	0.006 ± 0.003	1.73
CoT	0.070 ± 0.017	0.039 ± 0.009	1.79
Few-shot	0.065 ± 0.020	0.042 ± 0.010	1.54
ReAct-style	0.053 ± 0.025	0.033 ± 0.006	1.62

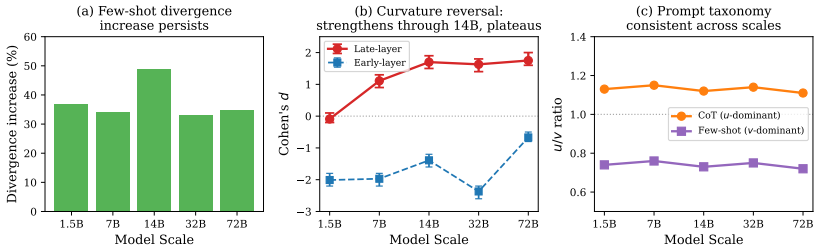


Figure 4: Geometric signatures across model scales (1.5B–72B). (a) Few-shot divergence increase persists across Qwen2.5 scales. (b) Full-trajectory curvature develops strong late-layer differences from 7B through 72B. Span-matched controls at 1.5B and 7B confirm that prompt-sensitive curvature persists after controlling the evaluated span, while its sign depends on depth and span definition. (c) $\mathbf{u}/\mathbf{v}$ ratios at 1.5B show the prompt-dependent ordering analyzed in the main text; cross-scale projection requires dimension-matched heads not trained here.

H Prefix vs. suffix position effects

I Long-context interference

Drift is sub-linear: the largest perturbation occurs from 0→3 passages, with diminishing marginal drift beyond 5. The representation remains perturbed even as marginal drift decreases.

Layer 27 shows the opposite curvature pattern, increasing with noise from 1.340 to 1.385, while earlier layers remain stable or decrease. Thus long-context interference has a depth-dependent geometric signature rather than producing uniform trajectory change across layers.

J Self-consistency and decision stability

Self-consistency operates by output averaging, not geometric smoothing: CoT generates $32\times$ more cross-sample endpoint dispersion than direct (0.554 vs. 0.017 at L14; dispersion measures variability across $n{=}8$ sampling runs, distinct from cross-paraphrase divergence which is identical for both, Table 1). Decision flip rate is lowest for direct (0.43), highest for few-shot (0.51); curvature predicts CoT stability ($r = -0.26$, $p{=}0.010$ at L27) but not few-shot, despite substantial geometric structure in the latter.

K Reflection curvature analysis

Reflection protocol. We use a two-pass reflection setup following Madaan et al. (2023). *Pass 1*: The model generates an initial answer using direct prompting ("Answer with just the final number") with greedy decoding. *Pass 2*: The model receives its Pass 1 output appended with the reflection prompt: "Review your answer step by step. If you find any errors, correct them. Give your final numerical answer." We extract hidden states at both passes and compare curvature. Of 200 questions, 174 were answered incorrectly in Pass 1; of these, 44.8% (78/174) were corrected in Pass 2, yielding final accuracy 47.5% (95/200). The $N{=}174$ subset used for AUROC analysis comprises all initially-wrong cases.

Gating protocol details. For the confidence baseline, confidence is defined as the log-probability of the final numeric answer token during generation. For retrospective gating evaluation, trajectories

Table 13: Question-endpoint trajectory divergence for Direct, suffix CoT, and prefix CoT. Position determines geometric influence.

Layer	Direct	CoT (suffix)	CoT (prefix)
7	0.348	0.348	0.397 (+14%)
14	0.308	0.308	0.343 (+12%)
21	0.286	0.286	0.307 (+7%)
27	0.285	0.284	0.279 (−2%)

Table 14: Last-token cosine drift vs. clean baseline by layer and number of irrelevant passages prepended.

Layer	0	1	3	5	7
7	0.000	0.005	0.011	0.017	0.019
14	0.000	0.004	0.008	0.011	0.013
21	0.000	0.004	0.007	0.009	0.010
27	0.000	0.004	0.007	0.008	0.009

with endpoint shift above threshold τ (L21: $\tau = 0.645$) are labeled "reconsider" rather than accepted. Because endpoint shift requires both passes, this is a post-reflection accept/abstain evaluation, not a pre-generation intervention. Threshold selection uses 5-fold stratified cross-validation (stratified by reflection outcome to ensure balanced folds), with no leakage between training and evaluation trajectories. AUROC is computed directly from endpoint shift over all 174 initially incorrect cases. The predictor uses only geometric signals (endpoint shift); no ground-truth labels are used at inference time.

L GEOMETRIC OVERLAP VS. FUNCTIONAL SEPARABILITY

We distinguish two senses of separability. Geometrically, the $\mathbf{u}$ and $\mathbf{v}$ projections overlap substantially with the top principal-variance directions of the hidden states (subspace angle $\approx 33^\circ$ to an SVD basis, i.e. substantially closer than orthogonal): the axes share high-variance geometric support. Functionally, however, they are distinguishable under the independent behavioral manipulations in §5.4, their update *directions* are near-orthogonal on actual displacements ($\theta_{\text{emp}} \approx 84^\circ$), and prompting versus training engage them asymmetrically. The $\mathbf{u}/\mathbf{v}$ projection is thus a *functional organization over shared geometric support*, not an orthogonal partition of representation space, consistent with reasoning and judgment being related but distinguishable.

We do not claim that the axes are uniquely identified. An independently constructed answer-choice/stage supervision objective changes their absolute coordinate realization but preserves the relative ordering $S_R > S_J$ (Appendix ad). Together with the probe-provenance results, this supports the distinction between coordinate realization and relational functional organization: the learned directions themselves can vary while the relative functional relation remains reproducible.

M BASE VS. INSTRUCT GEOMETRY

Alignment effects on the $\mathbf{u}/\mathbf{v}$ projection. Applying the fixed heads to both base and instruct hidden states (N=300 questions), we characterize both magnitude and directional changes under the fixed probe. *Magnitude:* at L7 the $\mathbf{u}/\mathbf{v}$ magnitude ratio shifts $0.70 \rightarrow 2.02$, driven by reduced $\mathbf{v}$ magnitude (t=−28.3, p<0.001); L14 shows a smaller ratio shift ($1.47 \rightarrow 1.66$) while L21 shows $\mathbf{v}$ decrease similar to L7. *Direction:* we measure the angle between mean $\mathbf{v}$-directions, $\theta = \arccos(\bar{\mathbf{v}}_{\text{base}} \cdot \bar{\mathbf{v}}_{\text{instruct}})$, with each $\bar{\mathbf{v}}$ a unit-normalized mean over N=300 questions (×3 paraphrases = 900 samples). This is a vector angle, not a sign-blind principal angle, so $\theta \in [0^\circ, 180^\circ]$. At L27, $\theta_{\mathbf{v}}$=96° and $\theta_{\mathbf{u}}$=35° (held-out means); the mean directions are concentrated (resultant length 0.92), so the rotation is a real property of the mean directions, not an averaging artifact.

The direction rotation is shared, not v-specific. The large $\mathbf{v}$ rotation relative to $\mathbf{u}$ (96° vs. 35°) is *not* evidence that alignment targets the judgment factor. An orthogonal Procrustes analysis with

Table 15: Mean curvature by layer and noise level. Curvature is stable or decreasing at L7–L21, but increases at L27.

Layer	0	1	3	5	7
7	1.689	1.627	1.636	1.633	1.634
14	1.494	1.425	1.425	1.422	1.427
21	1.470	1.357	1.341	1.318	1.321
27	1.340	1.354	1.375	1.376	1.385

Table 16: Reflection reduces curvature. $\Delta\kappa$ is the change from pass 1 to pass 2.

Layer	κ (Pass 1)	κ (Pass 2)	$\Delta\kappa$
7	1.700	1.629	-0.070
14	1.523	1.472	-0.052
21	1.512	1.420	-0.093
27	1.495	1.439	-0.056

5-fold held-out validation (Appendix n) shows that most of the angular gap is explained by a shared transformation: after removing R^*, the raw 61° **v**–**u** gap falls to 2.9° (95% reduction). The 96°/35° gap therefore reflects how the fixed $\mathbf{u}/\mathbf{v}$ readouts respond differently to shared representational reorientation, not a factor-specific rotation. Post-training's factor-*specific* effects are instead non-angular under the fixed probe: the **v** magnitude reduction above and the preserved behavioral dissociation, which shows a larger measured Γ on instruct (Appendix o).

Training stages: Base → SFT → DPO (pilot). As an exploratory pilot (N=25), we compare Mistral-7B (base), mistral-7b-sft-beta (SFT only), and Zephyr-7B-beta (SFT+DPO) checkpoints to characterize checkpoint-associated geometric signatures. Key findings: (1) The SFT checkpoint shows significantly higher curvature at all layers (+0.07 to +0.09, all $p<0.001$), contrary to a smoothing hypothesis. (2) The DPO checkpoint shows additional curvature at L16–L31 (+0.03 to +0.09, $p<0.05$), with the largest difference at L31 (+0.095). (3) Neither checkpoint transition significantly affects divergence. The stagewise curvature pattern (Base: 1.59–1.66 → SFT: 1.68–1.75 → DPO: 1.68–1.84) shows progressively higher curvature across these checkpoints, while the pilot design does not isolate the named training stages from other checkpoint differences.

Cross-family validation: Qwen vs. Mistral. To test which post-training signatures extend across model families, we compare Mistral-7B-v0.3 (base) vs. Mistral-7B-Instruct-v0.3 on N=50 GSM8K questions using the same geometric metrics.

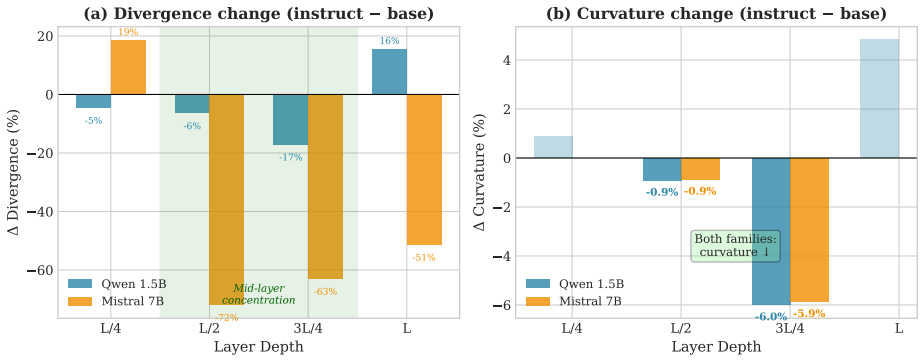


Figure 5: Cross-family training effects. (a) Divergence change shows mid-layer concentration in both families, with Mistral exhibiting larger effects. (b) Curvature decreases at 3L/4 for both Qwen (-6.0%) and Mistral (-5.9%).

Key findings:

Table 17: Base vs. instruct model geometry. The instruct checkpoint shows reduced divergence and curvature at intermediate layers. Instruct curvature values are from a separate extraction run and differ from Table 10 by ≤ 0.04.

	Divergence		Curvature	
Layer	Base	Instruct	Base	Instruct
7	0.354	0.338	1.694	1.709
14	0.321	0.300	1.544	1.530
21	0.338	0.280	1.601	1.505
27	0.249	0.288	1.365	1.431

- *Mid-layer concentration replicates*: Both families show the largest divergence reduction at intermediate depths (Qwen L21: -17%; Mistral L/2: -72%, 3L/4: -63%).
- *Curvature effects align*: Curvature decreases at 3L/4 with nearly identical magnitude (Qwen: -6.0%; Mistral: -5.9%).
- *Magnitude differences*: Mistral shows larger divergence effects overall, potentially reflecting differences in base model calibration or training data.

The Mistral result provides a second-family instance of the mid-depth pattern, while effect magnitudes remain architecture-dependent.

Comparison with CKA-based similarity. Base–instruct CKA similarity decreases with depth, independently indicating stronger representational change at later layers under this metric.

Rotation metric (code).

```
v_base = normalize(base_v_dirs.mean(0))  # in R^64
v_inst = normalize(inst_v_dirs.mean(0))
v_cos  = float(v_base @ v_inst)          # -0.10 at L27 => 96 deg
angle  = degrees(arccos(clip(v_cos,-1,1)))  # [0,180]
```

N SHARED VS. FACTOR-SPECIFIC REORIENTATION: PROCRUSTES ANALYSIS

The mean $\mathbf{v}$-direction rotates 96° under alignment while $\mathbf{u}$ rotates 35° (§5.6). A natural question is whether this gap reflects a $\mathbf{v}$-specific reorientation or a shared rotation of the representation space that the $\mathbf{u}/\mathbf{v}$ readouts happen to sample differently. We resolve this directly with an orthogonal Procrustes analysis.

Method. Using paired base/instruct hidden states at L27 over the same N inputs ($H_{\text{base}}, H_{\text{inst}} \in \mathbb{R}^{N\times H}$), we compute the optimal shared orthogonal transformation

$$R^* = \arg\min_{R:\, R^\top R = I} \|H_{\text{base}} R - H_{\text{inst}}\|_F = UV^\top, \quad H_{\text{base}}^\top H_{\text{inst}} = U\Sigma V^\top. \tag{12}$$

R^* is estimated in hidden space ($\mathbb{R}^{1536}$) and applied to base states before projection; residual angles are measured in the 64-dimensional projected space. Because P is a nonlinear MLP, we project per-state and then average: $\theta_{\text{res}} = \angle\big(\overline{P(R^* h_{\text{base}})}, \overline{P(h_{\text{inst}})}\big)$ for each projection $P \in \{P_{\mathbf{u}}, P_{\mathbf{v}}\}$, where $\bar{\cdot}$ denotes the sample mean.

The reorientation is shared (held-out validation). We validate this with 5-fold cross-validation over $N{=}300$ questions (three paraphrases each), splitting folds by question: R^* is fit on train questions and residual angles are evaluated on held-out questions. This prevents in-sample overfitting given that $N \ll H$ ($300 \ll 1536$). We also test low-rank Procrustes (ranks 16–256) to check robustness.

The raw $\mathbf{v}$–$\mathbf{u}$ gap is 60.8°; after removing the shared transformation the residual gap is 2.9° (95% reduction). Thus large coordinate movement can coexist with preserved functional organization: most of the apparent factor-specific rotation is explained by a shared change of representation frame. Across low-rank settings (16–256) the residual gap remains 0.7–2.9°, confirming the result is not

Table 18: Held-out Procrustes validation (5-fold CV, N=300). The raw $60.8°$ rotation gap falls to $2.9°$ after shared orthogonal alignment.

Direction	Raw rotation	Residual (after R^*)
$\mathbf{v}$	$95.7°$	$8.8°$
$\mathbf{u}$	$34.9°$	$5.9°$
Random (mean of 10)	–	$4.9°$

driven by overfitting. The raw $96°/35°$ gap reflects how the fixed $\mathbf{u}/\mathbf{v}$ readouts sample the shared transformation differently; it is a property of projection geometry, not a $\mathbf{v}$-specific training effect. Base–instruct CKA in the $\mathbf{u}$- and $\mathbf{v}$-projected spaces rises from $0.02/0.20$ to $0.67/0.69$ after aligning base states by R^*; on the raw hidden states CKA is unchanged (0.535), as it must be under an orthogonal map.

Consequence. The rotation magnitude is descriptive of post-training's large final-layer reorientation but is not evidence for a $\mathbf{v}$-specific effect. The functional interpretation of $\mathbf{v}$ rests instead on independent behavioral validation (§5.4, Appendix ad) and its persistence under post-training (Appendix o), neither of which depends on interpreting the raw direction angle as function-specific.

Sample efficiency. This alignment is likewise sample-efficient: with as few as 10 training pairs, the residual-gap CI upper bound stays below $14°$, versus a raw gap of $\sim 65°$ (Appendix u).

o Behavioral dissociation under post-training

The behavioral validation of §5.4 establishes the reasoning/judgment semantics of the fixed $\mathbf{u}/\mathbf{v}$ readout. We next ask whether this functional organization differs between the base and instruction-tuned checkpoints. Using the same fixed projection and the same stimuli, we repeat the reasoning/judgment manipulations on both models.

The fixed readout remains behaviorally informative across checkpoints. The same projection reads structured prompt-conditioned differences in both base and instruct models. At L27, the prompt-conditioned $\mathbf{u}/\mathbf{v}$ ordering is preserved across checkpoints (Spearman ρ=0.80), Few-shot retains the lowest $\mathbf{u}/\mathbf{v}$ ratio, and Few-shot vs. CoT separates strongly (d=1.05, $p<10^{-4}$).

The dissociation is preserved, with larger measured Γ on instruct. Running the full contrast on both checkpoints (N=200 questions per checkpoint, identical design and metric to Appendix ad):

Table 19: Behavioral dissociation on base vs. instruct models (same fixed projection, same stimuli, N=200). The signed dissociation is preserved, with a larger measured Γ on the instruct checkpoint.

Statistic	Base	Instruct
S_R (reasoning)	$+0.96$ $[0.91, 1.01]$	$+1.24$ $[1.21, 1.28]$
S_J (judgment)	-0.36 $[-0.40, -0.33]$	-0.34 $[-0.39, -0.29]$
$\Gamma = S_R - S_J$	$+1.33$ $[1.26, 1.39]$	$+1.59$ $[1.53, 1.64]$

The signed dissociation holds ($S_R>0$, $S_J<0$) and Γ increases from 1.33 to 1.59 (non-overlapping CIs). Both functional effects continue to exceed their matched lexical controls (reasoning excess $+0.65$, judgment excess $+0.13$; both $p<10^{-4}$), so the increase is not a lexical artifact.

The increase is not explained by a uniform baseline shift. The instruct model has an overall $+0.11$ shift toward $\mathbf{u}$-weighting (measured on the un-manipulated *direct* baseline), consistent with reduced $\|\mathbf{v}\|$ (Table 3). Subtracting this baseline shift, the reasoning manipulation becomes more $\mathbf{u}$-engaging ($+0.17$ net) and the judgment manipulation more $\mathbf{v}$-engaging (-0.09 net): the two contrasts move in opposite directions. Thus the larger measured Γ on instruct is not explained by a uniform magnitude rescaling, although the fixed probe was trained on the instruct checkpoint and we therefore do not interpret the increase alone as evidence that post-training strengthens the dissociation.

Consequence. The reasoning/judgment dissociation is not an artifact of the base model: its behavioral signature is preserved, with a larger measured Γ on the instruct checkpoint under this fixed instruct-trained probe, and this holds under the same fixed projection whose raw direction rotation is largely explained by shared reorientation (Appendix n). Post-training's factor-differentiated effects are therefore supported behaviorally and through magnitude under the fixed probe, whereas the raw direction rotation is largely explained by shared frame reorientation.

P SYMMETRIC PROBE-PROVENANCE EVALUATION

To separate checkpoint effects from probe-training provenance, we independently train matched JM heads on Base and Instruct trajectories using the same Phase-1 procedure (architecture, objective, episode structure, and training hyperparameters), and evaluate all four probe-source $\times$ evaluation-checkpoint combinations using the same operational reasoning and judgment contrasts. These heads are trained specifically for the symmetric provenance experiment and are distinct from the canonical fixed head used in Appendix o.

Because the independently trained probe frames differ substantially in absolute coordinate realization, we treat the sign and magnitude of individual S_R and S_J values as probe-dependent. We therefore test the relative separation $\Gamma = S_R - S_J$ as the candidate provenance-robust quantity.

Table 20: Symmetric probe-provenance evaluation at layer 27 (N=200). JM heads are trained independently on Base or Instruct trajectories using the same Phase-1 procedure, then evaluated with identical factorial behavioral contrasts.

Probe source	Evaluation	S_R	S_J	Γ
Base	Base	-3.34	-3.93	0.59
Base	Instruct	-1.99	-2.16	0.17
Instruct	Base	-0.29	-1.48	1.19
Instruct	Instruct	0.39	-1.56	1.95

All four Γ estimates are positive with bootstrap 95% CIs excluding zero: $[0.56, 0.62]$, $[0.13, 0.21]$, $[1.14, 1.23]$, and $[1.92, 1.99]$, respectively. Only the Instruct-probe/Instruct-evaluation condition recovers the primary signed pattern $S_R{>}0{>}S_J$; the other probe frames shift the absolute coordinates while preserving $S_R{>}S_J$. Thus probe provenance strongly affects coordinate realization but does not explain the relative behavioral dissociation.

We note that the Base-trained probe assigns strongly negative S_R (-3.34), i.e. under that coordinate frame the reasoning contrast is $\mathbf{v}$-weighted; the "reasoning-associated" and "judgment-associated" labels are therefore properties of the primary probe frame, not of any probe trained with this objective. This result has two implications. First, the $\Gamma{>}0$ finding is not an artifact of Instruct-trained probe bias: a probe trained entirely on Base trajectories still detects the dissociation on the Base checkpoint (Γ=0.59). Second, the substantial variation in absolute S_R and S_J across probe frames—spanning from -3.34 to $+0.39$—reinforces that the stable object is the *relative* functional relation rather than any particular signed coordinate realization.

Q REFLECTION GATING BY OUTCOME

Table 21: Endpoint shift (L21) by reflection outcome (N=200 questions, 174 initially wrong). The table reports L21 values; optimal thresholds τ below are layer-specific.

Outcome	n	Mean shift	Std
Correct $\rightarrow$ correct	17	0.251	0.293
Wrong $\rightarrow$ correct	78	0.088	0.169
Wrong $\rightarrow$ wrong	96	0.495	0.298
Correct $\rightarrow$ wrong	9	0.631	0.216

The optimal gate threshold varies with depth: L7: $\tau = 0.785$, L14: $\tau = 0.721$, L21: $\tau = 0.645$, L27: $\tau = 0.955$. At each layer τ falls near the median endpoint shift of the uncorrected group (L21: 0.646; L27: 0.973), i.e. the gate is calibrated to the uncorrected distribution rather than to an absolute shift magnitude; the layer-to-layer variation is therefore consistent with depth-dependent scaling of endpoint shift rather than a fixed absolute shift threshold.

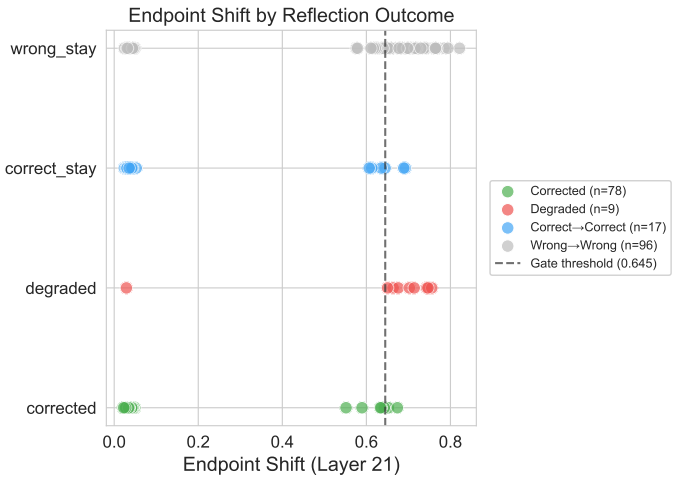


Figure 6: Endpoint shift distributions (L21) by reflection outcome (N=200, 174 initially wrong). Corrected trajectories (wrong→correct) show significantly smaller shift than uncorrected ones (wrong→wrong), directly visualizing the separation that yields AUROC=0.85.

R TOKEN BUDGET SCALING

Table 22: Accuracy and L21 curvature by prompt condition and token budget.

Condition		64	128	256	512
Direct	Acc.	0.12	0.12	0.12	0.12
	κ_{21}	1.512	1.512	1.512	1.512
CoT	Acc.	0.12	0.00	0.56	**0.80**
	κ_{21}	1.429	1.391	1.360	1.354
ReAct-style	Acc.	0.16	0.56	**0.76**	0.76
	κ_{21}	1.445	1.423	1.418	1.418

S MULTI-ROUND REFLECTION

Table 23: Multi-round reflection: accuracy and L21 curvature by round (N=25 pilot).

	Round 0	Round 1	Round 2	Round 3
Acc. (ungated)	0.12	0.52	0.52	0.36
Acc. (gated)	0.12	**0.56**	**0.56**	0.40
κ_{21}	1.512	1.420	1.393	1.383
Mean shift	n/a	0.262	0.335	0.380

Curvature monotonically decreases across rounds regardless of accuracy changes. Ungated accuracy degrades at round 3 (0.52 → 0.36) as correct answers are overwritten. Gating partially mitigates this (0.40 vs. 0.36), but effectiveness decreases as mean endpoint shift increases (0.262 → 0.380).

T INTERVENTION AND TEMPORAL STABILITY CONTROLS

Two intervention methods and the random null. Our initial intervention uses a fixed linear surrogate of the nonlinear projection head, omitting the state-dependent derivatives induced by its GELU nonlinearities. We add $\alpha\mathbf{d}$ at generation start, where α is set as a fraction of $\|\mathbf{h}\|$ (calibrated perturbation). Table 24 shows selection rates (percentage of samples where the model output changed) across layers, dose fractions, and directions (N=50). This intervention does not outperform random controls. The local-Jacobian intervention in §5.5 maps the same behaviorally anchored factor through the exact derivative of the nonlinear readout and yields $|\Delta_v|$=0.37 vs. $|\Delta_u|$=0.06 ($p<10^{-30}$). Although the paired v–u difference is precisely estimated, its magnitude is not unusual relative to arbitrary direction pairs: $|\Delta_u|$ and $|\Delta_v|$ lie at the 20th and 45th percentiles of the random-direction null, respectively, and their absolute gap lies at the 57th percentile of the random-pair-gap null.

Jacobian steering protocol. *Data split*: N=100 direction-construction, N=50 pilot, N=200 held-out confirmatory. *Target direction*: $d_v = \mathbb{E}[\phi_v(h_{J+}) - \phi_v(h_{J-})]$ from the construction set; control direction $d_u = \mathbb{E}[\phi_u(h_{R+}) - \phi_u(h_{R-})]$. *Intervention*: for hidden state h, compute exact Jacobian $J_v(h) = \partial\phi_v(h)/\partial h$ (including GELU derivatives), gradient $g_v(h) = J_v(h)^\top d_v$, normalized $\hat{g}_v(h) = g_v(h)/\|g_v(h)\|$, and intervene $h' = h \pm \alpha\|h\|\hat{g}_v(h)$. *Pilot*: layer $\in \{14, 21, 27\}$, $\alpha \in \{0.0025, 0.005, 0.01, 0.02, 0.05\}$; selected L27, α=0.05 before confirmatory evaluation. *Conditions*: $\pm v$, $\pm u$, $\pm$random (norm-matched), paired on each example. *Endpoint*: decision-logit contrast $L = \text{logit}(\texttt{COMMIT}) - \text{logit}(\texttt{WITHHOLD})$ prior to generation; $\Delta_v = L(+v) - L(-v)$. *Manipulation check*: $\langle \phi_v(h') - \phi_v(h), d_v \rangle > 0$ for $+v$, < 0 for $-v$ (both confirmed). *Success criteria*: (1) bidirectional v effect with 95% CI excluding 0; (2) $|\Delta_v| > |\Delta_u|$; (3) manipulation check passes; (4) the v–u gap exceeds the random-pair null. Criteria (1)–(3) are met; criterion (4) is not ($|\Delta_u|$ at 20th percentile, $|\Delta_v|$ at 45th percentile, gap at 57th percentile of random-pair gaps).

Table 24: Calibrated causal intervention grid. Entries show the percentage of samples with changed output. No consistent direction-specific advantage over norm-matched random perturbations is observed.

Layer	$\|\Delta\mathbf{h}\|/\|\mathbf{h}\|$	$\mathbf{u}+$	$\mathbf{u}-$	$\mathbf{v}+$	$\mathbf{v}-$	random
L14	2%	22%	26%	18%	22%	18%
	5%	26%	20%	28%	26%	26%
	10%	28%	20%	22%	30%	20%
	20%	38%	34%	44%	34%	30%
L27	2%	0%	2%	0%	2%	0%
	5%	0%	2%	0%	2%	0%
	10%	2%	2%	0%	2%	2%
	20%	16%	2%	2%	2%	4%

L14: all directions change generation at similar rates ($\mathbf{u}$ 26.7%, $\mathbf{v}$ 28.0%, random 23.5% averaged). L27: most conditions change generation in only 0–2% of samples, with one $\mathbf{u}+$ condition reaching 16% at the largest dose. No consistent direction-specific effect is observed.

Table 25: Temporal stability of projection heads at layer 14. Cos gap = inter-entity cosine gap.

Condition	cos gap	R (step/para)	Interpretation
Factorized $\mathbf{v}$ (D)	0.45	**1.07**	entity-stable
Unfactored z (F)	1.06	1.53	step-leaking
Shuffled labels (E)	0.00	n/a	no structure

Setup. We compare three training conditions for the projection heads using identity-tracking episodes: (D) *Factorized*: the $\mathbf{u}/\mathbf{v}$ projection with trained projection loss; (F) *Unfactored*: a single projection without the split; (E) *Shuffled*: factorized architecture with permuted labels. We measure (1) inter-entity cosine gap and (2) step-drift ratio $R = d_{\text{step}}/d_{\text{para}}$, where $R \approx 1$ indicates temporal stability.

Findings. The factorized $\mathbf{v}$-space achieves near-perfect temporal invariance (R=1.07), while the unfactored baseline leaks step information (R=1.53). Shuffled-label training yields zero inter-entity separation despite identical architecture (cos gap 0.00), confirming that correct supervision is necessary for the structured entity-identity factor. The unfactored condition attains a larger raw inter-entity gap than the factorized one (1.06 vs. 0.45), so the factorized projection's advantage is temporal stability rather than raw separation. On the synthetic identity-tracking task, both factorized and unfactored conditions show high causal selectivity (100–500$\times$ larger displacement for identity-changing perturbations), but only the factorized projection maintains this selectivity across reasoning steps.

Cross-seed stability. To validate that the $\mathbf{u}/\mathbf{v}$ projection is robust to initialization, we trained four independently initialized projection heads (the primary head and three additional seeds) on the same entity-tracking episodes. Across the primary head and three additional seeded heads, downstream clustering (NMI) is perfectly stable (1.000 ± 0.000), and the barrier gap shows low variance (1.034 ± 0.024, CV $= 2.4\%$). Raw weight matrices differ substantially across seeds despite equivalent pretext performance, indicating non-identifiability of the learned coordinates.

Cross-seed reproducibility of prompt ordering. We evaluated the CoT–Few-shot ordering with four independently trained JM projection heads using the same pretext objective. Three of four heads recover the same CoT $>$ Few-shot $\mathbf{u}/\mathbf{v}$ ordering (all $p \leq 0.01$). One seed reverses the ordering, indicating that the prompt taxonomy is reproducible across most probe realizations but not invariant to initialization. This is consistent with the broader probe analysis: relative behavioral dissociation is more stable than any particular coordinate realization or prompt-level ordering.

Behavioral dissociation across heads. We repeat the §5.4 contrast on the primary head (used throughout) and three additional seeded heads.

Table 26: Behavioral dissociation (N=200) on the primary head and three seeded heads.

Head	S_R	S_J	Γ [95% CI]
Primary	$+0.96$	-0.36	1.33 [1.26, 1.39]
Seed 42	$+0.36$	$+0.16$	0.20 [0.14, 0.25]
Seed 123	-0.17	-0.59	0.42 [0.37, 0.47]
Seed 456	$+0.42$	-0.04	0.46 [0.42, 0.50]

Γ>0 on all heads; the signed pattern (S_R>0, S_J<0) holds on the primary head and seed 456.

Subspace identifiability. Aligning each pair of seeded heads by orthogonal Procrustes, the residual Frobenius distance is 0.19 (mean) / 0.22 (max) for $\mathbf{u}$ and 0.27 / 0.28 for $\mathbf{v}$, against 1.03 ± 0.004 for random 64-D subspace pairs.

Projection coverage and interpretation. Across layers and questions the $\mathbf{u}$ and $\mathbf{v}$ update directions are near-orthogonal in output space (mean angle $\theta_{\text{emp}} \approx 84°$, max cross-coupling $\varepsilon_{\text{emp}} \approx 0.55$). Per-layer values are consistent: $\theta_{\text{emp}} \in [83.6°, 84.5°]$, $\varepsilon_{\text{emp}} \in [0.43, 0.55]$. The projections capture only $\sim$4% of displacement norm (median $\rho = 0.96$), and the unprojected residual has curvature comparable to the full trajectory ($\kappa(r)/\kappa(h) \approx 1$). We tested whether higher $\mathbf{v}$-displacement fraction correlates with lower early-layer curvature: the correlation is weak and not significant (Pearson $r = -0.12$, $p = 0.09$). The projections therefore function as empirical probes that capture behaviorally meaningful structure (dissociated by prompt condition, differentially affected by training, weakly coordinated with curvature) rather than as a curvature decomposition. The relative $\mathbf{u}/\mathbf{v}$ orderings are therefore probe-based observations rather than a decomposition of total trajectory geometry.

Random projection ablation. To test whether the $\mathbf{u}/\mathbf{v}$ taxonomy reflects meaningful learned structure (rather than arbitrary projection directions), we compare trained projection heads against 10 randomly-initialized (untrained) projections. We measure the Cohen's d for the CoT vs. few-shot $\mathbf{u}/\mathbf{v}$ displacement-ratio difference. At layer 27, the trained head yields $d = 1.10$ (CoT $>$ few-shot),

while **80% of random projections show the opposite pattern** (few-shot > CoT). This indicates that the trained taxonomy captures a specific geometric structure that random projections would typically miss or reverse. At layer 14, the trained head shows $d = 2.34$ with signal-to-noise ratio 0.56 relative to random variation. The consistency of the trained pattern (versus the high variance of random projections) supports the interpretation that the $\mathbf{u}/\mathbf{v}$ projection encodes meaningful condition-specific structure rather than arbitrary subspace partitioning.

Linear-surrogate intervention procedure. To test a coarse global back-projection baseline, we perform activation perturbations at L14 and L27 using a fixed linear surrogate of the nonlinear head. *Implementation*: During the forward pass, we register a hook at the target layer that modifies the hidden state at the *last token position* (generation-start) by adding $\alpha \cdot \mathbf{d}$. The direction $\mathbf{d}$ is defined as follows: for N=50 questions, collect the hidden state $\mathbf{h}$ at generation-start, project to get $(\mathbf{u}, \mathbf{v}) = \text{head}(\mathbf{h})$, then $\mathbf{d}_\mathbf{u} = \text{normalize}(\text{mean}(\mathbf{u}))$ and $\mathbf{d}_\mathbf{v} = \text{normalize}(\text{mean}(\mathbf{v}))$ in the 64-dimensional projected space. To add $\alpha\mathbf{d}$ back to the hidden state, we use an approximate back-projection P_f^T for factor $f \in \{\mathbf{u}, \mathbf{v}\}$, where $P_f = B_f W_2 W_1 W_0$ with $B_f \in \mathbb{R}^{64\times 128}$ selecting the factor-specific 64-dimensional block of the 128-dimensional head output and intermediate GELU nonlinearities omitted, giving $\Delta\mathbf{h} = \alpha(P_f^\mathsf{T}\mathbf{d}_f)$. This is a linear surrogate for back-projection rather than the exact Jacobian; we use it only to obtain a hidden-space direction of approximately correct scale. We test $\alpha \in \{0.02, 0.05, 0.10, 0.20\} \times \|\mathbf{h}\|$ (calibrated to each sample's hidden-state norm). A norm-matched random direction serves as control. No re-normalization is applied after perturbation; the perturbed state propagates through remaining layers. We generate up to 256 tokens with temperature $= 0$ and compare outputs to unperturbed baselines. Results are shown in Table 24. This linear surrogate does not yield consistent effects; local-Jacobian steering (§5.5) confirms that neither the linear nor the exact method produces factor-specific effects beyond random directions.

U SAMPLE COMPLEXITY OF RELATIONAL IDENTIFICATION

Two experiments show that the relational organization is both reproducible and sample-efficient.

Γ sample complexity. Using cached C1/C2 hidden states (N=200 full), we subsample at $N \in \{10, 20, 30, 50, 75, 100, 150\}$, drawing 1000 random subsamples without replacement per N and recording the subsample Γ; the table reports the median and 2.5–97.5% range across subsamples, and the fraction of subsamples with Γ>0.

Table 27: Γ sample complexity. CI is across subsamples drawn without replacement; the full-sample bootstrap CI is reported in Appendix ad.

N	Median Γ	2.5–97.5% range	Frac Γ>0
10	1.33	[1.03, 1.60]	100%
20	1.33	[1.14, 1.51]	100%
30	1.32	[1.17, 1.47]	100%
50	1.33	[1.22, 1.43]	100%
75	1.33	[1.25, 1.41]	100%
100	1.33	[1.26, 1.39]	100%
150	1.33	[1.29, 1.36]	100%

The relational ordering is positive in 100% of subsamples even at N=10, showing that the primary effect is recoverable from small calibration sets.

Procrustes sample complexity. Using paired base/instruct hidden states (N=100), we fit Procrustes on $N_{\text{train}} \in \{10, 20, 30, 50, 75\}$ and measure residual angles on held-out test data.

Full baseline (N=100): raw gap = 64.7°, residual gap = 0.35°. The residual-gap CI upper bound stays below 14° at every tested N_{train} (including N=10), versus raw gap ∼65°—most of the gap is explained by shared reorientation even with few training pairs.

Table 28: Procrustes sample complexity (N=100 paired states, separate from main N=300 experiment). Residual-gap CI upper bound stays below 14° at every N_{train}, versus raw gap $\sim 65^\circ$.

N_{train}	Residual Gap	95% CI	Raw Gap
10	2.1°	$[-18.0, +13.3]$	65.0°
20	2.3°	$[-9.3, +9.8]$	64.7°
30	0.4°	$[-15.6, +7.6]$	64.5°
50	1.2°	$[-9.4, +8.3]$	64.8°
75	1.7°	$[-8.1, +8.9]$	63.9°

Table 29: Predicting reflection success from geometric signatures (N=174 initially-wrong cases).

Metric	Value	Note
AUROC (L21 shift)	**0.850**	[0.79, 0.91] at N=174
Correction rate	44.8%	(78/174)
Mean shift (will correct)	0.088	± 0.17
Mean shift (won't correct)	0.495	± 0.30
Cohen's d	1.64	n/a

V PREDICTING REFLECTION SUCCESS

Using 200 reflection trials, we test whether endpoint shift at L21 predicts whether reflection will correct an initially-wrong answer. For the 174 initially-wrong cases (78 corrected, 96 uncorrected; 44.8% correction rate), lower endpoint shift strongly predicts successful correction. AUROC $= 0.850$ at L21. Layer-wise: L7: 0.825, L14: 0.847, L21: 0.850, L27: 0.830.

Baseline comparisons. To test whether endpoint shift captures information beyond generic scalar summaries, we evaluate all predictors on the same 174 initially-wrong reflection cases. Endpoint shift achieves AUROC $= 0.850$, substantially exceeding predictive entropy (0.553), sequence perplexity (0.534), mean answer-token log-probability (0.534), response-length change (0.520), curvature change (0.511), and initial curvature (0.408). Thus correction-relevant information is not recovered by the tested output-confidence, trajectory-shape, or length summaries; it is captured substantially more strongly by the geometric relation between the pre- and post-reflection endpoint representations.

Full-sample generalization. The analysis above conditions on initially-wrong cases, which would not be known at deployment time. Evaluating on all 200 cases (predicting whether pass-2 is correct, regardless of pass-1 outcome), AUROC remains high: L7: 0.802, L14: 0.838, L21: 0.837, L27: 0.819. The modest drop from 0.850 to 0.837 at L21 shows that the geometric signal persists without conditioning on first-pass correctness.

Cross-task monitoring boundary. A predictor defined on GSM8K endpoint shift transfers only weakly to CommonsenseQA (AUROC 0.607, 95% CI $[0.308, 0.857]$). Interpretation is limited by the small number of initially incorrect examples (n=19) and by a qualitative change in the reflection regime: whereas reflection predominantly corrects errors on GSM8K, on CommonsenseQA it produces 50 degradations but only 7 corrections. Unsupervised Procrustes alignment of the task-specific representation spaces changes AUROC only from 0.607 to 0.619, indicating that simple coordinate alignment does not resolve the cross-task degradation. We therefore treat cross-task monitoring as unresolved rather than claiming task-general transfer.

W PROJECTION GENERALIZATION AND BASELINES

Generalization test. To rule out memorization, we conduct a strict evaluation with entity-disjoint train/test splits: entities seen during probe training never appear during evaluation. Under this protocol, NMI remains 1.0 at all layers, confirming that the probes learn the *geometric structure* of entity identity rather than memorizing specific entity embeddings.

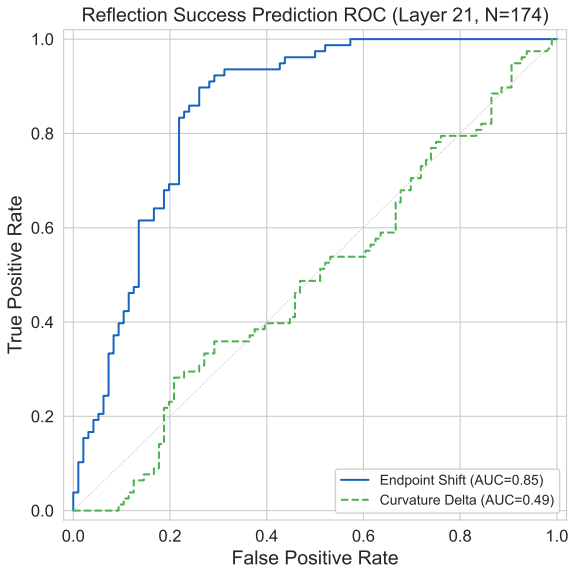


Figure 7: ROC curve for reflection success prediction using endpoint shift at layer 21 (N=174 initially-wrong cases; AUROC $= 0.85$).

Surface-feature baseline. A TF-IDF classifier on raw token text achieves 100% leaf-type accuracy (since task schemas are semantically distinct), but produces zero structural metrics: Spearman $\rho = 0$ for within-leaf ordering and barrier gap $= 0$. This shows that surface-level task identification alone does not recover the continuous ordering captured by the trained projection: despite perfect leaf-type classification, the TF-IDF baseline has zero within-leaf rank correlation and barrier gap.

X SYNTHETIC VALIDATION OF PROJECTION

We validate the projection structure on a synthetic multi-entity identity-tracking task where ground-truth leaf structure is known by construction; generalization to GSM8K is evaluated through independent behavioral validation and random-projection controls; local-Jacobian interventions separately test whether the behaviorally identified factors are selectively steerable (Appendix t). Each episode contains K=4 entities undergoing state transitions (e.g., location changes). Hidden states are extracted from Qwen2.5-1.5B-Instruct processing these episodes. The trained projection achieves perfect separation: NMI $= 1.000$, ARI $= 1.000$ for $\mathbf{v}$-clustering by entity identity, and Spearman $\rho = 0.955$ for $\mathbf{u}$-rank ordering within leaves. The barrier gap between same-entity and cross-entity $\mathbf{v}$-distances is 1.022 (normalized units). Ablation confirms that $\mathcal{L}_{\text{same}}$ is the most critical component for clustering quality, while $\mathcal{L}_{\text{order}}$ is essential for within-leaf ranking.

Y STATISTICAL ROBUSTNESS ANALYSES

Multiple comparison correction. For predefined families of multi-layer prompt-condition comparisons (divergence: 4 layers $\times$ 3 conditions; curvature: 4 layers $\times$ 3 conditions), we control false discovery rate using Benjamini-Hochberg at $\alpha = 0.05$. All curvature comparisons in Table 11 remain significant after correction (all raw $p < 0.001$); the span-matched late-layer effect (d= $+$ 0.43) and reflection prediction (AUROC $= 0.85$) are reported with their experiment-specific tests. Other analyses (intervention and random-null tests, Spearman correlations, and SFT/DPO comparisons) report their own statistical procedures as described in the corresponding sections.

Paraphrase semantic equivalence. Using sentence-transformer embeddings (MiniLM-L6-v2), we validate that rule-based paraphrases preserve semantic content. Across 150 paraphrase pairs: mean cosine similarity $= 0.986 \pm 0.017$, 100% above the 0.8 equivalence threshold, and 100% above the 0.9 high-quality threshold. The lowest-scoring paraphrase still achieves 0.902 similarity.

Paraphrase method robustness. Core findings replicate with GPT-4o-mini generated paraphrases (N=100, temperature 0.7): few-shot divergence increase persists across the tested scales (1.5B: +19–31%; 7B: +12–17%; 14B: +11–19%), suffix invariance holds ($\Delta < 0.001$), and CommonsenseQA shows stronger effect (+66–119%). Paraphrase quality remains high (mean MiniLM similarity 0.961 ± 0.035; 100% above the 0.8 equivalence threshold).

Span matching at 7B. Repeating the question-span control on Qwen2.5-7B preserves substantial prompt-sensitive curvature at mid-to-late depth (L20: d=−1.47, L27: d=−0.56), although the sign differs from the corresponding full-trajectory effect. Thus the larger-scale curvature signature is not eliminated by span matching, while its direction remains dependent on the evaluated trajectory span.

Turning-angle curvature. Replacing the second-difference formula with turning-angle curvature (angle between consecutive velocity vectors, direction-only) preserves substantial depth-dependent structure:

Comparison	L7	L14	L21	L27
Second-difference d (main text):				
CoT vs Direct	−4.3	−3.7	−1.6	+1.5
Few-shot vs Direct	+1.8	+1.9	−1.3	−0.7
Turning-angle d:				
CoT vs Direct	−4.3	−3.4	−3.7	−2.5
Few-shot vs Direct	+0.9	+1.5	−0.3	−0.3

The alternative metric preserves substantial depth structure but not every layerwise effect. Most notably, few-shot prompting shows the same transition from positive early-layer effects to negative later-layer effects under both metrics, whereas the final-layer CoT effect changes sign (+1.5 vs −2.5). We therefore treat depth-dependent curvature structure as robust while regarding individual layerwise effects as metric-dependent.

Normalization robustness. Normalization preserves the curvature contrast through early and mid depth (L7–L21), with consistent signs under raw and L2-normalized representations. At the final layer, however, the effect is normalization-sensitive: the raw-state contrast is negligible (d=0.02), whereas the normalized representation yields a moderate negative effect (d=−0.57). Thus the mid-depth curvature pattern is robust, while final-layer curvature should be interpreted as metric-dependent.

Z CROSS-ARCHITECTURE PILOT: TINYLLAMA

As a pilot replication, we test core geometric signatures on TinyLlama-1.1B-Chat (N=25 questions, layers 5/11/16/21 ≈ 25%/50%/75%/95% depth). This small-N pilot provides a qualitative cross-architecture stress test; larger-N validation is left for future work.

Curvature remains prompt-sensitive. Few-shot shows lower curvature than direct at all layers: L5 (1.86 vs 1.88), L11 (1.65 vs 1.67), L16 (1.62 vs 1.64), L21 (1.79 vs 1.85). This shows curvature sensitivity to prompting in this pilot, though the sign differs from Qwen's span-controlled late-layer effect.

Divergence patterns differ. Unlike Qwen, TinyLlama shows *lower* divergence for few-shot (0.004) than direct (0.023) at late layers. This reversal shows that the sign of the divergence response is model-family dependent in this pilot.

Implications. Curvature and divergence are both prompt-sensitive in this pilot, while their signs and magnitudes differ from the Qwen results.

AA CROSS-ARCHITECTURE PILOT (GEMMA-2B)

To further test cross-architecture transfer beyond Llama-family models, we replicate divergence experiments on Gemma-2B-it (Gemma Team et al., 2024) (N=15 questions, layers 4/9/13/17 $\approx$ 22%/50%/72%/94% depth across 18 layers).

Architecture-dependent divergence effects. Table 30 shows trajectory divergence by layer and condition. CoT reduces divergence relative to direct by -8.9% on average. Few-shot also reduces divergence (-9.1% average), unlike the *increase* observed in Qwen, with layer-dependent variation (ranging from -25.9% at L9 to $+9.7\%$ at L13).

Table 30: Gemma-2B trajectory divergence by layer and condition. CoT reduces divergence across the tested layers, while few-shot reduces divergence on average but shows layer-dependent variation, unlike Qwen's overall increase.

Layer	Direct	CoT	Few-shot	CoT Δ
4	0.0001	0.0001	0.0001	-5.7%
9	0.0035	0.0030	0.0026	-14.9%
13	0.0094	0.0086	0.0103	-8.3%
17	0.0073	0.0069	0.0071	-6.6%

Position sensitivity. A prefix–suffix divergence difference is also present: suffix tokens show lower divergence than prefix at deeper layers (L13: -4.1%, L17: -7.0%), while early layers show the opposite (L4: $+275\%$, L9: $+15\%$). This provides qualitative cross-architecture evidence for position sensitivity.

Implications. Combined with TinyLlama results (Appendix z), Gemma validation provides cross-architecture evidence for position-sensitive geometric effects, while the sign and magnitude of divergence effects remain architecture-dependent.

AB RESULT LEDGER: CLAIMS AND DATA SOURCES

To ensure transparency and reproducibility, Table 31 maps each main claim to its exact data source, readout position, sample size, effect size, and statistical test.

Notation. *Question-endpoint* (QE) refers to the last token of the question text before any suffix; *prompt-endpoint* (PE) refers to the last token of the full prompt (after CoT suffix, few-shot examples, etc.). *Full* span (F) computes metrics over all trajectory tokens; *question-span* (Q) restricts to question tokens only (controlling for prompt-length confounds). Effect sizes are d_z (paired standardized mean difference) for paired comparisons; other comparisons use Cohen's d.

Statistical corrections. All p-values for multi-layer comparisons (Tables 1, 10) are corrected using Benjamini-Hochberg FDR at α=0.05. Statistical tests use question-level pairing (3 paraphrases aggregated per question) to avoid pseudo-replication. Bootstrap 95% CIs are reported for AUROC estimates.

AC SHUFFLED-LABEL CONTROL (B2)

To test whether the base$\rightarrow$instruct rotation asymmetry ($\mathbf{v}$ rotates more than $\mathbf{u}$) is specific to functional supervision, we trained M=100 projection heads with shuffled leaf labels (identical architecture, optimization, and hidden-state statistics; only pretext labels permuted). Each shuffled head was evaluated on N=50 held-out GSM8K questions (3 paraphrases each, question-level bootstrap CIs).

Results. The real head shows $\Delta\theta_{\text{real}}$=59.8° ($\theta_V - \theta_U$, where θ is the base$\rightarrow$instruct mean-direction rotation). The shuffled null distribution has mean 39.7° $\pm$ 29.0°; 22/100 shuffled heads equal or

Table 31: Result ledger mapping claims to data sources. Readout: QE = question-endpoint, PE = prompt-endpoint. Span: F = full trajectory, Q = question-span only; '–' = not applicable.

Claim	Data	Readout	Span	N	Effect	Table
Divergence (§5.1)						
CoT ≈ Direct divergence	GSM8K	PE	F	100	$\Delta<0.001$	1
Few-shot ↑ divergence +33–49%	GSM8K	PE	F	100	$p<0.001$	1
Depth-wise convergence (direct)	GSM8K	PE	F	100	$0.338\rightarrow0.288$	1
Curvature (§5.2)						
Few-shot ↓ curvature L7–14	GSM8K	PE	F	100	$d_z = +1.90$ (vanishes under span matching)	11
Late-layer span-matched (L27)	GSM8K	QE	Q	100	$d = +0.43$, p=0.012	text
Exemplar relevance control	GSM8K	QE	Q	100	$d = +0.08$, ns	text
CoT minimal curv. change	GSM8K	PE	F	100	$\lvert\bar{\Delta}\kappa\rvert\leq0.016$	10
Position effects (Appendix h)						
Suffix CoT near-invariant QE	GSM8K	QE	–	100	$\Delta<0.001$	App.h
Prefix CoT ↑ 7–14% (L7–L21)	GSM8K	QE	–	100	sig.	App.h
u/v projection (§5.3)						
CoT relatively $\mathbf{u}$-weighted (L14)	GSM8K	PE	F	100	0.053 vs 0.047	2
Few-shot lower $\mathbf{u}/\mathbf{v}$ (L14)	GSM8K	PE	F	100	0.042 vs 0.031	2
Random proj. ablation (L27)	GSM8K	PE	F	100	80% opposite	text
Reflection (Appendix v)						
Endpoint shift predicts success	GSM8K	PE	F	174	AUROC=0.85	App.v
Endpoint shift, full sample	GSM8K	PE	F	200	AUROC=0.84	App.v
Behavioral validation (App.ad)						
Reasoning $\mathbf{u}$-heavy (S_R)	GSM8K	PE	–	200	+0.96 [0.91,1.01]	App.ad
Judgment $\mathbf{v}$-heavy (S_J)	GSM8K	PE	–	200	−0.36 [-0.40,-0.33]	App.ad
Signed dissociation (Γ)	GSM8K	PE	–	200	+1.33 [1.26,1.39]	App.ad
Lexical excess (paired)	GSM8K	PE	–	200	$p<10^{-4}$	App.ad
Procrustes control, 5-fold held-out (App.n)						
$\mathbf{v}$ residual after R^*	GSM8K	PE	–	300	8.8°	App.n
$\mathbf{u}$ residual after R^*	GSM8K	PE	–	300	5.9°	App.n
$\mathbf{v}$–$\mathbf{u}$ gap reduction	GSM8K	PE	–	300	95%	App.n
Instruct-C validation (App.o)						
Γ under fixed primary probe	GSM8K	PE	–	200	$1.33\rightarrow1.59$	App.o
Direct-S baseline shift	GSM8K	PE	–	200	+0.11	App.o
Refusal orthogonality (App.af)						
$\mathbf{v}$ vs. refusal/abstention	AdvBench/SQuAD2	PE	–	100/100	cos≈0	App.af
Shuffled-label control (App.ac)						
Rotation not identifying	GSM8K	PE	–	100	p_{emp}=0.23	App.ac
Cross-scale (§5.7)						
Late-layer format effect (1.5B span-matched)	GSM8K	QE	Q	100	$d = +0.43$	text
7B span-matched curvature (L20)	GSM8K	QE	Q	100	$d=-1.47$	App.y
7B span-matched curvature (L27)	GSM8K	QE	Q	100	$d=-0.56$	App.y
Intervention (§5.5)						
Jacobian $\lvert\Delta_v\rvert$ vs $\lvert\Delta_u\rvert$ (L27)	GSM8K	PE	–	200	0.37 vs 0.06; gap at 57th pct of random pairs (null)	App.t
Linear-surrogate grid	GSM8K	PE	–	50	$\mathbf{u}$≈$\mathbf{v}$≈random	App.t
Probe provenance and further tasks						
Symmetric probe (Γ, 4 cells)	GSM8K	PE	–	200	0.17–1.95, all > 0	App.p
ProofWriter (Γ, canonical head)	PW	PE	–	300	+0.07 [0.03,0.10]	App.ae
Cross-architecture						
Gemma-2B CoT ↓ div.	GSM8K	PE	F	15	−8.9%	30
TinyLlama curv. responsive	GSM8K	PE	F	25	qualitative	App.z
Robustness checks (App.t, n, y)						
Γ across seed/dimension heads	GSM8K	PE	–	200	$\Gamma>0$ 8/8; signed 5/8	App.t
Subspace identifiability	–	–	–	3 seeds	0.19–0.27 vs 1.03	App.t
CKA after R^*	GSM8K	PE	–	300	0.67/0.69	App.n
Turning-angle curvature (full traj.)	GSM8K	PE	F	100	depth structure preserved; L27 CoT sign flips	App.y
Policy wording (S_J)	GSM8K	PE	–	200	rule 3/3 $S_J<0$; verification 0/3 $S_J<0$	App.ad
Cross-task (Γ, CSQA)	CSQA	PE	–	200	+0.10, not signed	App.ad
Dim. subsets (primary head)	GSM8K	PE	–	200	top-8 PCs Γ=1.36	App.ad
Retrained d=16–128 heads	GSM8K	PE	–	200	Γ 0.49–0.99, signed 3/4	App.ad

exceed the real head's asymmetry. Empirical p-value: $p_{\text{emp}}=(1+22)/(100+1)=0.23$. The real head lies at the 78th percentile.

Interpretation. The rotation asymmetry alone does not establish semantic meaning; shuffled-label heads can exhibit comparable or larger $\mathbf{v}>\mathbf{u}$ rotation. This motivates the independent behavioral validation (Appendix ad) as the primary construct-validity test. Training geometry is descriptive; behavioral dissociation provides the primary construct-validity evidence.

AD INDEPENDENT BEHAVIORAL VALIDATION (C)

We test whether independent reasoning and judgment manipulations preferentially engage $\mathbf{u}$ and $\mathbf{v}$, respectively, using controlled behavioral contrasts with matched lexical controls. This validation uses

the Qwen2.5-1.5B base checkpoint; replication on the instruct checkpoint (Appendix o) shows the dissociation is preserved, with a larger measured Γ.

Design. *C1 (Reasoning):* Same GSM8K question; vary whether derivation is supplied (R$-$) or must be derived independently (R$+$). Lexical control: same R$-$ demand, bullet vs. numbered formatting. *C2 (Judgment):* Same reasoning trace; vary decision policy: permissive (J$-$: "return unless explicit contradiction") vs. conservative (J$+$: "commit only if fully supported"). Lexical control: same J$+$ policy, different wording.

Stimulus wording. The exact instruction suffixes used in each condition:

R$-$ (low reasoning demand): *"Using the derivation above, provide the final numerical answer."*

R$+$ (high reasoning demand): *"Independently derive the solution step by step, then provide the final numerical answer."*

$\mathbf{R}_{\text{lex}}$ (lexical control): Same text as R$-$, with bullet formatting instead of numbered steps.

J$-$ (permissive policy): *"Policy: Return the proposed answer unless there is an explicit arithmetic contradiction in the reasoning."*

J$+$ (conservative policy): *"Policy: Commit to the proposed answer only if every step is fully supported. If any step is uncertain or potentially wrong, respond with UNSURE."*

$\mathbf{J}_{\text{lex}}$ (lexical control): *"Policy: Only accept the proposed answer when all reasoning steps are completely justified. When evidence is insufficient, respond with UNSURE."*

We read the L27 residual state at the final token of the complete user message, immediately before assistant generation; all treatment text precedes this position. N=200 questions, 6 stimuli each (1200 total).

Relative functional organization transfers across tasks. The signed geometry is task-dependent, but its relative organization is more stable. We evaluate on CommonsenseQA (N=200 dev questions; supplied explanations from ECQA (Aggarwal et al., 2021); decision-rule wording v4 with v5 as lexical control) and ProofWriter (N=300 depth-5 OWA test questions; deductive reasoning with supplied proof chains vs. independent derivation; Appendix ae).

Table 32: Cross-task comparison of behavioral dissociation.

Statistic	GSM8K	CommonsenseQA	ProofWriter
S_R	$+0.96$ [0.91, 1.01]	$+0.79$ [0.74, 0.83]	$+0.49$ [0.45, 0.52]
S_J	-0.36 [-0.40, -0.33]	$+0.69$ [0.65, 0.72]	$+0.42$ [0.40, 0.44]
Γ	$+1.33$ [1.26, 1.39]	$+0.10$ [0.04, 0.17]	$+0.07$ [0.03, 0.10]

On CommonsenseQA, the reasoning–judgment ordering remains positive (Γ=0.10, 95% CI $[0.04, 0.17]$) despite a reversal of the signed judgment contrast. We obtain the same relative ordering on a third, deductive-reasoning domain: on ProofWriter (N=300), the canonical head yields S_R=0.486 and S_J=0.418 (Γ=0.068, 95% CI $[0.035, 0.101]$). The independently constructed answer-choice/stage head (Set B) also preserves this relation (S_R=0.365, S_J=0.111, Γ=0.255, 95% CI $[0.235, 0.275]$). Thus absolute signed coordinates do not transfer universally, whereas the relative relation $S_R > S_J$ reproduces on CommonsenseQA and ProofWriter, with small margins, and across distinct supervision constructions.

Task-dependent judgment readout. The judgment-associated readout is task-dependent rather than globally signed. On GSM8K, the fixed-trace judgment contrast is $\mathbf{v}$-weighted ($S_J < 0$), whereas on CommonsenseQA and ProofWriter the same operational contrast becomes $\mathbf{u}$-weighted ($S_J > 0$). Thus the transferable structure is the relative dissociation between reasoning and judgment manipulations ($\Gamma > 0$), not a universal sign convention for the judgment factor. Absolute signed realization is task-dependent, while relative functional organization is the more stable object.

Metric. For each question i, define $S_{R,i}=\log\big((\|\Delta\mathbf{u}_{R,i}\|+\epsilon)/(\|\Delta\mathbf{v}_{R,i}\|+\epsilon)\big)$, where $\Delta\mathbf{u}_{R,i}=\mathbf{u}_i(\mathrm{R}+)-\mathbf{u}_i(\mathrm{R}-)$ is the displacement in $\mathbf{u}$-space from R$-$ to R$+$ (analogously for $\mathbf{v}$), with $\epsilon=10^{-6}$ for numerical stability. Similarly, $S_{J,i}$ measures displacement from J$-$ to J$+$. We report $S_R=\frac{1}{N}\sum_i S_{R,i}$ and $S_J=\frac{1}{N}\sum_i S_{J,i}$, with primary statistic $\Gamma=S_R-S_J$. Primary prediction: $\Gamma>0$ (relative dissociation); a stronger signed prediction is $S_R>0$ and $S_J<0$. For the primary GSM8K evaluation, all inference is question-level ($N=200$); the six stimuli per question are repeated conditions, not independent observations. We report 95% bootstrap CIs (1000 resamples) and paired t-tests for functional-minus-lexical contrasts.

Results. $S_R=+0.96$ (95% CI $[0.91, 1.01]$): reasoning manipulation is $\mathbf{u}$-heavy. $S_J=-0.36$ (95% CI $[-0.40, -0.33]$): judgment manipulation is $\mathbf{v}$-heavy. $\Gamma=+1.33$ (95% CI $[1.26, 1.39]$): strong signed dissociation.

Lexical controls: $S_{R,\mathrm{lex}}=+0.44$ (CI $[0.39, 0.50]$), $S_{J,\mathrm{lex}}=-0.31$ (CI $[-0.33, -0.29]$). Paired excess tests are computed at the question level (not from the difference of aggregate means) and confirm functional effects exceed lexical: $\overline{S_R-S_{R,\mathrm{lex}}}=+0.52$ (CI $[0.45, 0.59]$, $p<10^{-4}$); $\overline{|S_J|-|S_{J,\mathrm{lex}}|}=+0.08$ (CI $[0.05, 0.11]$, $p<10^{-4}$). The judgment functional-over-lexical excess is smaller than reasoning's but reliably positive; the primary evidence for factor dissociation is the opposite-signed engagement ($S_R>0$, $S_J<0$) and the large Γ, not the magnitude of either lexical excess alone.

Policy wording. We evaluate six J+/J$-$ pairs in total: three decision-rule phrasings (including the primary pair) and three verification phrasings.

Table 33: S_J under alternative policy wordings ($N=200$).

Ver.	Type	J$-$ wording	J$+$ wording	S_J [95% CI]
v0	rule (primary)	Return the proposed answer unless there is an explicit arithmetic contradiction in the reasoning.	Commit to the proposed answer only if every step is fully supported. If any step is uncertain or potentially wrong, respond with UNSURE.	-0.36 [-0.40, -0.33]
v4	rule	Return the proposed answer unless it is explicitly contradicted.	Return the proposed answer only if it is fully supported; otherwise respond UNSURE.	-0.38 [-0.42, -0.35]
v5	rule	Accept the proposed answer unless there is a clear contradiction.	Accept the proposed answer only when all steps are justified; otherwise respond UNSURE.	-0.23 [-0.26, -0.20]
v1	verify	Accept the answer by default; only reject if there's an obvious calculation error.	Verify each step rigorously. Return UNSURE if any calculation lacks full justification.	$+0.84$ [$+0.79$, $+0.89$]
v2	verify	Trust the reasoning and return the answer unless you find a clear mistake.	Apply strict verification. Only confirm if every step is unambiguously correct; otherwise abstain from answering.	$+0.52$ [$+0.47$, $+0.56$]
v3	verify	Be lenient. Output the proposed answer unless the logic is clearly flawed.	Be cautious. Confirm the answer only when fully certain. Express doubt with UNSURE.	$+0.79$ [$+0.74$, $+0.83$]

Decision-rule phrasings give $S_J<0$ in 3/3 (-0.23 to -0.38); verification phrasings give $S_J>0$ in 3/3 ($+0.52$ to $+0.84$). Verification instructions are consistent with added reasoning demand and stronger $\mathbf{u}$ engagement; across these tested wordings, the $\mathbf{v}$ preference is specific to the decision-rule formulation.

Strength of behavioral identification. On GSM8K, the relative dissociation is positive with a confidence interval excluding zero, the stronger signed pattern holds ($S_R>0$, $S_J<0$), and both functional contrasts exceed their matched lexical controls. On CommonsenseQA, only the relative dissociation transfers: $\Gamma>0$ with confidence interval excluding zero, but the signed judgment association does not.

Factorial test of independence. We cross reasoning demand (R−/R+) and judgment policy (J−/J+) in a 2×2 design (N=200) to test whether the factors are independently engaged. Simple effects stratified by policy:

- *Reasoning in both strata:* $S_R^{J-} = +1.34$ (permissive policy), $S_R^{J+} = +0.18$ (conservative policy). Both positive: reasoning remains $\mathbf{u}$-weighted regardless of judgment policy.
- *Judgment in both strata:* $S_J^{R-} = -0.30$ (derivation supplied), $S_J^{R+} = +0.21$ (model must derive). The judgment effect is $\mathbf{v}$-weighted when reasoning is held fixed but reverses to $\mathbf{u}$-weighted when the model must actively reason.

The asymmetric pattern indicates that the factors are distinguishable but not independent: active reasoning changes the geometry through which judgment policy is expressed. We interpret this as a conditional rather than fully independent organization.

Together, these contrasts behaviorally identify a relative reasoning–judgment organization in the $\mathbf{u}/\mathbf{v}$ readout, with the judgment-associated geometry conditional on reasoning state.

Dimensionality. Within the primary 64-D head, the top-8 principal components of the projected outputs carry the full dissociation (Γ=1.36 vs. 1.33 at d=64; bottom components $\Gamma \leq 0.08$), and random 8-, 16-, and 32-dimensional subsets retain 87–95% of Γ with the signed pattern intact. Heads retrained from scratch at $d \in \{16, 32, 64, 128\}$ give Γ=0.49–0.99 with the signed pattern in three of four. Across all eight heads we trained (Table 26, plus these four), $\Gamma > 0$ throughout and the signed pattern holds in five. The judgment effect is generally weaker and more variable than the reasoning effect ($|S_J|$: 0.01–0.59 vs. $|S_R|$: 0.17–0.96), although this ordering is not universal across probe realizations. Dimensionality itself is not the controlling factor: d=64 heads from two training runs give Γ=1.33 and 0.99.

AE PROOFWRITER CROSS-TASK EVALUATION

We evaluate the behavioral dissociation on ProofWriter (Tafjord et al., 2021), a deductive reasoning dataset requiring multi-step logical inference from explicit rules.

Dataset and selection. We use the depth-5 OWA (open-world assumption) test split (N=300 questions with valid proof chains). Each question provides a context of facts and rules, a hypothesis to evaluate, and a ground-truth proof chain. We select questions where the proof chain is non-empty and the answer is either True or False.

Factorial design. We construct a 2×2 factorial analogous to GSM8K:

- **R**− (supplied reasoning): The model receives the ground-truth proof chain before making a judgment.
- **R**+ (independent reasoning): The model must derive the proof independently before judging.
- **J**− (permissive policy): "Return the answer unless there is an explicit logical contradiction."
- **J**+ (conservative policy): "Commit only if every inference step is fully supported."

This yields 4 stimuli per question (1200 total).

Extraction and metric. We read the L27 residual state at the final token of the user message, immediately before generation. All conditions use the same extraction point. We compute S_R, S_J, and Γ as defined in the main text. Bootstrap CIs use 1000 resamples at the question level (N=300).

Results. Both heads yield $\Gamma > 0$:

- **Set A** (canonical head): S_R=0.486, S_J=0.418, Γ=0.068 (95% CI $[0.035, 0.101]$).
- **Set B** (answer-choice/stage head): S_R=0.365, S_J=0.111, Γ=0.255 (95% CI $[0.235, 0.275]$).

Both 95% bootstrap CIs exclude zero. The independently constructed Set B head also preserves the relative ordering, showing that this result is not restricted to the original supervision construction.

Interpretation. The signed geometry differs from GSM8K: both S_R and S_J are positive, indicating $\mathbf{u}$-weighted engagement for both manipulations. However, the relative ordering $S_R > S_J$ is preserved. This confirms that the transferable structure is the relative dissociation, not a universal sign convention.

AF THE $\mathbf{v}$ FACTOR IS DISTINCT FROM REFUSAL/ABSTENTION DIRECTIONS

A natural question is whether $\mathbf{v}$ merely re-packages the refusal or abstention directions studied in the activation-steering literature (Arditi et al., 2024; Zou et al., 2023a). We test this directly.

Method. Following the standard difference-of-means construction, we build a *refusal* direction from harmful-minus-harmless prompts (N=100 pairs from AdvBench (Zou et al., 2023b)) and an *abstention* direction from unanswerable-minus-answerable questions (N=100 pairs from SQuAD 2.0 (Rajpurkar et al., 2018)). The refusal and abstention directions are constructed directly in the raw L27 hidden space ($\mathbb{R}^{1536}$) as $\mathbf{d} = \bar{\mathbf{h}}_{\text{harmful}} - \bar{\mathbf{h}}_{\text{harmless}}$ (and analogously for abstention). To compare them with the 64-dimensional JM factors, we map each factor direction back to hidden space using the same linear surrogate P_f^{T} defined in Appendix t, normalize the resulting hidden-space direction, and compute cosine similarity with the refusal or abstention direction. As in the intervention analysis, this back-projection is a linear surrogate rather than the exact Jacobian of the nonlinear head.

Result: negligible alignment. Neither direction aligns with $\mathbf{v}$ or $\mathbf{u}$:

Direction	$\cos(\mathbf{v}, \cdot)$	Angle	$\cos(\mathbf{u}, \cdot)$	Angle
Refusal	-0.011	90.6°	-0.036	92.0°
Abstention	$+0.003$	89.8°	-0.021	91.2°
Random (mean $\lvert\cos\rvert$)	0.019 ± 0.023		0.019 ± 0.023	

All measured cosines fall within the null distribution of random projections (N=100): $|\cos(\mathbf{v}, d_{\text{refusal}})| = 0.011$ ranks at the 34th percentile; $|\cos(\mathbf{v}, d_{\text{abstain}})| = 0.003$ ranks at the 7th percentile. The tested surrogate directions show negligible alignment with both safety-style directions.

Interpretation. This is a novelty control, not a semantic one: $\mathbf{v}$'s judgment role is established behaviorally (Appendix ad, Γ=1.33), by manipulating commit-versus-abstain policy on reasoning problems. The present result shows that the reasoning-task judgment readout is not simply aligned with these tested refusal or abstention directions. Commitment policy and safety refusal are distinct operational constructs; relating their representation geometry is left to future work.