MECHANISTIC INTERPRETABILITY WORKSHOP · 43RD ICML · SEOUL, SOUTH KOREA · 2026

Geometry of Ordinal Representations in Language Models

Do the geometric motifs behind counting generalize across ordinal tasks: and across model architectures?

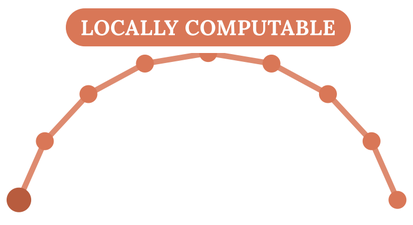
Saksham Bassi* **Sharvi Tomar***

*Equal contribution · Correspondence: stomar2@illinois.edu

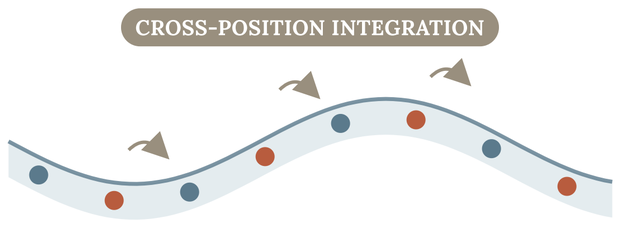
THE CENTRAL FINDING

A Taxonomy of Ordinal Geometry

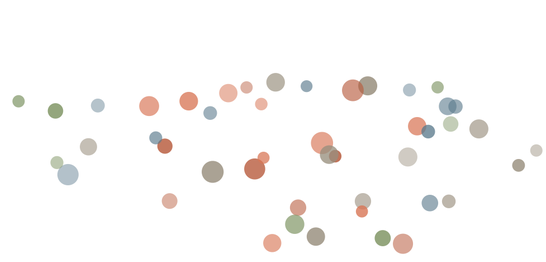
Across four tasks and three model families, how cleanly a variable is geometrically represented tracks **how directly it maps onto token identity** (not how accurately it can be decoded).



Bracket depth · Table column
1D manifold
pre-organized at the embedding layer
max twist ≤ 0.13



Indentation · Table row
2:4D manifold
actively constructed by attention heads
twist up to 0.80



Numeric magnitude
4:5D, diffuse
no clear ordinal axis
train R^2 > 0.93, but no manifold

"Tasks where the ordinal variable is directly computable from token identity produce clean manifolds; tasks requiring cross-position integration or semantic extraction don't: regardless of how well a linear probe can decode them."

01 Motivation

Gurnee et al. (2026) showed Claude 3.5 Haiku represents character counts on **curved 1D manifolds**, with SAE features acting as place cells and attention heads geometrically twisting the manifold to compute distances. We test whether this generalizes across **four ordinal tasks** and **three open-weight model families**.

KEY FINDINGS

- The manifold hypothesis only **partly generalizes**: locally-computable tasks get clean 1D manifolds, others don't.
- Geometric computation is **architecture-dependent**: Qwen3-4B twists 4:10× harder than Gemma at similar probing accuracy.
- High twist alone isn't enough: Qwen's indentation twisters preserve order (ρ≥0.77); its numeric twisters don't (ρ≤0.23).
- Patching manifold subspaces drops probe accuracy **28×+ more** than random-direction controls, in every task & model.

FOUR ORDINAL TASKS

Bracket depth
`( [ { } ] )`
Nesting depth 0:10 · 3,000 seqs (synthetic + The Stack)

Indentation
```
if x:
    y = 1
```
Python indent level 0:5 (×4 spaces) · 3,000 samples, The Stack

Table position
`| a | b | c |`
Row 0:14, column 0:8 · 2,000 synthetic tables

Numeric magnitude
`"1,204 units"`
$\lfloor \log_{10}(\text{value}) \rfloor$, 0:8 · 5,000 samples, OpenWebText

THREE MODELS

Model	Params	Layers	Heads	d_{model}
Gemma-2-2B	2.6B	26	8	2304
Gemma-2-9B	9.2B	42	16	3584
Qwen3-4B	4.0B	36	32	2048

Qwen3-4B adds RoPE positional encoding. Gemma Scope SAEs (16K-feature dictionary) exist only for the Gemma models: so feature-tiling and strong twisting are demonstrated on different models.

THE STUDY, BY THE NUMBERS

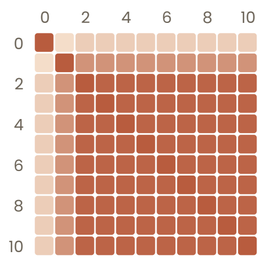
4 ordinal tasks · **3** model families · **13k** labeled examples · **16K** SAE dictionary size

02 Methods

1. **Probe** — Logistic / ridge regression per layer (4k train / 1k test).
2. **Discover manifold** — PCA on per-value mean activations; count PCs for 90% variance.
3. **Find feature families** — Gemma Scope SAEs at 4 layers; selective if peak > 3×mean, monotonic if non-decreasing.
4. **Analyze head geometry** — Project centroids into Q-space; score twist, alignment, and ordinality.
5. **Ablate subspace** — Patch top-3 PCA directions vs. k random controls; compare probe-accuracy drop.

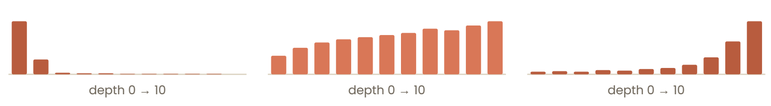
twist = $1-\text{corr}(d_r, d_Q)$ · **align** = $\text{Var}(\text{PC1}_Q)/\text{Var}(\text{tot})$ · **ord** = $|\rho(\text{value}, \text{PC1})|$

FIG 1 · MANIFOLD GEOMETRY



Pairwise cosine similarity, bracket depth (Gemma-2-2B, L8). Depth 0 sits in its own direction; depths 2:10 collapse to nearly the same direction: a 1D manifold with a single outlier.

FIG 2 · SAE PLACE-CELL TILING

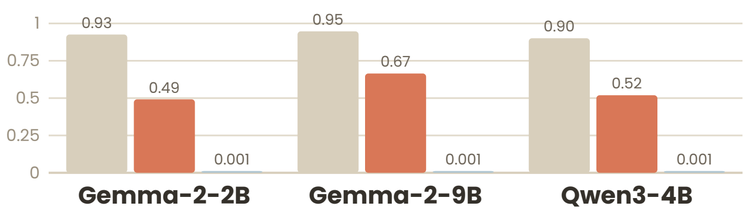


Depth=0 detector (F9213) · **Monotonic ↑ (F302)** · **Peaks at depth 10 (F13542)**

Discrete SAE features tile the bracket-depth manifold like place cells along a line: one per landmark position. For numeric magnitude, by contrast, **zero** of the top features are monotonic at any layer, in either Gemma model.

Cross-task sharing: recurring features across tasks are all value=0 "beginning-of-structure" detectors. No head appears in the top-20 twisters for all four tasks.

SUBSPACE ABLATION · INDENTATION



Patching the top-3 manifold directions causes a large probe-accuracy drop; patching random directions barely moves it: a **≥28×** ratio in every task:model pair (full results in Table 4 of the paper).

03 Results

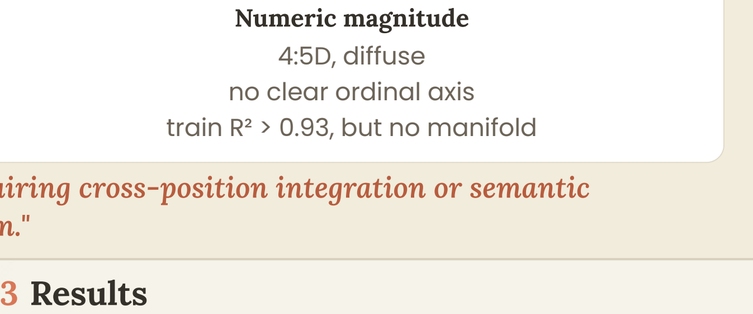
PROBING ACCURACY / R^2 (BEST LAYER, TEST SET)

	Gemma-2-2B	Gemma-2-9B	Qwen3-4B
Brackets	0.538	0.525	0.573
Table cols	0.983	0.996	0.992
Indentation	0.943	0.947	0.944
Table rows	0.746	0.821	0.803
Numeric (R^2)	0.536	0.549	0.688

lower ■ higher

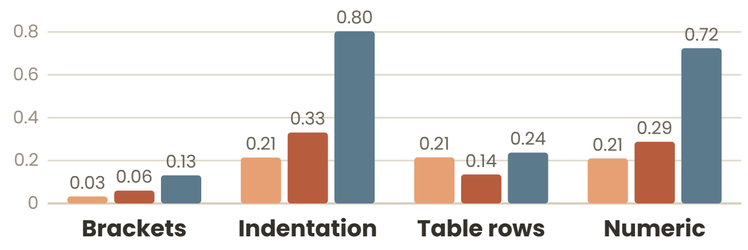
The ordinal variable is linearly decodable in every task and model. Bracket depth looks low (0.53:0.57) but that's **overfitting**: train accuracy exceeds 99%.

FIG 4 · MANIFOLD DIMENSIONALITY ACROSS LAYERS



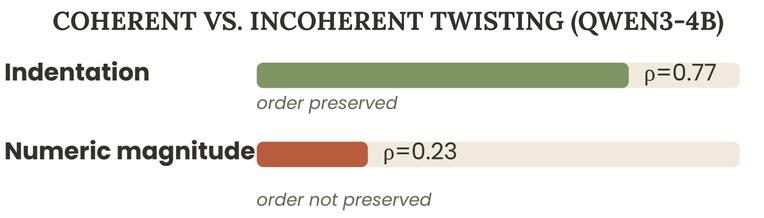
Indentation, PC1 variance explained. Qwen3-4B collapses to a near-1D manifold at mid-layers (step-jump to ~100%) while both Gemma models stay diffuse (45:65%): the same architecture-dependence shown across tasks in Fig. 4 of the paper.

TWISTING MAGNITUDE BY TASK (MAX TWIST SCORE)



Qwen3-4B twists 4:10× harder than either Gemma model on tasks that require manifold construction (indentation, numeric): possibly reflecting RoPE or head count.

COHERENT VS. INCOHERENT TWISTING (QWEN3-4B)



High twist doesn't always mean ordinal computation: Qwen's indentation heads rearrange the manifold while preserving order; its numeric heads just rearrange it.

THE RESULTS, BY THE NUMBERS

28×+ manifold vs. random ablation drop · **4:10×** Qwen3-4B vs. Gemma twist ratio · **0.77** Qwen indentation ordinality ρ · **0/20** monotonic SAE features, numeric

LIMITATIONS

- Bracket-depth probe overfits heavily (99% train vs. 53:57% test); read its manifold structure with that caveat.
- Synthetic bracket data always closes with ')' regardless of opener: a minor labeling inconsistency.
- No pretrained SAE exists for Qwen3-4B, so feature-tiling and strong twisting are shown on different models.
- Twist score uses per-value *means* and may miss token-level transformations.
- Subspace ablation shows informational concentration, not behavioral causality on generation.

CONCLUSION

A tentative taxonomy, organized by how directly the ordinal variable maps onto token identity: **locally computable** tasks (brackets, table columns) get clean, pre-organized 1D manifolds; **cross-position integration** tasks (indentation, table rows) get 2:4D manifolds actively constructed by attention heads; **semantic extraction** (numeric magnitude) gets a diffuse 4:5D representation despite train R^2 > 0.93. Geometric computation is architecture-dependent: Qwen3-4B's heads twist far harder than Gemma's when construction is required: yet the twist concentration ratio (≈5:8×) is similar across models, so the gap looks global rather than head-specialized.

FUTURE WORK

- Train SAEs on Qwen3-4B to test whether stronger twisting yields cleaner feature tiling.
- Apply information-geometric tools (Park et al., 2026) directly to the curved manifolds.
- Scale to larger models: Gemma-2-27B, Qwen3-14B.
- Test whether patching these subspaces causally shifts downstream generation.