A Hyperbolicity Atlas of Large Language Model Hidden States

Anonymous ACL submission

Abstract

Hidden states of large language models (LLMs) are typically analyzed as flat Euclidean vectors through probing, anisotropy, and intrinsic-dimension lenses. Yet much of what these models process (taxonomies, program structure, multi-step derivations) is hierarchical, and whether hidden-state distances reflect such hierarchy remains untested. We present the first systematic study of Gromov hyperbolicity (GH), a distance-based measure of tree-likeness, in prompt-token hidden states of open-weight LLMs. Across model families, scales, layer depths, and domains spanning math, code, commonsense, and truthfulness benchmarks, we construct a hyperbolicity atlas charting the normalized statistic δ_{rel} along each axis. Four findings emerge: (i) depth dominates, with middle layers forming a high-δ_{rel} plateau and final layers dropping toward tree-like structure; (ii) scale is weak and non-monotonic; (iii) matched-scale families occupy distinct regimes, with specialization gaps exceeding scale effects; and (iv) domains interact with specialization, e.g., HumanEval shifts from the highest- to the lowest-δ_{rel} domain between base and coder checkpoints. GH thus serves as a lightweight diagnostic that localizes where hierarchical distance structure emerges and which comparisons warrant deeper analysis.

1 Introduction

Transformer language models store token information in Euclidean vectors,yet many things they process are hierarchical: lexical taxonomies, entailment relations, syntactic trees, discourse dependencies, program blocks, and multi-step derivations. Hyperbolic geometry is useful for analyzing such structure because negatively curved spaces can represent trees and hierarchies with low distortion (Nickel and Kiela, 2017, 2018; Ganea et al., 2018). This observation has already shaped hyperbolic word and sentence embeddings (Dhingra et al., 2018; Tifrea et al., 2018), multilingual and multimodal representation guidelines (Sawhney et al., 2024), and recent hyperbolic fine-tuning or architecture proposals for LLMs (Yang et al., 2026; He et al., 2026; Patil et al., 2025). For contemporary LLMs, however, an important question remains open: even when the model is not trained in a hyperbolic manifold, do its internal hidden states exhibit tree-like metric organization?

Existing representation analyses answer related but different questions. Probing classifiers test whether a feature is recoverable; anisotropy and intrinsic-dimension analyses characterize spread or local complexity; structural probes show that some linguistic relations are geometrically recoverable (Ethayarajh, 2019; Hewitt and Manning, 2019; Valeriani et al., 2023). These methods do not directly test whether distances among hidden states behave like a hierarchy. Gromov Hyperbolicity (GH) provides a simple complementary test: it asks whether distances among hidden-state vectors behave more like distances in a tree. To our knowledge, this paper is the first systematic study of GH for prompt-token hidden states in modern open-weight LLMs.

We use "atlas" literally: we map the statistic across axes instead of giving one number per model. For a model m, layer ℓ, domain d, and prompt x, we measure the normalized statistic

$$\mathcal{A}(m, \ell, d, x) = \delta_{\mathrm{rel}}(X_{m,\ell,x}), \qquad (1)$$

where $X_{m,\ell,x}$ is the set of prompt-token hidden states for that layer. This map is necessary because the axes need not move together. A larger model may change the final layer without changing the middle layers; a domain may be tree-like for one specialization but not for another; and a family difference at matched parameter count can exceed a scale difference within a family. The empirical question is therefore not whether "LLMs are hyperbolic," but where lower or higher GH appears across models, layers, and domains.

We organize the study around four research questions: (1) Does normalized hyperbolicity change when scale grows from 1.5B to 7B within a family? (2) How does hyperbolicity vary across transformer layers within a model? (3) At matched scale, do different 7/8B families exhibit distinct final-layer profiles? (4) For a fixed model, how does hyperbolicity differ across math, code, commonsense, and truthfulness prompts?

To answer these questions, we compute prompt-only hidden states for ten open-weight checkpoints from Llama, Qwen2.5, Qwen2.5-Math, Qwen2.5-Coder, and DeepSeek-R1-Distill-Qwen families on MATH500, HumanEval, WinoGrande, and TruthfulQA. For each prompt and layer, we sample quadruples of token hidden states, estimate four-point GH, and normalize by the largest pairwise distance in that set. This design keeps the comparisons aligned: within-family scale contrasts, layerwise trajectories, matched 7/8B family comparisons, and fixed-model domain comparisons are all produced by the same measurement protocol. It also avoids a common shortcut: generated responses are excluded so that prompt geometry is not confounded with decoding length, answer correctness, or model-specific generation behavior.

The resulting map yields four main findings. First, scale alone is weak and non-monotonic: Qwen2.5-7B-Base is more tree-like than Qwen2.5-1.5B on some domains but less tree-like on others, while the math track moves upward in δ_{rel}. Second, depth dominates the geometry: middle layers form a high-δ_{rel} plateau, and final layers usually drop toward more tree-like structure. Third, matched 7/8B families occupy distinct value ranges, with family and specialization gaps larger than the pooled scale shifts. Fourth, domains interact with specialization: HumanEval is high-δ_{rel} for Qwen2.5-7B-Base but becomes the lowest-δ_{rel} domain for Qwen2.5-7B-Coder-Instruct.

Contributions. Our main contributions are as follows:

- We introduce the first multi-axis hyperbolicity atlas for LLM hidden states, systematically charting GH across parameter scale, layer depth, model family, and input domain under a unified measurement protocol.
- We establish that depth is the dominant axis: all studied models exhibit a high-δ_{rel} middle-layer plateau followed by a final-layer drop toward tree-like structure, a pattern robust across families and domains.
- We show that model specialization reshapes hidden-state geometry in domain-selective ways, with family and specialization gaps exceeding within-family scale effects, demonstrating that GH is a practical, lightweight diagnostic for comparing representational organization across LLMs.

2 Background

2.1 Hyperbolicity as Tree-Likeness

Gromov hyperbolicity measures how close a metric space is to a tree metric (Gromov, 1987). For four points a, b, c, d with pairwise distance $d(\cdot,\cdot)$, define

$$s_1 = d(a,b) + d(c,d), \tag{2}$$
$$s_2 = d(a,c) + d(b,d), \tag{3}$$
$$s_3 = d(a,d) + d(b,c). \tag{4}$$

Let $s_{(1)} \le s_{(2)} \le s_{(3)}$ be the sorted values. The four-point hyperbolicity of the quadruple is

$$\delta(a,b,c,d) = \frac{1}{2}\left(s_{(3)} - s_{(2)}\right). \tag{5}$$

A tree metric has $\delta = 0$ for every quadruple. Lower values therefore indicate a more tree-like metric structure, while larger values indicate stronger deviation from tree-likeness.

Raw δ depends on the size of the distances, so it is hard to compare models with different hidden dimensions and activation norms directly. We report a diameter-normalized statistic

$$\delta_{\text{rel}}(X) = \frac{\widehat{\delta}(X)}{\operatorname{diam}(X)}, \tag{6}$$

where X is a set of hidden-state vectors, $\widehat{\delta}$ is the sampled four-point estimate, and $\operatorname{diam}(X)$ is the maximum observed pairwise distance in the sampled set. In the rest of the paper, lower δ_{rel} means that the token-state distance pattern is more tree-like after normalization.

2.2 Prompt Hidden-State Sets

For model m, layer ℓ, prompt x, and token position t, let $h_{m,\ell}(x,t) \in \mathbb{R}^{d_m}$ denote the hidden state. For each prompt and layer, we collect the hidden states of all prompt tokens:

$$X_{m,\ell,x} = \{h_{m,\ell}(x,t) : t \in T(x)\}, \tag{7}$$

where $T(x)$ is the set of prompt-token positions. We use Euclidean distances over the model hidden-state coordinates and estimate δ_{rel} separately for each sample and layer. This prompt-only design isolates geometry induced by the input distribution from geometry induced by model-specific generated text length, decoding behavior, and answer correctness.

3 Experimental Design

3.1 Models, Data, and Scope

We analyze ten open-weight checkpoints from the Llama, Qwen2.5, Qwen2.5-Math, Qwen2.5-Coder, and DeepSeek-R1-Distill-Qwen families (Grattafiori et al., 2024; Yang et al., 2024; Hui et al., 2024; Guo et al., 2025). The 1.5B models include Qwen2.5-1.5B, Qwen2.5-1.5B-Instruct, Qwen2.5-Math-1.5B, and DeepSeek-R1-Distill-Qwen-1.5B. The 7/8B set includes Llama-3.1-8B, Llama-3-8B-Instruct, Qwen2.5-7B-Base, Qwen2.5-7B-Instruct, Qwen2.5-7B-Coder-Instruct, and Qwen2.5-Math-7B.

The datasets cover mathematical reasoning (MATH500, derived from MATH and verifier work), code generation (HumanEval), commonsense pronoun resolution (WinoGrande), and truthfulness (TruthfulQA) (Hendrycks et al., 2021; Lightman et al., 2024; Chen et al., 2021; Sakaguchi et al., 2021; Lin et al., 2022). The sample counts are 500 for MATH500, 164 for HumanEval, 1,267 for WinoGrande, and 817 for TruthfulQA for each model, yielding the same dataset composition across checkpoints. The measurements are prompt-only: generated responses in the local result directory are not used for the main atlas because generation length and answer correctness would add another uncontrolled axis.

3.2 Measurement Protocol

For every model, dataset example, and extracted layer, we compute prompt-token hidden states without gradient updates. One measurement uses the prompt-token hidden states from a single example at a single layer. For that set, we sample 2,000 quadruples of token states and compute the normalized four-point estimate in Eq. 6. We retain the embedding layer and every available transformer layer, giving 818,904 valid sample-layer measurements.

The main analysis uses medians and interquartile ranges because per-example δ_{rel} distributions are skewed and often bounded near 1 in middle layers. For paired layer comparisons we use the Wilcoxon signed-rank test over matched examples. For unpaired family and domain comparisons we use Kruskal–Wallis tests and report median shifts or ranges. We interpret the results as distributions, not as yes-or-no labels. A lower median does not prove that a model stores an explicit symbolic tree; it means the sampled token-state distances are closer to a tree metric under the four-point test.

4 Results

Lower δ_{rel} means the prompt-token distance pattern is closer to a tree metric after diameter normalization. All tables report medians with interquartile ranges unless otherwise stated.

4.1 RQ1: Scale Effects Are Conditional

Figure 1 and Table 1 show that parameter scale alone is a weak predictor of prompt geometry. Pooled over domains, the base track shifts by only -0.029 from 1.5B to 7B, while the math track shifts by +0.035; the pooled Cliff's Δ is +0.08. The small aggregate effect hides domain-specific reversals. For the base track, the larger model is less tree-like on MATH500 and HumanEval (+0.024 and +0.021), but more tree-like on WinoGrande and TruthfulQA (-0.022 and -0.051). For the math track, the 7B model is higher-δ_{rel} in every domain, with the largest shift on HumanEval (+0.093), suggesting that more math-specialized capacity does not automatically produce more tree-like final prompt geometry.

The direction of the base-track shift is especially informative. MATH500 and HumanEval are the two domains with the most explicit formal structure, yet they do not both become more tree-like with scale; in fact, both move upward in δ_{rel} for Qwen2.5-7B-Base. Conversely, WinoGrande and TruthfulQA move downward, even though they are less obviously tree-structured tasks. This indicates that the statistic is not simply tracking whether a dataset name sounds hierarchical. It is tracking how a particular family reorganizes prompt-token distances after training and scaling. The weak pooled effect is therefore informative: scale matters less than the combination of family, domain, and layer.

The layer profiles explain why final-layer scale summaries should be read cautiously. Both tracks quickly enter a middle-layer plateau near $\delta_{\mathrm{rel}} \approx 1$, and most of the visible scale effect occurs in the

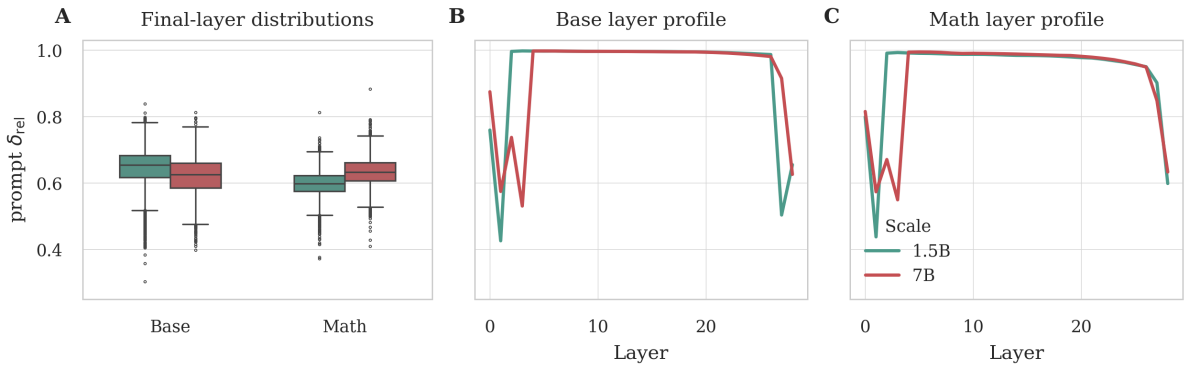

Figure 1: RQ1 scale comparison within the Qwen2.5 base and Qwen2.5-Math tracks. Panel A shows final-layer distributions pooled over domains; panels B–C show median layer profiles.

Track	Dataset	1.5B	7B	7B–1.5B
Base	MATH500	0.526 [0.486, 0.568]	0.549 [0.519, 0.591]	+0.024
	HumanEval	0.655 [0.636, 0.687]	0.676 [0.640, 0.712]	+0.021
	WinoGrande	0.670 [0.649, 0.694]	0.648 [0.624, 0.670]	-0.022
	TruthfulQA	0.655 [0.628, 0.680]	0.604 [0.573, 0.633]	-0.051
Math	MATH500	0.581 [0.539, 0.621]	0.600 [0.566, 0.630]	+0.019
	HumanEval	0.600 [0.572, 0.634]	0.694 [0.670, 0.725]	+0.093
	WinoGrande	0.604 [0.585, 0.627]	0.644 [0.620, 0.667]	+0.039
	TruthfulQA	0.593 [0.573, 0.619]	0.624 [0.605, 0.646]	+0.031

Table 1: RQ1 within-family scale comparison by domain. Entries are final-layer medians with interquartile ranges; the last column is the 7B minus 1.5B median shift.

first few layers and in the final exit from the plateau. Thus scaling does not translate the entire curve upward or downward. It changes when the model enters the high-δ_{rel} middle-layer range and how sharply it leaves that range; both changes depend on family and domain.

Takeaway. Parameter count alone is not a reliable explanation for GH. Scale should be reported together with model family, domain, and layer.

4.2 RQ2: Depth Dominates the Atlas

Layer depth is the most stable factor in the study. For Qwen2.5-7B-Base, Figure 2A shows a three-part trajectory: early layers depend on the input, middle layers form a long high-δ_{rel} plateau, and final layers drop sharply. Table 2 quantifies the same pattern. The middle median is almost identical across domains (0.995–0.996), while final medians spread from 0.549 on MATH500 to 0.676 on HumanEval. The final-minus-middle shift is negative for every domain, ranging from -0.319 on HumanEval to -0.446 on MATH500.

The embedding layer shows why the final drop is not just inherited from the input representation. MATH500, HumanEval, and WinoGrande begin with relatively high embedding-layer medians, whereas TruthfulQA begins much lower (0.310) and then rises to 0.604 at the final layer. In other words, the network can either sharpen tree-like structure relative to the embedding layer or weaken it, while still ending far below the middle-layer plateau. Across Figure 2B, every model–domain pair has a negative final-minus-middle shift, but the magnitude varies strongly. Llama and Qwen-Coder variants often show drops near -0.5, whereas Qwen2.5-Math-7B on HumanEval drops by only -0.29. Depth is therefore a robust effect, but the size of the final-layer drop still depends on the model and the domain.

This result also clarifies the role of the middle-layer plateau. Because the plateau appears across domains with very different prompt lengths and lexical distributions, it is unlikely to be a narrow artifact of one dataset. At the same time, the plateau does not make the statistic useless: final layers consistently leave the plateau and separate models and domains again. The simplest interpretation is computational rather than architectural. Middle layers maintain many competing relational constraints at once, making the resulting distance pattern less compatible with a single tree metric. Final layers then compress the representation toward the information needed by the next-token objective, and that compression can produce a more tree-like distance

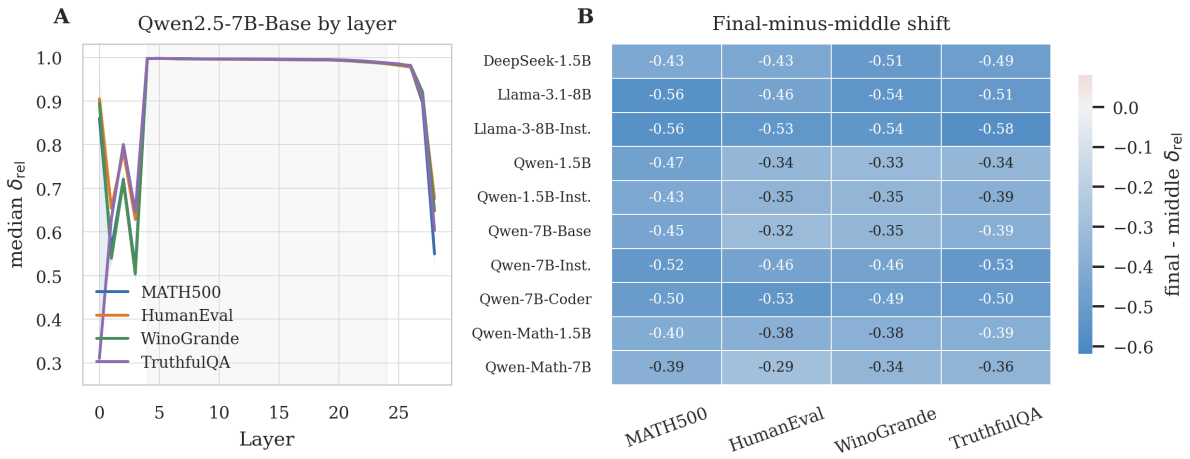

Figure 2: RQ2 depth effects. Panel A shows the layer trajectory for Qwen2.5-7B-Base. Panel B shows the median final-minus-middle shift for every model–domain pair. Negative values mean the final layer is more tree-like than the middle-layer median.

Dataset	Embedding	Middle	Final	Final–Middle
MATH500	0.860 [0.384, 0.908]	0.996 [0.995, 0.996]	0.549 [0.519, 0.591]	-0.446
HumanEval	0.905 [0.880, 0.962]	0.995 [0.995, 0.995]	0.676 [0.640, 0.712]	-0.319
WinoGrande	0.895 [0.876, 0.911]	0.995 [0.995, 0.996]	0.648 [0.624, 0.670]	-0.347
TruthfulQA	0.310 [0.296, 0.339]	0.996 [0.996, 0.996]	0.604 [0.573, 0.633]	-0.392

Table 2: RQ2 depth summary for Qwen2.5-7B-Base. Middle is the per-example median over relative layer positions 0.35–0.70.

structure.

Takeaway. The main geometric change happens across depth: most models move from a high-δ_{rel} middle layer to a lower-δ_{rel} final layer.

4.3 RQ3: Matched Families Occupy Different Regimes

At matched scale, family differences are larger than the within-family scale effects in RQ1. Figure 3 and Table 3 show that pooled matched-model medians range from 0.429 for Llama-3-8B-Instruct to 0.633 for Qwen2.5-Math-7B, with a Kruskal–Wallis test giving $p < 10^{-300}$. The cell-level spread is also substantial: TruthfulQA on Llama-3-8B-Instruct is 0.405, while HumanEval on Qwen2.5-Math-7B is 0.694. This range is much larger than the pooled scale shifts of -0.029 and +0.035.

The matched-family ordering is not merely "Llama versus Qwen." Llama models are consistently low-δ_{rel}, but Qwen variants separate according to specialization. Qwen2.5-7B-Base and Qwen2.5-Math-7B have the highest final-layer values, especially on HumanEval and WinoGrande. Qwen2.5-7B-Instruct lowers δ_{rel} across all domains, and Qwen2.5-7B-Coder-Instruct is closer to the Llama models than to Qwen2.5-7B-Base. Architecture, tokenizer, pretraining data, and post-training are mixed together here, so we should not assign the gap to one cause. The practical point is clear: models at the same approximate scale can organize prompt hidden states in measurably different ways.

The within-domain ordering strengthens this conclusion. On HumanEval, the gap between Qwen2.5-Math-7B and Qwen2.5-7B-Coder-Instruct is 0.256, far larger than any within-family scale shift in Table 1. On TruthfulQA, Llama-3-8B-Instruct reaches 0.405 while Qwen2.5-Math-7B remains at 0.624, again showing that matched-size families can end in very different final-layer geometries. The repeated pattern is that post-training and specialization can move a model into a different geometric range without changing its nominal scale class.

Takeaway. Matched-size models should not be treated as interchangeable in hidden-state geometry. Family and specialization can matter more than parameter count.

4.4 RQ4: Domains Interact with Specialization

Domain effects are visible even when the model is fixed. For Qwen2.5-7B-Base, Figure 4A and Ta-

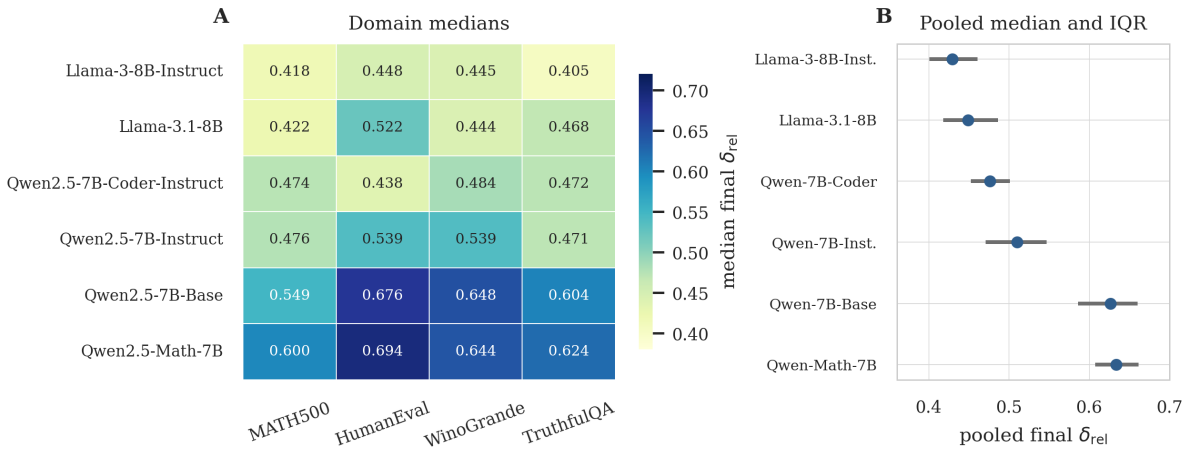

Figure 3: RQ3 matched 7/8B family comparison. Panel A reports final-layer domain medians; panel B summarizes the pooled final-layer median and IQR for each model.

Model	MATH500	HumanEval	WinoGrande	TruthfulQA
Llama-3.1-8B	0.422 [0.373, 0.482]	0.522 [0.435, 0.590]	0.444 [0.418, 0.471]	0.468 [0.435, 0.493]
Llama-3-8B-Instruct	0.418 [0.385, 0.473]	0.448 [0.388, 0.556]	0.445 [0.423, 0.468]	0.405 [0.377, 0.430]
Qwen2.5-7B-Base	0.549 [0.519, 0.591]	0.676 [0.640, 0.712]	0.648 [0.624, 0.670]	0.604 [0.573, 0.633]
Qwen2.5-7B-Instruct	0.476 [0.448, 0.500]	0.539 [0.506, 0.574]	0.539 [0.514, 0.568]	0.471 [0.441, 0.502]
Qwen2.5-7B-Coder-Instruct	0.474 [0.444, 0.500]	0.438 [0.417, 0.464]	0.484 [0.456, 0.515]	0.472 [0.457, 0.490]
Qwen2.5-Math-7B	0.600 [0.566, 0.630]	0.694 [0.670, 0.725]	0.644 [0.620, 0.667]	0.624 [0.605, 0.646]

Table 3: RQ3 matched-scale family summary. Each cell reports final-layer median δ_{rel} with interquartile range. Lower values indicate more tree-like metric structure.

ble 4 show a clear ordering: MATH500 is most tree-like (0.549), TruthfulQA is intermediate (0.604), WinoGrande is higher (0.648), and HumanEval is least tree-like (0.676). A Kruskal–Wallis test over domains gives $p = 1.5 \times 10^{-247}$. This ordering is not explained by sample size alone, because the HumanEval distribution remains high despite being the smallest dataset.

Specialization changes the domain ordering rather than shifting all domains by a constant amount. Instruction tuning lowers Qwen2.5-7B-Base fairly broadly, for example from 0.676 to 0.539 on HumanEval and from 0.604 to 0.471 on TruthfulQA. Coder specialization produces the sharpest interaction: HumanEval becomes the lowest-δ_{rel} domain for Qwen2.5-7B-Coder-Instruct (0.438), a -0.238 shift from the base model. By contrast, Qwen2.5-Math-7B is not more tree-like than the base model on MATH500; its MATH500 median is 0.600 versus 0.549 for Qwen2.5-7B-Base. Thus a domain label and a model specialization label do not map one-to-one onto lower GH. The strongest regularity is relational: code prompts become tree-like in the coder model, while math specialization keeps a higher-δ_{rel} final geometry across domains.

The domain effect is also visible in the full distributions. For Qwen2.5-7B-Base, HumanEval is not merely shifted upward by a few outliers; its quartiles sit above the MATH500 distribution, as shown in Figure 4A. For Qwen2.5-7B-Coder-Instruct, the same domain becomes the lowest-median condition, which means the reversal is not produced by dataset size or by a fixed property of HumanEval prompts. Instead, code specialization appears to make code-prompt final-layer distances more tree-like. This fits the intuition that code has nested structure, but the comparison with Qwen2.5-Math-7B shows that this structure appears clearly only in the relevant specialized model.

Prompt length is a possible confound because longer prompts provide more token states in each sample. We therefore bin examples by prompt-length quartile within each model and dataset as a control analysis. For Qwen2.5-7B-Base, HumanEval and WinoGrande remain higher than MATH500 across quartiles, and the within-domain slopes are modest. For Llama-3-8B-Instruct, medians are lower overall and still do not collapse into a single length-driven trend. Length can add local variation, but it does not explain the large family and specialization effects above.

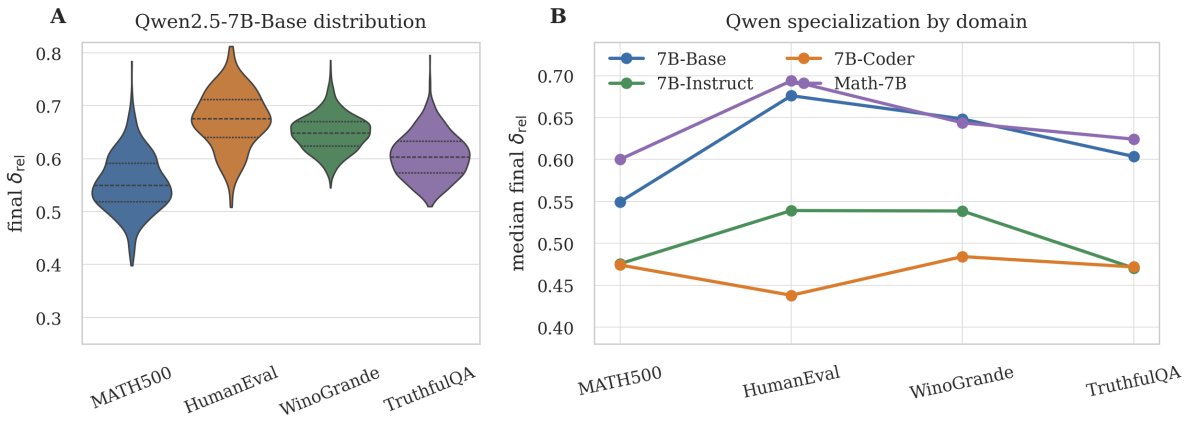

Figure 4: RQ4 domain effects. Panel A shows final-layer distributions for Qwen2.5-7B-Base. Panel B compares domain medians across four matched Qwen 7B variants.

Dataset	7B-Base	7B-Instruct	7B-Coder	Math-7B
MATH500	0.549 [0.519, 0.591]	0.476 [0.448, 0.500]	0.474 [0.444, 0.500]	0.600 [0.566, 0.630]
HumanEval	0.676 [0.640, 0.712]	0.539 [0.506, 0.574]	0.438 [0.417, 0.464]	0.694 [0.670, 0.725]
WinoGrande	0.648 [0.624, 0.670]	0.539 [0.514, 0.568]	0.484 [0.456, 0.515]	0.644 [0.620, 0.667]
TruthfulQA	0.604 [0.573, 0.633]	0.471 [0.441, 0.502]	0.472 [0.457, 0.490]	0.624 [0.605, 0.646]

Table 4: RQ4 domain summary for matched Qwen 7B variants. Each cell reports final-layer median δ_{rel} with interquartile range.

Takeaway. Domain effects are real, but they depend on the model. The same dataset can look less tree-like in one model and more tree-like in another.

4.5 Cross-RQ Patterns

Taken together, the four RQs reveal three regularities. First, hyperbolicity is primarily a trajectory rather than an endpoint. The final layer is important because it is closest to the prediction interface, but the most repeatable phenomenon is the movement from early variability to a high-δ_{rel} middle plateau and then to a model-specific final drop. This is why scale and family comparisons are most interpretable when paired with layer profiles.

Second, specialization changes which domains receive a tree-like final organization. The coder model does not simply lower all medians equally; it selectively makes HumanEval the lowest-δ_{rel} domain among the Qwen variants. Instruction tuning produces a broader reduction, while math specialization keeps relatively high final-layer δ_{rel} even on MATH500. This pattern suggests that lower hyperbolicity is not a generic marker of task difficulty or formal structure. It is better understood as an alignment between a model's learned representational organization and the prompt distribution.

Third, family differences at matched scale dominate simple parameter-count explanations. The Llama models and Qwen-Coder have lower final-layer values than Qwen2.5-7B-Base and Qwen2.5-Math-7B, even though all are in the 7/8B range. Thus, a useful hyperbolicity analysis should report at least the model family, layer range, and data domain; one number per model loses the main pattern.

5 Discussion

A depth-centered view of hidden-state geometry. The most robust pattern is the three-part layer trajectory. Embedding and early layers are sensitive to domain and tokenization. Middle layers are close to the upper end of the normalized statistic, suggesting a highly non-tree-like organization under the four-point metric. Final layers then often move toward lower δ_{rel}, but the degree of movement depends on family and domain. This supports a depth-centered view: hyperbolicity is not a fixed property of a model, but a trajectory through the computation.

Why might middle layers be high-δ_{rel}? One interpretation is that middle layers mix multiple incompatible organizational principles. They must preserve lexical identity, local syntax, long-range context, and task cues before the model has fully compressed the prompt into a prediction-oriented representation. A collection of token states that si-

multaneously reflects many such axes may be less tree-like than either the input embedding space or the final task-facing representation. The final-layer drop can then be read as compression toward a simpler distance pattern, although the present measurements do not identify the mechanism causally.

Specialization changes geometry. The Qwen variants show that specialization is not merely a performance label. Coder specialization strongly lowers HumanEval δ_{rel}, while math specialization does not make every domain more tree-like and can even raise final-layer δ_{rel} relative to the base track. This suggests that training or adaptation changes which domains receive a tree-like final organization. The same domain can therefore have simple distances for one model and more complex distances for another.

Interpreting low hyperbolicity. Low δ_{rel} should not be read as direct evidence of an explicit symbolic tree. It only states that sampled distances among prompt-token hidden states are closer to a tree metric. The statistic is nevertheless useful because it depends only on distances, is cheap to summarize as a distribution, and is comparable across models. It complements probing, intrinsic dimension, anisotropy, and causal interventions.

What the map explains. The GH map is most informative when it is used comparatively. A low final-layer value for one model does not by itself say whether the model is better at a task, but a systematic reversal across specializations can reveal how training changes the geometry assigned to a domain. For example, HumanEval is high-δ_{rel} for Qwen2.5-7B-Base and Qwen2.5-Math-7B, but low-δ_{rel} for Qwen2.5-7B-Coder-Instruct. The important observation is not that code is intrinsically tree-like, but that code prompts become tree-like in a model whose training history makes code structure central. Similarly, the middle-layer plateau is not a performance claim; it is a repeatable part of the layer trajectory where token distances are far from a single tree metric before the final layers separate again.

Implications for future analyses. These results suggest two practical reporting norms for geometric studies of LLM hidden states. First, final-layer summaries should be accompanied by at least a coarse depth profile, because middle and final layers answer different questions. Second, model comparisons should be matched by family, scale, and domain whenever possible. Otherwise, a domain effect can be mistaken for specialization, and a specialization effect can be mistaken for scale. Hyperbolicity is therefore better used as one coordinate in a larger representation-analysis suite than as a standalone score.

6 Related Work

Hyperbolic representation learning. Hyperbolic representation learning encodes hierarchical structure with low distortion (Nickel and Kiela, 2017, 2018) and supports neural operations on hyperbolic manifolds (Ganea et al., 2018). In NLP, it has been applied to word and sentence embeddings (Dhingra et al., 2018; Tifrea et al., 2018) and to multilingual or multimodal representation design (Sawhney et al., 2024). We instead ask whether standard LLM hidden states already exhibit tree-like metric structure.

Geometry of language-model representations. Prior work studies anisotropy and layerwise geometry (Ethayarajh, 2019), syntactic structure in hidden states (Hewitt and Manning, 2019), and intrinsic-dimension or neighborhood variation across layers and domains (Valeriani et al., 2023). These analyses show that hidden states are structured, but do not directly test whether token-state distances behave like a hierarchy.

Hyperbolic and geometric analysis for LLMs. Recent LLM work studies hyperbolic embeddings and adaptation (Yang et al., 2026), mixture-of-curvature models (He et al., 2026), broader hyperbolic methods (Patil et al., 2025), and geometry-based adversarial-prompt detection (Yung et al., 2025). Most of this work introduces or exploits geometry, whereas we use Gromov hyperbolicity as a diagnostic on fixed models.

7 Conclusion

We presented a prompt-only hyperbolicity atlas for LLM hidden states across ten models, four datasets, and all layers. Normalized Gromov hyperbolicity is highly structured: depth dominates, matched families differ, domains interact with specialization, and scale alone is weak and non-monotonic. Hyperbolicity is therefore a useful diagnostic when interpreted as a model–layer–domain property rather than as a direct claim about symbolic hierarchy.

Limitations

This study is restricted to prompt-token hidden states; generated responses, longer contexts, and different decoding settings may yield different patterns. The statistic depends on tokenization, quadruple sampling, and distance choice (we use Euclidean distance with diameter normalization; cosine distances and alternative normalizers remain unexplored). A fixed sampling budget keeps comparisons fair but may introduce estimation variance for very short or very long prompts. Low δ_{rel} should not be read as proof of an explicit symbolic tree; it is a metric proxy for tree-likeness and should be combined with probing and causal analyses. Direct links between GH, task accuracy, and geometry-changing interventions are left to future work.

References

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. 2018. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 59–69.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic neural networks. *Advances in neural information processing systems*, 31.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mikhael Gromov. 1987. Hyperbolic groups. In *Essays in group theory*, pages 75–263. Springer.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Neil He, Rishabh Anand, Hiren Madhu, Ali Maatouk, Smita Krishnaswamy, Leandros Tassiulas, Menglin Yang, and Rex Ying. 2026. Helm: Hyperbolic large language models via mixture-of-curvature experts. *Advances in Neural Information Processing Systems*, 38:142604–142635.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30.

Maximillian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR.

Sarang Patil, Zeyong Zhang, Yiran Huang, Tengfei Ma, and Mengjia Xu. 2025. Hyperbolic large language models. *arXiv preprint arXiv:2509.05757*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Ramit Sawhney, Shrey Pandit, Vishwa Shah, Megh Thakkar, and Shafiq Joty. 2024. Adapt: A set of guidelines for hyperbolic multimodal multilingual nlp. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1757–1771.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. 2018. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2023. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Menglin Yang, Aosong Feng, Bo Xiong, Jiahong Liu, Irwin King, Rex Ying, and 1 others. 2026. Hyperbolic fine-tuning for large language models. *Advances in Neural Information Processing Systems*, 38:29613–29648.

Canaan Yung, Hanxun Huang, Christopher Leckie, and Sarah Erfani. 2025. Geometry-guided adversarial prompt detection via curvature and local intrinsic dimension. *arXiv preprint arXiv:2503.03502*.