LINGUISTIC CARTOGRAPHY: METRIC SPATIAL AND HIERARCHICAL PLATIAL REPRESENTATIONS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can encode where a place lies on the Earth's surface, but geographic representation is not only about the coordinate-based metric space. A city is also nested in a hierarchical platial structure: state/province, country, subregion and continent. We study how these two geographic forms coexist in the neural representations of language models. We build GPH-30K, a primary source of geographic hierarchy with 30,000 globally distributed populated places with both latitude/longitude information and a complete administrative path. This benchmark dataset supports metric spatial experiments and hierarchical platial experiments. On an open-weight Llama/Qwen model panel, coordinate probes recover latitude/longitude while hierarchy probes recover platial path from the same residual stream. Layer-wise parent-recovery maps show that the place hierarchy is not a single flat attribute: country-to-subregion relations emerge earlier, admin1-to-country relations peak late, and place-to-admin1 remains the hardest edge. Final-layer embedding geometry also varies by model family and generation: Qwen-2.5 models are consistently more spatial metric-leaning, whereas Llama and Qwen3 are closer to balanced or platial hierarchy-leaning. Finally, on a conflict diagnostic set HGCD where coordinate distance and administrative-tree distance make opposite predictions, smaller Qwen/Llama checkpoints are mostly spatial, while large Llama and Qwen-2.5 checkpoints can become mixed or platial. These results support a coexistence view of linguistic cartography: language models contain both continuous metric spatial representations and discrete hierarchical platial representations, which emerge at different depths and dominate geometry under different contexts. This work offers insights into the neural representations of our physical world in LLMs toward geospatial intelligence.

1 INTRODUCTION

A growing body of evidence suggests that language models trained only on text develop coherent internal representations of the world they describe (Huh et al., 2024). On geography in particular, linear probes recover latitudes and longitudes from residual-stream activations with high accuracy (Gurnee & Tegmark, 2024), suggesting that language models carry map-like metric information as a linear "world model". But geographic representation is not only about the metric space. A place can be represented by coordinates, by linguistic co-occurrence patterns, by a regional hierarchy such as the "country–subregion–continent" label, or by a prompt-specific answer format. These views often agree in ordinary data, but they may disagree across borders and enclaves. We investigate how language models represent both *metric spatial* structures and *hierarchical platial* structures, when each representation becomes available internally, and which one is emphasized when the two structures disagree.

The distinction between *space* and *place* roots in geography (Tuan, 1977; Goodchild, 2011); we use *spatial* for the geodesic distance-based metric structure and *platial* for the place-based hierarchical structure (Gao, 2022). A metric spatial representation organizes locations by coordinate proximity: e.g., San Diego (U.S.) and Tijuana (Mexico) are spatially close despite across the international border. A hierarchical platial representation organizes the same place entities by their membership relations: San Diego is part-of/within "California–United States–Northern America". An interesting

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.

platial reasoning arises: Place A is nearer to B than is C, if A and B are in the same higher-level place hierarchy but C is not (Goodchild, 2015). Platial closeness doesn't necessarily indicate spatial proximity (such as the example of San Diego-Los Angeles-Tijuana in Figure 1). Prior work shows that the metric spatial structure (coordinates) is linearly recoverable from hidden states in language models (Gurnee & Tegmark, 2024). We ask two fundamental research questions: **Is the hierarchical platial structure also linearly recoverable from the same geographic entities? How do the spatial and platial structures coexist (if they do coexist) in large language models (LLMs)?**

Prior work shows that the hidden states in LLMs encode world structure and recover geographic coordinates with linear probes, while geospatial NLP studies document both geographic competence and systematic regional distortions in model outputs (Huh et al., 2024; Gurnee & Tegmark, 2024; Manvi et al., 2024b;a; Fulman et al., 2024; Decoupes et al., 2025). GIScience and cognitive-map research, however, emphasize that geographic knowledge is not purely metric: places are also organized by borders, regions, and part-whole hierarchies (Tuan, 1977; Goodchild, 2011; Stevens & Coupe, 1978; Hirtle & Jonides, 1985). We connect these theoretical perspectives with linear probing, parent-retrieval maps, and representational-similarity analyses to test whether metric spatial and hierarchical platial neural representations coexist in the same residual stream. Appendix B.1 gives a detailed discussion of related literature.

Following the proposed questions, our research design consists of two parts. First, to verify whether LLMs recover the platial structure at the global scale, we build a benchmark dataset GPH-30K from primary geospatial sources: 30,000 global populated places nested under a hierarchy of **first-order administrative units (admin1)**, **countries**, **United Nations subregions**, and **continents**. Because every record has both metric coordinates (latitude/longitude) and platial hierarchy labels, it enables us to test whether metric coordinates and platial hierarchy labels are both linearly available across different types of places rather than only in a small diagnostic set. Second, we use HGCD (Hierarchical-Geospatial Conflict Diagnostic), a curated conflict location set where coordinate proximity and administrative membership deliberately disagree. For example, near-border place pairs that are physically close but hierarchically separated, to test which structure LLMs emphasize.

This research offers four major contributions:

- We construct two complementary geographic benchmarks: GPH-30K, a 30,000-location global place hierarchy with coordinates and administrative paths, and HGCD, a conflict diagnostic set where metric proximity and administrative membership deliberately disagree.
- Metric spatial coordinates and hierarchical platial labels are both available at global scale in the same residual stream of language models.
- Hierarchy levels are not recovered uniformly: different tree edges in the place hierarchy emerge at different depths and require different linear rank.
- When the metric spatial and the hierarchical platial structures disagree, different model series and model sizes show different preferences: smaller LLMs are mostly metric spatial on HGCD, while some large LLMs become mixed or platial. The target is therefore not a binary choice between "metric spatial" and "hierarchical platial" representations, but an account of how both structures coexist and shift across model series.

2 METHODS

2.1 DATASETS: COVERAGE AND CONFLICT

We build two complementary datasets. A broad-coverage dataset is GPH-30K, a geographic place hierarchy set reconstructed from global populated-place records: GeoNames country/admin1 metadata, Natural Earth admin metadata, and REST Countries regional metadata (GeoNames, 2026; Natural Earth, 2026; REST Countries, 2026). It contains 30,000 locations with coordinates, admin1 labels, countries, United Nations subregions, and continents, organized as the platial hierarchy:

$$\text{Earth} \rightarrow \text{continent} \rightarrow \text{subregion} \rightarrow \text{country} \rightarrow \text{admin1} \rightarrow \text{place}.$$

GPH-30K covers 226 countries and territories, 3,624 admin1 units, 24 subregions, and all six populated continents. We select from 68,277 eligible GeoNames populated places using a coverage-preserving allocation: every represented country receives a minimum quota when possible, additional examples follow a square-root country-size allocation, and a soft maximum country share prevents a few large countries from dominating the dataset. We also require globally unique bare place

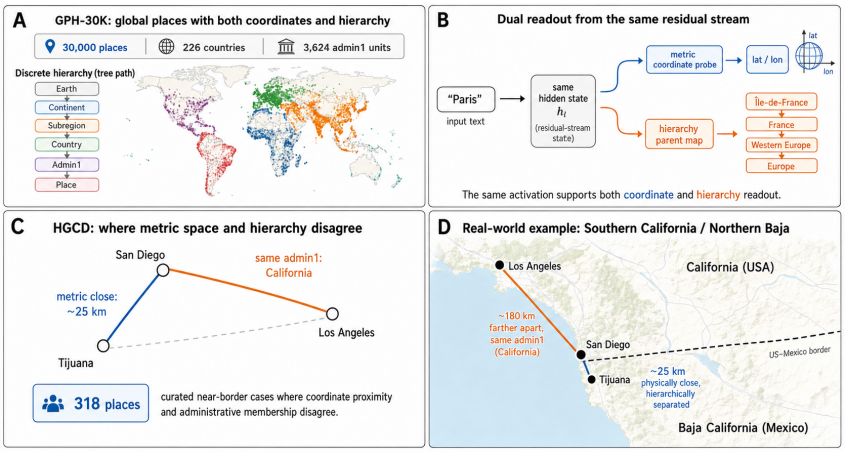

Figure 1: Overview of the Metric Spatial–Hierarchical Platial coexistence problem. **(B)** The same residual-stream activation supports two readouts: a metric coordinate probe for latitude/longitude and hierarchy parent maps for admin1-to-continent structure. **(C)** HGCD isolates conflict cases where metric proximity and administrative hierarchy disagree. **(D)** The San Diego-Los Angeles–Tijuana- example makes the conflict concrete: San Diego is physically closer to cross-border Tijuana but administratively closer to Los Angeles.

names, valid coordinates, complete country/admin1 metadata, and complete continent/subregion labels. This construction keeps the dataset large while making bare-prompt experiments auditable and geographically balanced enough for place hierarchy probes.

HGCD is a spatial/platial conflict diagnostic dataset. It is a curated set of place pairs drawn from GeoNames and OpenStreetMap (GeoNames, 2026; OpenStreetMap contributors, 2026) where metric space and administrative place predictions deliberately diverge. Conflict pairs include places less than ten kilometres apart that lie in different countries or states—Baarle-Hertog in Belgium and Baarle-Nassau in the Netherlands—and places more than a hundred kilometres apart inside one country or state. HGCD spans 318 unique cities across 69 countries and all six populated continents. We use GPH-30K to test whether both spatial/platial neural representations are available in different language models, and HGCD to ask what happens when they disagree.

2.2 MODELS AND PROMPTS

The broad-coverage experiments use open-weight Llama/Qwen checkpoints, in total 15 models spanning family, generation, and scale: Qwen-2.5-0.5B, 1.5B, 3B, 7B, 14B, 32B, and 72B; Qwen3-4B, 8B, 14B, and 32B; Llama-3.2-3B, Llama-3.1-8B, Llama-3.1-70B, and Llama-2-70B (Touvron et al., 2023; Grattafiori et al., 2024; Yang et al., 2024; 2025). This panel keeps the main GPH-30K claims in language model families that are both widely used and recent. Additional Pythia, OPT, and Mistral checkpoints are reported as external-family robustness checks in Appendix C.3. The main GPH-30K experiments use bare place-name prompts: `"{place}"`. Hierarchy-edge recovery also uses parent-free node prompts: a child node is represented by its own place name, not by a template that includes its parent. The HGCD conflict diagnostic uses the same bare-prompt representation so that the disagreement test measures hidden-state geometry rather than a prompt-specific output format.

2.3 CORE MEASUREMENTS

We use three complementary measurements, each answering a different question about a hidden state. *Metric probes* ask whether latitude and longitude are linearly decodable from the residual stream of language models. *Hierarchy probes* ask whether parent labels in the administrative place tree are linearly decodable from the same residuals. *Linear hierarchy maps* ask whether a learned linear transformation can map a child-node place representation to its parent-node place representation, so that nearest-neighbour retrieval recovers the correct parent node? Finally, *distance-matrix statistics* address whether the geometry of the embedding is shaped more by metric distance or by place

hierarchical-tree distance. We avoid making decoder-token readouts part of the main evidence, because tokenization and unembedding-basis choices are not needed for the hierarchy claim.

Let $\mathbf{D}_{\text{emb}}^{(\ell)}$ denote the pairwise cosine distance at layer ℓ, $\mathbf{D}_{\text{spatial}}$ the log-geodesic distance matrix, and $\mathbf{D}_{\text{platial}}$ the hierarchical-platial tree-path distance matrix. We report the platial/spatial alignment of LLMs' internal representations:

$$\begin{aligned}\Delta r^{(\ell)} = &\operatorname{corr}(\mathbf{D}_{\text{emb}}^{(\ell)}, \mathbf{D}_{\text{platial}} \mid \mathbf{D}_{\text{spatial}}) \\ &- \operatorname{corr}(\mathbf{D}_{\text{emb}}^{(\ell)}, \mathbf{D}_{\text{spatial}} \mid \mathbf{D}_{\text{platial}})\end{aligned} \tag{1}$$

using Spearman partial correlations: negative Δr means the embedding geometry is more aligned with metric distance than administrative hierarchy at that layer, positive means more aligned with hierarchy. We complement Δr with the distance-based redundancy analysis (dbRDA) ΔR^2, which partitions variance at the location level. Since pair-level Δr suffers from non-independence (Legendre & Fortin, 2010; Guillot & Rousset, 2013), we report bootstrap CIs over locations rather than over pairs, and we rely on location-level dbRDA together with instance-level probing and parent-recovery tools (Legendre & Anderson, 1999; Belinkov, 2022; Hewitt & Liang, 2019) when forming claims.

To compare neural representational shapes across language models or prompts, we compute the Pearson correlation of the upper-triangular pairwise-cosine vectors of two distance matrices:

$$r_{\text{RSA}}(\mathbf{D}_A, \mathbf{D}_B) = \text{Pearson}\big(\text{triu}(\mathbf{D}_A),\, \text{triu}(\mathbf{D}_B)\big). \tag{2}$$

Here triu denotes the strict upper triangle, excluding the diagonal. This is the standard Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008); among the alternatives (SVCCA (Raghu et al., 2017), PWCCA (Morcos et al., 2018), CKA (Kornblith et al., 2019)), it has the practical advantage of not requiring matched hidden dimensionality.

For the subspace intervention, we use iterative null-space projection (INLP) (Ravfogel et al., 2020) rather than decoder-token or unembedding directions. The metric-spatial subspace is fit by repeatedly training held-out latitude/longitude ridge probes and removing their coefficient directions. The hierarchical-platial subspace is fit by repeatedly training a one-vs-rest country classifier and removing leading right singular directions of the classifier weight matrix. We then project place, admin1, and country residuals orthogonally to the learned subspace and re-evaluate coordinate probes, hierarchy-label probes, and admin1→country parent-disjoint recovery. A rank-matched random subspace calibrates the effect of removing uninformative directions.

The PPMI–SVD baselines turn the corpus-statistical explanation into falsifiable tests. If metric spatial or hierarchical platial structure is already latent in text co-occurrence, then shallow distributional embeddings built without any transformer should recover part of the same structure. We use two baselines. On HGCD, Wikipedia-outlink pair indicators give an observed place-place co-occurrence graph (Appendix D.1). On GPH-30K, a full-coverage place-context matrix uses explicit geographic descriptors—continent, subregion, country, admin1, timezone, and feature type—but no coordinate tokens (Appendix D.3). This second baseline is intentionally strong: it tests whether hierarchy metadata alone is sufficient to induce approximate metric structure, not whether every pretraining mention supplies the same fields. In both cases we compute PPMI with context smoothing (Levy et al., 2015), take the top singular vectors scaled by $\sqrt{\sigma}$, and run the same probe and distance-geometry analyses. For mediation, we add the resulting corpus distance $\mathbf{D}_{\text{cooc}}$ to the distance regression $\mathbf{D}_{\text{emb}} \sim \mathbf{D}_{\text{spatial}} + \mathbf{D}_{\text{platial}} + D_{\text{cooc}}$ (Appendix D.4).

2.4 EVIDENCE HIERARCHY

The main text follows the argument in the order needed for our central claim. First, we use GPH-30K to show that metric coordinates and administrative parents of places are both recoverable at scale. Second, a layer sweep asks when each level of the hierarchical-place tree becomes available in relative depth. Third, we test whether final-layer pairwise geometry retains the administrative place tree even when coordinate probes succeed. Fourth, we use a decoder-independent hierarchy-recovery experiment: a linear map from child-node hidden states must retrieve the correct parent-node hidden state. Finally, HGCD is used only for its intended role: a conflict diagnostic where metric distance and administrative-tree distance disagree.

The corpus-statistical hypothesis is evaluated alongside this hierarchy: before attributing either geographic view to a special neural mechanism, we ask whether low-order co-occurrence already

contains the relevant signal and whether language models make metric spatial and hierarchical platial information jointly available in the same residual stream, recover hierarchy edges as in-support relations, and learn a transferable platial relation that generalizes to unseen parents within context.

3 EXPERIMENT RESULTS: METRIC SPATIAL AND HIERARCHICAL PLATIAL REPRESENTATIONS

We first ask which geographic information is linearly present in the residual stream. Unlike a pure coordinate-probing study, every location here has both a metric target and a place in an administrative hierarchical tree. This lets us evaluate spatial/platial representation coexistence directly: latitude/longitude measure continuous metric space, while admin1, country, subregion, and continent labels measure discrete hierarchical platial structure. HGCD is introduced after the large-scale test on GPH-30K, to ask what happens when the two views make opposing predictions.

3.1 BROAD COVERAGE: METRIC SPATIAL AND HIERARCHICAL PLATIAL VIEWS ARE BOTH AVAILABLE

For each model in the open-weight Llama/Qwen model panel, we extract last-token residuals for 30,000 locations in GPH-30K under the bare prompt and evaluate five final-layer probes: ridge regression for latitude/longitude, and linear classifiers for continent, subregion, country, and admin1, respectively. One novelty of this work lies in the comparison between the spatial metric and the platial hierarchy in neural representations of LLMs. Table 1 shows that the coarse and middle platial hierarchy levels are clearly available at scale, while the finest level is much weaker. Llama-3.1-70B is the strongest broad model on the coordinate and admin1 probes (coord $R^2 = 0.812$, continent 0.902, admin1 0.260), and Llama-2-70B is nearly tied while slightly leading subregion (0.835) and country recovery (0.675). Qwen-2.5 improves substantially from 0.5B to 14B/72B (coord $R^2 = 0.510 \rightarrow 0.766/0.783$), although the 32B and 72B points are not monotone on every label. Qwen3 also improves with scale, but Qwen3-32B (coord $R^2 = 0.734$, country 0.599) remains below the large Qwen-2.5 and Llama checkpoints on broad probes. Across the full series, continent and subregion accuracy are robustly above chance, country accuracy reaches the 0.67 range for the strongest models, and admin1 remains the weakest tier. We therefore treat the admin1 probe as partially recoverable rather than as a strong recovery result like that in other higher-level place levels, and we let coarser tiers (continent, subregion, country) carry the headline coexistence claim.

Table 1: Broad-coverage metric-spatial and hierarchical-platial probes on GPH-30K locations. All entries use final-layer bare-prompt residuals. Coord R^2 is the mean latitude/longitude ridge-regression R^2; Cont., Subreg., Country, and Admin1 are linear-probe accuracies for administrative hierarchy labels. Blocks separate small and large checkpoints; arrows indicate whether larger values are better.

Small models (0.5B–8B)						**Large models (14B–72B/70B)**					
Model	Coord R^2 ↑	Cont. ↑	Subreg. ↑	Country ↑	Admin1 ↑	Model	Coord R^2 ↑	Cont. ↑	Subreg. ↑	Country ↑	Admin1 ↑
Qwen-2.5-0.5B	0.510	0.740	0.598	0.436	0.153	Qwen-2.5-14B	0.766	0.874	0.793	0.624	0.236
Qwen-2.5-1.5B	0.646	0.812	0.696	0.519	0.193	Qwen-2.5-32B	0.762	0.870	0.786	0.612	0.234
Qwen-2.5-3B	0.698	0.838	0.735	0.553	0.209	Qwen-2.5-72B	0.783	0.878	0.796	0.623	0.240
Qwen-2.5-7B	0.735	0.854	0.760	0.581	0.225	Qwen3-14B	0.724	0.851	0.762	0.592	0.226
Qwen3-4B	0.626	0.808	0.698	0.532	0.200	Qwen3-32B	0.734	0.856	0.764	0.599	0.228
Qwen3-8B	0.690	0.838	0.740	0.573	0.217	Llama-3.1-70B	**0.812**	**0.902**	0.833	0.672	**0.260**
Llama-3.2-3B	0.767	0.873	0.800	0.634	0.241	Llama-2-70B	0.801	0.899	**0.835**	**0.675**	0.257
Llama-3.1-8B	**0.800**	**0.895**	**0.829**	**0.672**	**0.257**						

3.2 HIERARCHY EDGES EMERGE AT DIFFERENT DEPTHS

Final-layer probes establish the availability of spatial/platial neural representations, but flat label probes are not the right tool for identifying place hierarchy levels: continent, subregion, country, and admin1 labels are nested, so a single strengthening place-identity signal can make all four curves rise together. We therefore use the edge-conditioned parent recovery. For each edge in the GPH-30K platial tree, we train a linear map from child-node residuals to parent-node residuals and retrieve the parent by cosine similarity. To remove easy coarse cues, we also score each edge within the relevant context: place→admin1 among admin1 units in the same country, admin1→country among countries in the same subregion, and country→subregion among subregions in the same continent. Node prompts are parent-free (e.g., an admin1 node is prompted by its own name, not "admin1, country"), preventing the parent label from being copied from the prompt text.

Skipped Ancestor Prediction (SAP) versus Chained Hierarchy Prediction (CHP). The adjacent-edge experiment is intentionally strict: it asks whether each local tree relation can be recovered as a child-to-parent map. We also test non-adjacent ancestor recovery to distinguish local-edge chaining from skipped prediction to coarse ancestors. Figures 2–3 visualize the layer-wise place ancestor experiment, and Table 6 (Appendix C.2) shows the corresponding summary statistics. SAP maps are strong across the Qwen/Llama model series. Across all full-series checkpoints, SAP place→country recovery reaches median Cond.@1 0.649 [0.448, 0.792] against by chance 0.124, with the large-model subset reaching 0.689 [0.480, 0.792]. Coarser place→subregion and place→continent maps are even stronger. However, the CHP path, place→admin1→country, is consistently weaker (full-series median 0.581, large-subset median 0.632), although still far above chance. It suggests that non-adjacent administrative levels are directly connected in the residual representation rather than only recoverable through consecutive hierarchy edges.

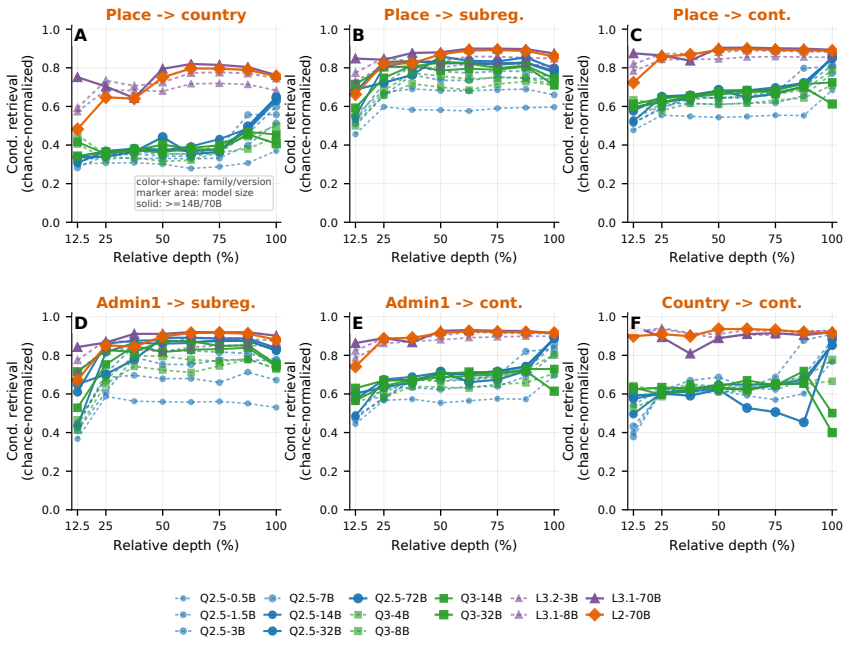

Figure 2: Layer-wise non-adjacent ancestor recovery on GPH-30K. Each panel shows one skipped child-to-ancestor map using chance-normalized context-conditioned retrieval across the Qwen/Llama series; solid lines are the large checkpoints and thin dashed lines are the small/mid checkpoints. Skipped ancestor information is available across hierarchical tree depth and differs by edge type: place→subregion and admin1→subregion emerge early, while place→country and place→continent become strongest late for several Qwen models.

What the linear map generalizes to. The Cond.@1 column in Table 3 uses stratified 5-fold cross-validation, so each parent label can appear in both the train and the test fold. The right-most column reports the *parent-disjoint* retrieval test: we hold parents out at training time using `GroupKFold` on the parent label, so the held-out parent appears only as a within-context retrieval candidate. The parent-disjoint result is mixed and tightens what the relational claim should say. Admin1→country is the one edge where the linear map clearly generalizes to held-out countries across the Qwen/Llama series. In the small/mid checkpoints, parent-disjoint Cond.@1 is 0.169 [0.140, 0.214] against a chance of 0.112; in the large checkpoints, the corresponding median is 0.150 [0.114, 0.185]. A split-seed confidence interval on the per-seed-mean accuracy across five `GroupKFold` country-partition seeds per large model shows the strongest new generalization in Qwen-2.5-32B/72B (0.198 ± 0.007 and 0.199 ± 0.009), while Llama-3.1-70B and Llama-2-70B remain above chance but weaker (0.148 ± 0.005 and 0.135 ± 0.007); Appendix Table 16 reports the full large-model split summary. As a causal check, on three representative Qwen/Llama models we use iterative null-space projection (Ravfogel et al., 2020) to remove model-specific metric-spatial subspaces until the held-out coordinate signal is nearly gone (Table 2). The required ranks are large ($k = 406, 250, 290$), and final 5-fold coordinate recovery drops from $+0.690$–$+0.800$ to negative R^2 (-0.131–-0.061). Despite this, country accuracy changes only from 0.581→0.562, 0.574→0.563, and 0.673→0.672; admin1 accuracy is unchanged at three decimals; and the admin1→country parent-disjoint map is preserved

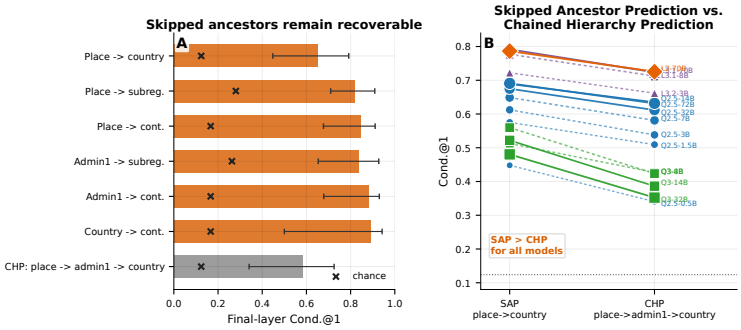

Figure 3: Final-layer skipped ancestor prediction and chained hierarchy prediction. **(A)** Median Cond.@1 across the Qwen/Llama series, with bars spanning the model range and crosses marking chance. SAP maps are well above chance; the gray row is the CHP place→admin1→country path. **(B)** Per-model paired comparison for place→country. Every checkpoint has higher SAP recovery than CHP recovery, showing that country information is readable from place states and not solely obtained by chaining adjacent hierarchy maps.

Table 2: Selective INLP erasure on three representative Llama/Qwen checkpoints. Coord R^2 is the mean held-out latitude/longitude ridge-regression R^2; Country and Admin1 are 5-fold classifier accuracies. A1→C PD is admin1→country parent-disjoint Cond.@1 averaged over three `GroupKFold` country-partition seeds. INLP-metric and INLP-hier remove learned metric or country-classifier subspaces; random is rank-matched.

Model	Erasure	Coord R^2 ↑	Country ↑	Admin1 ↑	A1→C PD ↑
Qwen-2.5-7B	none	+0.735	0.581	0.224	0.208
	INLP-metric (k=406)	−0.061	0.562	0.224	0.219
	INLP-hier (k=240)	+0.649	0.519	0.209	0.217
	random (k=406)	+0.734	0.576	0.221	0.211
Qwen3-8B	none	+0.690	0.574	0.216	0.174
	INLP-metric (k=250)	−0.131	0.563	0.216	0.171
	INLP-hier (k=240)	+0.607	0.507	0.209	0.179
	random (k=250)	+0.689	0.572	0.216	0.177
Llama-3.1-8B	none	+0.800	0.673	0.259	0.153
	INLP-metric (k=290)	−0.073	0.672	0.259	0.172
	INLP-hier (k=240)	+0.754	0.605	0.248	0.147
	random (k=290)	+0.798	0.673	0.260	0.153

(0.208→0.219, 0.174→0.171, 0.153→0.172). A rank-matched random subspace has no comparable effect. The surviving admin1→country relational claim is therefore not a downstream consequence of linear coordinate predictability. The reverse direction is intentionally reported more narrowly: a deeper rank-240 country-INLP control weakens country accuracy by about 10–12% relative but does not erase it, so we do not claim a clean bidirectional double dissociation.

For place→admin1, parent-disjoint accuracy falls back to the chance level in both groups (0.066 and 0.061 vs. chance 0.071), so the random-CV recovery on this edge is largely parent-label memorization rather than a transferable relation. Country→subregion also sits below chance under the strict parent-out split (0.139 and 0.128 vs. chance 0.263), which we read as the country-to-subregion linear map being too low-capacity for the 24-subregion candidate set once the held-out subregion is unseen at training time, rather than as evidence against any country→subregion structure. Subregion→continent is unstable under parent-out because continent-out is almost the same as candidate-out for that edge. Appendix E.1 reports the per-model parent-disjoint and continent-held-out tables that this column summarizes. The relational hierarchy claim is therefore sharpest at admin1→country, where it survives parent-disjoint generalization, and is bounded to in-support recovery for the other tree edges.

3.3 GPH-30K: SPATIAL AND PLATIAL STRUCTURES COEXIST AT LARGE

The probe results above show that coordinates and parent labels of places are linearly recoverable. We then ask a related question: whether the *global geometry* of the same final-layer residuals is organized more by metric distance or by platial tree distance. On a stratified sample of 5,000 locations from GPH-30K, we sample one million location pairs per model and compare three pairwise

Table 3: Parent-conditioned hierarchy-edge recovery on GPH-30K across the Qwen/Llama series. Rows report median [min–max] separately for small/mid and large checkpoints. Cond.@1 retrieves within the relevant parent context; Global@1 retrieves among all parent candidates at that tree depth. Emergence is the first sampled relative depth reaching 90% of the model's best chance-normalized conditional retrieval. Parent-out @1 holds parent labels out during training using `GroupKFold`; arrows indicate whether larger or smaller values are better.

Edge	Models	Cond.@1 ↑	Global@1 ↑	Chance ↓	Emergence ↓	Parent-out @1 ↑
Place→admin1	0.5B–8B	0.136 [0.098–0.198]	0.080 [0.026–0.157]	0.071	25% [11%–100%]	0.066 [0.058–0.075]
Place→admin1	14B–72B/70B	0.164 [0.136–0.222]	0.108 [0.083–0.179]	0.071	62% [12%–88%]	0.061 [0.058–0.068]
Admin1→country	0.5B–8B	0.612 [0.448–0.796]	0.511 [0.305–0.757]	0.112	100% [25%–100%]	**0.169 [0.140–0.214]**
Admin1→country	14B–72B/70B	0.706 [0.592–0.820]	0.634 [0.491–0.780]	0.112	100% [12%–100%]	**0.150 [0.114–0.185]**
Country→subreg.	0.5B–8B	0.737 [0.664–0.845]	0.666 [0.562–0.814]	0.263	25% [12%–75%]	0.139 [0.111–0.155]
Country→subreg.	14B–72B/70B	0.735 [0.646–0.832]	0.686 [0.540–0.810]	0.263	25% [12%–38%]	0.128 [0.075–0.217]
Subreg.→cont.	0.5B–8B	0.771 [0.750–0.875]	0.771 [0.750–0.875]	0.167	52% [11%–100%]	0.000 [0.000–0.208]
Subreg.→cont.	14B–72B/70B	0.750 [0.458–0.792]	0.750 [0.458–0.792]	0.167	12% [12%–100%]	0.000 [0.000–0.083]

distances: cosine distance between final-layer residuals, log geodesic distance, and unweighted tree-path distance in the platial administrative hierarchy: We report partial Spearman correlations, so $r_{\text{spatial}|\text{platial}}$ measures metric-spatial alignment after controlling for hierarchical-platial distance, and $r_{\text{platial}|\text{spatial}}$ measures hierarchical-platial alignment after controlling for metric-spatial distance.

Figure 4 and Table 7 (Appendix C.2) show that global final-layer geometry is not identical to linear probe availability. The Qwen-2.5 $\Delta = r_{\text{platial}|\text{spatial}} - r_{\text{spatial}|\text{platial}}$ values remain negative across 0.5B–72B (-0.042 to -0.090), meaning metric distance has stronger partial alignment than tree distance even at large scale. Llama shifts in the other direction: Llama-2-70B is the most tree-favored model ($\Delta = +0.070$), and Llama-3.1-70B is mildly tree-favored ($+0.016$). Qwen3 is mixed: Qwen3-14B is mildly platial ($+0.031$), while Qwen3-32B is mildly spatial (-0.030). We deliberately phrase this as a final-layer geometry profile rather than a causal scaling law because the absolute partial correlations are small and the pair-sampling design still inherits pairwise non-independence. Even with that caution, a feature can be linearly decodable without being the only dominant global pairwise-distance geometry of the final layer (Table 1), so we interpret Table 7 as a descriptive comparison of final-layer global shape, not as evidence that any model or family is exclusively metric spatial or exclusively hierarchical platial. GPH-30K establishes that coordinate and hierarchical platial information are both available at scale; HGCD then becomes a targeted disagreement test rather than the source of the paper's coverage.

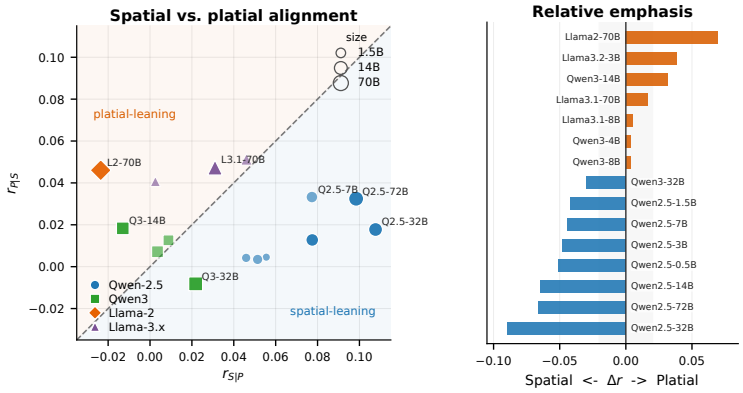

Figure 4: Large-sample final-layer spatial-vs-platial alignment on GPH-30K. The left panel plots partial alignment with $\mathbf{D}_{\text{spatial}}$ against partial alignment with $\mathbf{D}_{\text{platial}}$; the right panel $\Delta = r_{\text{platial}|\text{spatial}} - r_{\text{spatial}|\text{platial}}$. Qwen-2.5 are spatial-leaning through 72B, Llama-70B checkpoints are mixed/platial, and Qwen3 changes sign between 14B and 32B.

3.4 HGCD: METRIC SPATIAL AND HIERARCHICAL PLATIAL STRUCTURES DISAGREE AT TIMES

GPH-30K establishes broad availability. HGCD asks a different question: when coordinate distance and administrative-tree distance make opposite predictions, which geometry is emphasized by a hidden state? Figure 5 first gives the layer-wise diagnostic that motivates this conflict test: spatial coordinate information emerges early, while hierarchical-platial label prediction is strongest near the final layers. We then compute Δr on the 318-location conflict set, where negative values mean

the embedding geometry is closer to metric distance and positive values mean it is closer to the administrative tree after controlling for the other distance matrix.

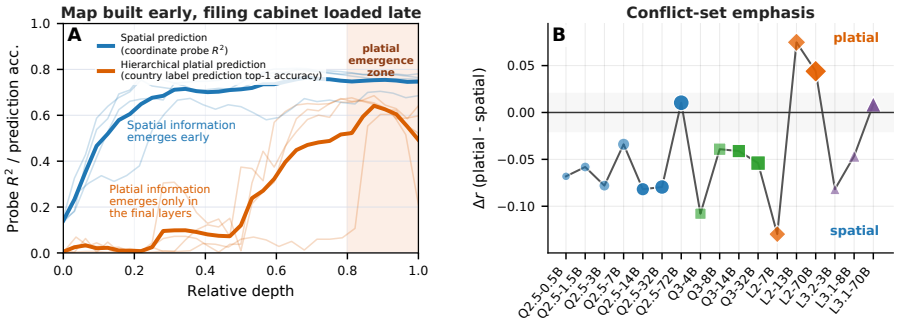

Figure 5: HGCD conflict-set summaries. **(A)** Layer-wise spatial and hierarchical-platial prediction curves for the 7B/8B models on HGCD, showing coordinate prediction across depth and late country-label prediction. **(B)** On the full 318 location-conflict set, negative $\Delta r = r_{\text{platial}|\text{spatial}} - r_{\text{spatial}|\text{platial}}$ indicates stronger metric than tree alignment. Smaller Qwen/Llama checkpoints are mostly spatial-leaning, Llama-2-70B is platial-leaning, and Llama-3.1-70B/Qwen-2.5-72B are mixed.

The location conflict diagnostic set HGCD is deliberately smaller than GPH-30K, so we use it only after the large-scale coverage analysis results. Table 14 (Appendix D.6) shows that the smaller Qwen/Llama checkpoints are mostly spatial-leaning on HGCD. The large models are more heterogeneous: Llama-2-70B becomes platial ($\Delta r = +0.044$), while Llama-3.1-70B and Qwen-2.5-72B are mixed ($+0.009$ and $+0.010$). Qwen-2.5-14B/32B and Qwen3-14B/32B remain spatial. This does not contradict the GPH-30K hierarchy results: hierarchy information is recoverable at scale, but when physical proximity and administrative membership are forced to disagree, final-layer geometry can follow different views depending on family, generation, and scale. The layer-wise diagnostic in Figure 5A helps explain why conflict behavior is not reducible to broad availability: spatial and hierarchical-platial predictions have different depth profiles before the final-layer conflict geometry is measured.

4 DISCUSSION

In this research, we construct a global hierarchical platial structure, extract residual-stream activations for 30,000 locations, and ask whether a single hidden state in LLMs supports both metric spatial and hierarchical platial information. The answer is yes. Coordinates are linearly recoverable, coarse hierarchy labels are linearly recoverable, and several child-to-parent relations can be retrieved by linear maps. The strongest relational result is admin1→country: it remains above chance when parent countries are held out during training, which separates it from simple memorization of known child–parent pairs. The results also show why coordinate probes alone are incomplete. Final-layer embedding geometry can be spatial-leaning, balanced, or mildly platial-leaning depending on model family, generation, and scale. HGCD turns this into a direct disagreement test: when physical proximity and administrative membership make opposing predictions, smaller Qwen/Llama checkpoints mostly emphasize metric proximity, but Llama-2-70B becomes platial and Llama-3.1-70B/Qwen-2.5-72B are mixed. Broad availability therefore does not determine which geographic view dominates a particular representation.

The corpus baselines make the interpretation more concrete. Distance-sensitive Wikipedia links recover a map-like embedding on HGCD, and full-coverage place-context PPMI recovers much of the GPH-30K hierarchy while also approximating coordinates without coordinate tokens. These baselines show that ordinary text already contains the raw material for both views. The transformer-specific finding is that trained LLMs keep continuous metric structure and discrete hierarchical platial structure jointly available in the same residual stream, with different hierarchy edges emerging at different depths and different geometric views winning under conflict. This work offers insights into the neural representations of our physical world in LLMs toward geospatial intelligence.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy P. Lillicrap, Piotr Wojciech Mirowski, Alexander Pritzel, Martin J. Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Hu Zhang, Ross Goroshin, Neil C. Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan Kumaran. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557:429 – 433, 2018.

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. Evaluation of geographical distortions in language models. *Machine Learning*, 114(12):263, 2025. doi: 10.1007/s10994-025-06916-9.

Nir Fulman, Abdulkadir Memduhoğlu, and Alexander Zipf. Distortions in judged spatial relations in large language models. *The Professional Geographer*, 76(6):703–711, 2024. doi: 10.1080/00330124.2024.2372792. URL https://doi.org/10.1080/00330124.2024.2372792.

Song Gao. Towards place-based GIS. In *New thinking in GIScience*, pp. 51–58. Springer, 2022.

GeoNames. Geonames geographical database. https://www.geonames.org/, 2026.

Michael F. Goodchild. Formalizing place in geographic information systems. In *Communities, Neighborhoods, and Health*, pp. 21–33. Springer, 2011.

Michael F Goodchild. Space, place and health. *Annals of GIS*, 21(2):97–100, 2015.

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

Gilles Guillot and François Rousset. Dismantling the mantel tests. *Methods in Ecology and Evolution*, 4(4):336–344, 2013. doi: https://doi.org/10.1111/2041-210x.12018.

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

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, November 2019. doi: 10.18653/v1/D19-1275.

Stephen C. Hirtle and John Jonides. Evidence of hierarchies in cognitive maps. *Memory & Cognition*, 13(3):208–217, 1985. doi: 10.3758/BF03197683.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Dhruva Karkada, Daniel J. Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations. *arXiv preprint arXiv:2602.15029*, 2026. doi: 10.48550/arXiv.2602.15029.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

Pierre Legendre and Marti J. Anderson. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, 69(1):1–24, 1999. doi: https://doi.org/10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2.

Pierre Legendre and Marie-Josée Fortin. Comparison of the mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, 10(5):831–844, 2010. doi: 10.1111/j.1755-0998.2010.02866.x.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. doi: 10.1162/tacl_a_00134. URL https://aclanthology.org/Q15-1016/.

Max M. Louwerse and Rolf A. Zwaan. Language encodes geographical information. *Cognitive Science*, 33(1):51–73, 2009. doi: 10.1111/j.1551-6709.2008.01003.x.

Rohin Manvi, Samar Khanna, Marshall Burke, David B. Lobell, and Stefano Ermon. Large language models are geographically biased. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34654–34669. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/manvi24a.html.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David B. Lobell, and Stefano Ermon. GeoLLM: Extracting geospatial knowledge from large language models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=TqL2xBwXP3.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Edvard I. Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, 31:69–89, 2008. doi: 10.1146/annurev.neuro.31.061307.090723.

Natural Earth. Natural earth. https://www.naturalearthdata.com/, 2026.

John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*. Oxford University Press, 1978.

OpenStreetMap contributors. Openstreetmap. https://www.openstreetmap.org, 2026.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems 30*, pp. 6076–6085. Curran Associates, Inc., 2017.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

REST Countries. REST Countries. https://restcountries.com/, 2026.

Albert Stevens and Patty Coupe. Distortions in judged spatial relations. *Cognitive Psychology*, 10(4):422–437, 1978. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(78)90006-3. URL https://www.sciencedirect.com/science/article/pii/0010028578900063.

Waldo R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1):234–240, 1970. doi: 10.2307/143141.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Yi-Fu Tuan. *Space and Place: The Perspective of Experience*. University of Minnesota Press, 1977.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

A APPENDIX GUIDE

The appendix is organized as a verification path for the paper rather than as a chronological experiment log. A reader who wants to audit the main claim should read the sections in order: first the benchmark and protocol definitions, then the co-occurrence and geometry evidence, and finally the hierarchy robustness checks.

Anonymous reproducibility repository. Code, data files, cached experiment summaries, and figure-generation scripts are available at: `https://anonymous.4open.science/r/ICLR_LinguisticCartography-2E89`.

Table 4: Appendix organization. Sections are ordered as a verification path rather than as a chronological experiment log.

Appendix block	Purpose	Sections
Related work / scope	Preserve the full literature context and state what the representation-based evidence can and cannot support.	B.1, B.2
Benchmark / protocols	Document GPH-30K and HGCD construction, probe definitions, distance choices, release contents, and compute setup.	C.1–C.7
Co-occurrence baselines	Test whether shallow corpus-statistical signals can recover metric-spatial or hierarchical-platial structure.	D.1–D.4
Qwen/Llama series geometry	Summarize the scale, family, depth, HGCD conflict, and hierarchy-PCA analyses that extend the main figures.	D.5–D.7
Hierarchy robustness	Bound the parent-recovery claim with parent-held-out splits, large-model seed summaries, selective subspace erasure, rank sweeps, and external-family checks.	E.1–E.3

Evidence tiers. The appendix uses three tiers. **Tier 1** is load-bearing for the paper's central claim: large-scale coordinate probes, hierarchy-label probes, linear parent recovery, and HGCD metric-vs-tree geometry. **Tier 2** checks robustness to distance metrics, co-occurrence baselines, data splits, and model family. Decoder-token diagnostics are intentionally omitted from this version because the main hierarchy claim is decoder-independent.

B RELATED WORK AND SCOPE

B.1 RELATED WORK

This paper builds on three lines of work. First, coordinate-probing studies show that language-model activations linearly encode latitude, longitude, and time across multiple scales (Gurnee & Tegmark, 2024), while geospatial NLP work documents both geographic competence and spatial distortions in model outputs (Manvi et al., 2024b;a; Fulman et al., 2024; Decoupes et al., 2025). We keep the coordinate-probe question but add a second structure: the administrative hierarchy containing each place.

Second, geography distinguishes metric *space* from contextual *place*; we refer to place-based hierarchical organization as *hierarchical platial* structure (Tuan, 1977; Goodchild, 2011). Human spatial judgments and neural navigation systems also mix metric and categorical structure (Stevens & Coupe, 1978; Hirtle & Jonides, 1985; O'Keefe & Nadel, 1978; Moser et al., 2008; Banino et al., 2018). Our contribution is to make the same distinction operational inside LLM hidden states, using one dataset that gives each place both coordinates and a complete administrative path.

Third, our measurements use standard interpretability tools: linear probes (Alain & Bengio, 2016; Belinkov, 2022), structural probing ideas (Hewitt & Liang, 2019), representational similarity (Kriegeskorte et al., 2008), and linear maps between related residual-stream representations. The main evidence is decoder-independent: we use hidden-state probes, parent retrieval, and pairwise geometry rather than country-token output scores.

Finally, language itself contains geographic structure. Linguistic co-occurrence statistics are distance-sensitive (Tobler, 1970; Louwerse & Zwaan, 2009), and distributional symmetries can shape representation geometry (Karkada et al., 2026). We therefore treat co-occurrence as a testable baseline. Appendices D.1–D.4 report the corresponding PPMI–SVD controls.

The geographic distinction motivating the paper, prior work on coordinate representations, hierarchy-oriented probing tools, and the corpus-statistical view that motivates our PPMI–SVD controls.

Space, place, and hierarchical platial structure. GIScience distinguishes *space*, a metric coordinate frame, from *place*, an administrative, social, or contextual unit (Tuan, 1977; Goodchild, 2011). We use *hierarchical platial* for this place-based organization when it is instantiated as an administrative hierarchy. Cognitive-map work shows that human spatial judgments are not purely metric: borders, regions, and category structure distort remembered distances (Stevens & Coupe, 1978; Hirtle & Jonides, 1985). Neuroscience gives a complementary picture in which hippocampal-entorhinal systems encode navigational variables but also support task- and context-dependent structure (O'Keefe & Nadel, 1978; Moser et al., 2008; Banino et al., 2018). We use this distinction as a computational question: does a language model represent only metric coordinates, only administrative membership, or both?

Geography in language models. Gurnee & Tegmark (2024) show that LLM hidden states linearly encode latitude, longitude, and time across multiple spatial and temporal scales. Their empirical style also motivates our presentation: build a labeled dataset, extract residual-stream activations, train simple linear readouts, and then test robustness and interpretation. Other work studies geospatial knowledge and bias in model outputs, including geographic attribute extraction, regional distortion, and uneven coverage across places (Manvi et al., 2024b;a; Fulman et al., 2024; Decoupes et al., 2025). Our paper differs by using one global place dataset with both coordinates and administrative paths, then asking how metric spatial and hierarchical platial views coexist in the same hidden states.

Hierarchical representations. Hierarchical structure is a natural target for probing because it requires relations between levels rather than a single flat label. Structural probes (Hewitt & Liang, 2019) and probing surveys (Belinkov, 2022) highlight both the value and the limitations of readout-based evidence. We therefore use hierarchy probes only as an availability test and put more weight on child-to-parent maps: a linear transformation from a child-node representation should retrieve the parent-node representation. This keeps the main hierarchy claim decoder-independent and avoids country-token or unembedding-basis artifacts.

Representation comparisons and feature subspaces. Representational similarity analysis compares the geometry induced by two distance matrices (Kriegeskorte et al., 2008); CKA, SVCCA, and PWCCA provide alternative ways to compare representation spaces (Kornblith et al., 2019; Raghu et al., 2017; Morcos et al., 2018). We use RSA because it can compare model representations, geodesic distances, administrative-tree distances, and corpus baselines without requiring matched dimensionality. Linear probes (Alain & Bengio, 2016; Belinkov, 2022) provide the basic coordinate and label readouts, while selective subspace erasure tests whether metric information is necessary for the hierarchy readouts.

Corpus-distributional geography. Language statistics carry spatial signal (Tobler, 1970; Louwerse & Zwaan, 2009); nearby places are more likely to be discussed together, and administrative contexts frequently mention countries, regions, and local units. Recent theory argues that statistical symmetries can shape learned representational geometry (Karkada et al., 2026). We therefore treat co-occurrence not as a nuisance but as an explicit hypothesis. On HGCD, if co-occurrence induces the metric signal, then place-place links should decay with distance, PPMI–SVD embeddings should recover coordinates, and degree-preserving rewiring should remove the effect. On GPH-30K, if ordinary geographic text supplies hierarchy context, then a place-context PPMI–SVD baseline should recover

both hierarchy labels and approximate coordinates without coordinate tokens. Appendices D.1–D.4 test these predictions.

B.2 SCOPE NOTES

The evidence in this paper is representational. Linear probes, parent retrieval, RSA, and dbRDA show what information is available in hidden states and how residual-stream geometry is organized; they do not by themselves prove how a model uses that information during generation. For that reason, the main text avoids decoder-facing country-token readouts. Tokenization choices, multi-token country names, and the decoder basis can all affect output-token experiments without changing the hidden-state geometry measured here.

GPH-30K is large and built from primary geospatial sources, but it is not a census of every place on Earth. The construction requires valid coordinates, complete hierarchy metadata, and a unique English or romanized display name, so it is biased toward well-documented populated places. The primary model analysis is also bounded to Qwen/Llama series checkpoints. Pythia, OPT, and Mistral runs are reported as external-family robustness checks rather than as part of the main scaling or generation analysis.

The hierarchy claim has two strengths. Admin1→country is the strongest relational result because it remains above chance under parent-disjoint country splits. The other edges show in-support recovery but are weaker under parent-held-out tests: place→admin1 falls to chance, country→subregion is difficult when a subregion is removed from training, and subregion→continent is constrained by having only six parent candidates. Finally, administrative taxonomies are political artifacts rather than neutral ground truth. We use standardized public sources for reproducibility, while noting that disputed territories and heterogeneous subnational systems remain unavoidable limitations of any global administrative hierarchy.

C BENCHMARK AND PROTOCOLS

This block defines the broad-coverage hierarchy, the diagnostic set, the probe/evaluation protocols, distance sensitivities, and compute setup. It is placed first so later controls can refer back to stable definitions.

C.1 GPH-30K CONSTRUCTION DETAILS

GPH-30K is built from primary geospatial sources rather than from a prior LLM probing dataset. We start with the GeoNames `cities5000` populated-place dump and join each record to GeoNames `countryInfo.txt` and `admin1CodesASCII.txt`. Natural Earth admin-0 metadata supplies continent and United Nations regional labels, with REST Countries used as a fallback for small territories. We keep only feature-class `P` populated places with valid coordinates, a country code, an admin1 code and name, and complete continent/subregion labels on one of the six populated continents.

The eligible pool contains 68,277 locations. From this pool we select 30,000 locations with a deterministic coverage-preserving policy: globally duplicated bare place names are removed; each country receives a minimum quota where possible; additional locations are allocated roughly in proportion to the square root of the eligible country count; and a soft country-share cap prevents the largest countries from dominating the sample. The released records include GeoNames IDs, names, ASCII names, coordinates, timezone, country/admin1 codes, hierarchy labels, and prompt strings. The final dataset contains 30,000 unique place names, 226 countries and territories, 3,624 admin1 units, 24 subregions, 5 United Nations regions, and all six populated continents.

Table 5: GPH-30K continental coverage after coverage-preserving sampling from eligible GeoNames populated places.

Continent	Africa	Asia	Europe	North America	Oceania	South America
Locations	4,787	10,153	8,013	3,552	780	2,715

This dataset is intentionally a populated-place hierarchy: it supports coordinate probes and hierarchy probes on exactly the same examples. HGCD is retained as the smaller conflict diagnostic where metric and hierarchy predictions are deliberately opposed.

Audit checks. The repository includes a dataset card and audit script (`data/GPH30K_DATASET_CARD.md`, `scripts/audit_gph30k.py`). The audit verifies that all 30,000 records have unique bare place names and unique GeoNames IDs; all coordinates lie in valid latitude/longitude bounds; all required hierarchy fields are present; all six populated continents are represented; and the largest country share remains below the predefined 8% cap. The largest country is India with 2,291 locations (7.64%). The sorted GeoNames-ID checksum prefix is `4c0677c5c68ef87e`.

C.2 GPH-30K PLACE ANCESTOR RECOVERY

Here, we report more details on the non-adjacent ancestor recovery experiment results on GPH-30K dataset and the spatial-vs-platial alignment on 5,000 stratified locations.

Table 6: Non-adjacent ancestor recovery on GPH-30K. Rows report median [min–max] separately for the small/mid and large Qwen/Llama series checkpoints. SAP maps recover coarse ancestors from child-node residuals without putting the parent label in the child prompt. The last row is CHP: it composes two separately trained adjacent maps, place→admin1 and admin1→country, then retrieves the country. SAP place→country is consistently stronger than CHP, showing that country information is present in place states while adjacent hierarchy composition is only partial.

Edge	Models	Cond.@1 ↑	Global@1 ↑	Chance ↓	Emergence ↓
Place→country	0.5B–8B	0.594 [0.448–0.777]	0.505 [0.331–0.720]	0.124	57% [11%–100%]
Place→country	14B–72B/70B	0.689 [0.480–0.792]	0.611 [0.346–0.738]	0.124	88% [12%–100%]
Place→subreg.	0.5B–8B	0.803 [0.710–0.890]	0.727 [0.575–0.834]	0.281	25% [12%–39%]
Place→subreg.	14B–72B/70B	0.844 [0.791–0.909]	0.782 [0.735–0.856]	0.281	25% [12%–50%]
Place→cont.	0.5B–8B	0.831 [0.737–0.901]	0.831 [0.737–0.901]	0.167	94% [12%–100%]
Place→cont.	14B–72B/70B	0.874 [0.677–0.911]	0.874 [0.677–0.911]	0.167	50% [12%–100%]
Admin1→subreg.	0.5B–8B	0.817 [0.653–0.909]	0.747 [0.526–0.869]	0.263	30% [25%–39%]
Admin1→subreg.	14B–72B/70B	0.871 [0.810–0.927]	0.826 [0.735–0.887]	0.263	25% [12%–50%]
Admin1→cont.	0.5B–8B	0.849 [0.745–0.928]	0.849 [0.745–0.928]	0.167	89% [25%–100%]
Admin1→cont.	14B–72B/70B	0.906 [0.679–0.930]	0.906 [0.679–0.930]	0.167	38% [12%–100%]
Country→cont.	0.5B–8B	0.894 [0.721–0.942]	0.894 [0.721–0.942]	0.167	94% [12%–100%]
Country→cont.	14B–72B/70B	0.881 [0.500–0.934]	0.881 [0.500–0.934]	0.167	62% [12%–100%]
CHP: Place→country	0.5B–8B	0.524 [0.340–0.712]	0.417 [0.226–0.653]	0.124	100%
CHP: Place→country	14B–72B/70B	0.632 [0.352–0.725]	0.506 [0.215–0.668]	0.124	100%

Table 7: Large-sample spatial-vs-platial alignment on 5,000 stratified locations for the Qwen/Llama series. $r_{\text{spatial}|\text{platial}}$ is the partial Spearman correlation between embedding distance and geodesic distance after controlling for administrative-tree distance; $r_{\text{platial}|\text{spatial}}$ is the analogous tree-distance correlation after controlling for geodesic distance. $\Delta = r_{\text{platial}|\text{spatial}} - r_{\text{spatial}|\text{platial}}$; small magnitudes should be read as relative geometry descriptors.

Model	Size	$r_{S\|P}$ ↑	$r_{P\|S}$ ↑	Δ platial ↑
Qwen-2.5-0.5B	0.5B	0.056	0.005	-0.051
Qwen-2.5-1.5B	1.5B	0.046	0.004	-0.042
Qwen-2.5-3B	3B	0.051	0.003	-0.048
Qwen-2.5-7B	7B	0.077	0.033	-0.044
Qwen-2.5-14B	14B	0.077	0.013	-0.065
Qwen-2.5-32B	32B	0.108	0.018	-0.090
Qwen-2.5-72B	72B	0.098	0.032	-0.066
Qwen3-4B	4B	0.009	0.013	+0.004
Qwen3-8B	8B	0.004	0.007	+0.004
Qwen3-14B	14B	-0.013	0.018	+0.031
Qwen3-32B	32B	0.022	-0.008	-0.030
Llama-3.2-3B	3B	0.003	0.041	+0.038
Llama-3.1-8B	8B	0.046	0.051	+0.005
Llama-3.1-70B	70B	0.031	0.047	+0.016
Llama-2-70B	70B	-0.024	0.046	+0.070

C.3 EXTERNAL-FAMILY GPH-30K ROBUSTNESS

The main text uses Llama/Qwen as the primary model panel. For robustness, we also ran the same final-layer GPH-30K coverage and large-sample geometry protocols on Pythia, OPT, and Mistral checkpoints below 14B parameters (Biderman et al., 2023; Zhang et al., 2022; Jiang et al., 2023). These external-family runs are not used to define the main claims, but they show that broad coordinate and hierarchy availability is not limited to the Llama/Qwen panel.

Table 8: External-family broad-coverage GPH-30K probes. These Pythia, OPT, and Mistral runs are robustness checks; the main claims use the Llama/Qwen panel in Table 1.

Model	n	Coord R^2 ↑	Cont. ↑	Subreg. ↑	Country ↑	Admin1 ↑
Llama-3.2-3B	30000	0.767	0.873	0.800	0.634	0.241
Llama-3.1-8B	30000	0.800	0.895	0.829	0.672	0.257
Mistral-7B	30000	0.773	0.880	0.808	0.653	0.250
OPT-1.3B	30000	0.561	0.754	0.628	0.462	0.175
OPT-2.7B	30000	0.617	0.783	0.662	0.497	0.198
OPT-6.7B	30000	0.663	0.807	0.699	0.533	0.211
Pythia-1B	30000	0.492	0.726	0.585	0.419	0.146
Pythia-2.8B	30000	0.589	0.781	0.652	0.482	0.177
Pythia-6.9B	30000	0.609	0.790	0.665	0.494	0.181
Qwen-2.5-0.5B	30000	0.510	0.740	0.598	0.436	0.153
Qwen-2.5-1.5B	30000	0.646	0.812	0.696	0.519	0.193
Qwen-2.5-3B	30000	0.698	0.838	0.735	0.553	0.209
Qwen-2.5-7B	30000	0.735	0.854	0.760	0.581	0.225
Qwen3-4B	30000	0.626	0.808	0.698	0.532	0.200
Qwen3-8B	30000	0.690	0.838	0.740	0.573	0.217

Table 9: External-family large-sample metric-spatial versus hierarchical-platial geometry. These runs use the same 5,000-location, one-million-pair protocol as Table 7.

Model	n loc.	n pairs	$r_{\text{spatial}\|\text{platial}}$ ↑	$r_{\text{platial}\|\text{spatial}}$ ↑	Δ platial ↑
Llama-3.2-3B	5000	1000000	0.003	0.041	0.038
Llama-3.1-8B	5000	1000000	0.046	0.051	0.005
Mistral-7B	5000	1000000	-0.010	0.054	0.063
OPT-1.3B	5000	1000000	-0.004	0.037	0.041
OPT-2.7B	5000	1000000	-0.002	0.038	0.040
OPT-6.7B	5000	1000000	-0.015	0.037	0.052
Pythia-1B	5000	1000000	-0.029	0.029	0.059
Pythia-2.8B	5000	1000000	-0.012	0.035	0.047
Pythia-6.9B	5000	1000000	-0.006	0.037	0.042
Qwen-2.5-0.5B	5000	1000000	0.056	0.005	-0.051
Qwen-2.5-1.5B	5000	1000000	0.046	0.004	-0.042
Qwen-2.5-3B	5000	1000000	0.051	0.003	-0.048
Qwen-2.5-7B	5000	1000000	0.077	0.033	-0.044
Qwen3-4B	5000	1000000	0.009	0.013	0.004
Qwen3-8B	5000	1000000	0.004	0.007	0.004

C.4 HGCD FULL BENCHMARK DETAILS

Dataset splits. All 318 locations are used as a single diagnostic set; we do not train/test split the HGCD benchmark itself. The benchmark is used for distance-geometry diagnostics rather than trained classifier probes.

Release. On publication we will release an anonymised repository containing the GPH-30K records, HGCD locations and conflict pairs, experiment scripts, figure/table-generation scripts, cached result JSONs, and the PPMI–SVD artifacts used for the corpus baselines. The repository will include dataset cards and audit scripts for the released geographic records.

C.5 PROBE PROTOCOL DETAILS

An overview of the measurement protocol is shown in Figure 6.

GPH-30K hierarchy probes. The broad-coverage GPH-30K probes use 5-fold stratified cross-validation on 30,000 locations. Coordinate targets use ridge regression on raw latitude and longitude

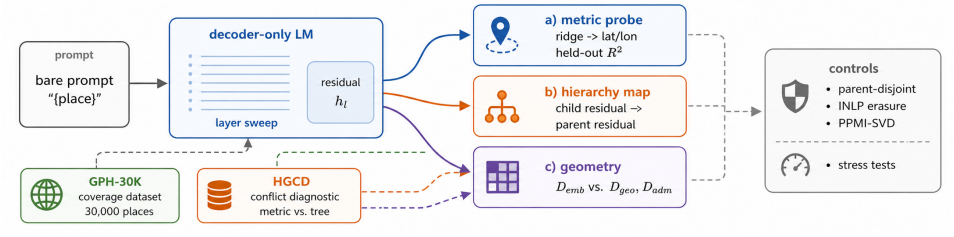

Figure 6: Measurement protocol. From parent-free place prompts, we extract layer-wise residual-stream activations and apply three decoder-independent readouts: metric coordinate probes, child-to-parent hierarchy maps, and distance-matrix geometry. Parent-disjoint recovery, selective INLP erasure, and PPMI–SVD co-occurrence baselines serve as controls on the hierarchy interpretation.

with train-fold standardization of inputs and targets; in the reported main runs the ridge penalty is set to the hidden dimension (`-ridge-alpha -1`). Continent and subregion label probes use all classes. Country and admin1 probes require at least 20 and 10 examples per class, respectively, yielding 168 country classes and 728 admin1 classes in the main table. Label probes use train-fit input standardization, train-fit PCA with 128 components, and a ridge classifier with $\alpha = 1.0$. These flat probes are used only to establish availability of hierarchy labels.

Layer-wise hierarchy-edge recovery. For each adjacent tree edge type (place→admin1, admin1→country, country→subregion, subregion→continent), and for the non-adjacent ancestor edges in Table 6, we cache bare-prompt residuals for all child nodes and all candidate parent nodes at the nearest transformer block outputs to a 12.5% relative-depth grid. Node prompts are parent-free: a place node is prompted by the place name, an admin1 node by the admin1 name alone, a country node by the country name, and so on. Within each fold, child residuals are standardized and reduced with train-fit PCA using up to 256 components, then a ridge map with $\alpha = 10.0$ is fit from child residuals to standardized corresponding parent residuals. The predicted parent vector is inverse-transformed back to the parent residual space. Target parent residuals are standardized but not PCA-reduced. At test time, the mapped child vector is scored by cosine similarity against the raw candidate parent residuals. We report both global retrieval among all parents at that tree depth and context-conditioned retrieval among sibling candidates (e.g., admin1 candidates within the same country). For the CHP diagnostic, we fit place→admin1 and admin1→country maps separately inside each fold, compose their predictions, and retrieve countries from the resulting vector. Emergence is the first sampled depth where chance-normalized conditional retrieval reaches 90% of that model's best sampled value. No decoder vocabulary, first-token country label, or unembedding row is used in this evaluation.

C.6 ADMIN-DISTANCE WEIGHTING SENSITIVITY

The default HGCD administrative distance uses $(w_{\text{country}}, w_{\text{state}}, w_{\text{county}}) = (1.0, 0.5, 0.25)$. To check that the spatial-dominance dbRDA conclusion does not depend on this choice, we re-run the variance-partition analysis on all 4 models under 5 schemes:

Table 10: HGCD administrative-distance weighting sensitivity. Entries are dbRDA ΔR^2 values; negative values indicate stronger metric-spatial than hierarchical-platial alignment.

Scheme	**dbRDA ΔR^2 (platial ↑, spatial ↓)**			
	Llama-2-7B	Llama-3.1-8B	Mistral-7B	Qwen-2.5-7B
default $(1.00, 0.50, 0.25)$	-0.035	-0.060	-0.045	-0.039
uniform $(1.00, 1.00, 1.00)$	-0.276	-0.264	-0.210	-0.223
country-only $(1.00, 0.00, 0.00)$	-0.036	-0.060	-0.045	-0.039
state-only $(0.00, 1.00, 0.00)$	-0.283	-0.271	-0.208	-0.223
heavy-country $(1.00, 0.25, 0.05)$	-0.036	-0.060	-0.045	-0.039

ΔR^2 values; negative = spatial dominant. All 20 cells (4 models × 5 schemes) are negative.

The sign of ΔR^2 is negative (spatial dominant) under *every* weighting scheme on *every* model tested. Effect sizes vary: default and heavy-country give $\Delta R^2 \in [-0.060, -0.035]$, uniform and state-only give $\Delta R^2 \in [-0.28, -0.21]$ (which treats any cross-state pair as equally distant). The spatial-dominance conclusion does not depend on the specific $w_{\text{country}}, w_{\text{state}}, w_{\text{county}}$ choice.

C.7 IMPLEMENTATION DETAILS

The small and mid-size broad-coverage experiments run on single- or few-GPU workers with float16/bfloat16 model weights and temporary HuggingFace caches on local storage. The 70B/72B experiments use multi-GPU A100 inference with automatic device mapping, fp16/bf16 weights, and offload when needed. Statistical analyses (partial Spearman, 999-permutation dbRDA, bootstrap CIs, ridge/ridge-classifier probes) run on CPU after embedding extraction.

D SPATIAL SIGNAL AND GEOMETRY

This block asks whether HGCD contains a recoverable metric-spatial signal, how broadly the depth profile holds, and how metric spatial and hierarchical platial geometry are distributed.

D.1 CORPUS CO-OCCURRENCE HYPOTHESIS: WIKIPEDIA OUTLINK DECAY

The PPMI–SVD baseline operationalises a corpus-statistical hypothesis inspired by symmetry-based accounts of representation geometry (Karkada et al., 2026). If language statistics are distance-sensitive, then even a shallow embedding built from place-place co-occurrence should contain metric geography. We test three predictions on the Wikipedia outlink graph over the 318 HGCD articles: linked-place probability should decay with distance; a PPMI–SVD factorisation should recover coordinate-correlated geometry; and a degree-preserving rewire should remove that signal.

The Wikipedia outlink graph over the 318 HGCD articles gives the pair-level indicator $I(\ell_i, \ell_j) = 1$ iff either article links to the other. This is not a full training-corpus measurement; it is a benchmark-local sufficiency test of whether a low-order text graph already carries the latent geography. Across all $50{,}403$ pairs, 3.19% are linked, with fraction-linked declining monotonically with log-geodesic distance:

Table 11: Wikipedia outlink probability declines with metric-spatial distance on HGCD. Bins are formed over pairwise log-geodesic distance.

log-km decile	$\log(d+1)$ range	# linked / # pairs	frac. linked ↓
0 (nearest)	$[0, 6.82]$	$578/5041$	11.47%
1	$[6.82, 7.57]$	$326/5040$	6.47%
2	$[7.57, 8.40]$	$223/5040$	4.42%
3	$[8.40, 8.81]$	$75/5040$	1.49%
4–9 (farthest)	$[8.81, 9.90]$	$404/30242$	$\sim 1.3\%$ floor

Spearman $\rho(I, \log d_{\text{spatial}}) = -0.14$ at $p < 10^{-200}$; partial Spearman controlling for administrative-tree distance is -0.126. The outlink indicator also contains hierarchical signal: partial Spearman with administrative-tree distance after controlling for geodesic distance is -0.621. The PPMI–SVD factorisation converts these low-order corpus statistics into a coord-correlated embedding *without* any neural training, with dbRDA assigning slightly more unique variance to metric distance than to administrative-tree distance on this conflict set.

D.2 DEGREE-PRESERVING NULL-GRAPH CONTROL FOR THE PPMI–SVD BASELINE

The decay and PPMI results could still be an artifact of generic graph topology: some cities may simply have many links, and high-degree nodes could dominate the SVD. We therefore test the third prediction of the corpus-statistical hypothesis by preserving each node's degree while randomising which geographic pairs are linked.

We generate $N_{\text{null}} = 50$ null graphs from the observed 318-node HGCD outlink graph by repeated double-edge swaps: pick edges (a, b) and (c, d) uniformly at random and replace with (a, c) and (b, d) if neither creates a self-loop nor a duplicate edge. Each null graph preserves each node's degree exactly. For each null we rebuild the PPMI–SVD ($d = 100$) and compute both the dbRDA ΔR^2 and the 5-fold-CV coord probe R^2.

Table 12: Degree-preserving null-graph control for the HGCD Wikipedia-outlink PPMI–SVD baseline. The real graph preserves metric-spatial signal; degree-matched rewiring removes it.

	real graph	**null (mean $\pm$ std, $N = 50$)**
dbRDA ΔR^2 (spatial $\downarrow$)	-0.018	-0.002 ± 0.004
coord probe $R^2 \uparrow$	$+0.371$	-0.025 ± 0.009

one-sided null p-values: $p(\text{null } \Delta R^2 \leq \text{real}) = 0.02$; $p(\text{null coord } R^2 \geq \text{real}) = 0$ (none of 50 nulls reach real coord R^2).

This rules out degree-sequence / topology-only explanations of the PPMI–SVD corpus baseline: the geodesic correlation is specifically driven by how outlinks are *placed* in geographic space, not by the coarse connectivity distribution.

D.3 GPH-30K PLACE-CONTEXT PPMI–SVD BASELINE

The HGCD outlink experiment is intentionally local: it asks whether a real article-link graph over the 318 diagnostic places already contains metric signal. To test the hierarchical platial side at the same scale as the main coverage experiments, we build a full GPH-30K place-context matrix. Rows are the 30,000 GPH-30K places. Context columns are textual descriptors that geographic discourse routinely co-occurs with place names: continent, subregion, country, admin1, and, in the auxiliary variant, timezone and GeoNames feature code. No latitude, longitude, geohash, distance, or coordinate-bin tokens are included. We compute PPMI with context smoothing, take a $d = 100$ truncated SVD, and evaluate the resulting place embeddings with the same coordinate, hierarchy-probe, and pairwise-geometry analyses used for LLM hidden states.

Table 13: GPH-30K place-context PPMI–SVD baseline. Context fields include explicit hierarchy descriptors and therefore make this a strong distributional sufficiency baseline, not a natural-corpus measurement. S and P denote metric-spatial and hierarchical-platial distance.

Context	**Coord.** $R^2 \uparrow$	**Cont.** $\uparrow$	**Subreg.** $\uparrow$	**Country** $\uparrow$	**Admin1** $\uparrow$	$r_{S\|P} \uparrow$	$r_{P\|S} \uparrow$	Δ platial $\uparrow$
Hierarchy	0.964	1.000	0.999	0.879	0.311	-0.009	0.652	+0.661
Hierarchy+aux	0.960	1.000	0.998	0.878	0.265	-0.023	0.580	+0.603

The result is deliberately strong but not tautological. Recovering continent, subregion, country, and admin1 from context fields that include those descriptors shows that low-order co-occurrence can represent the full hierarchy at 30K scale. The nontrivial observation is that the same embeddings recover latitude/longitude with mean $R^2 \approx 0.96$ without coordinate tokens. This follows from spatial autocorrelation in the hierarchy and auxiliary contexts: countries, subregions, timezones, and feature types occupy structured regions of the Earth. The pairwise geometry is strongly tree-dominant, however ($r_{\text{platial}|\text{spatial}} \gg r_{\text{spatial}|\text{platial}}$), so this baseline does not replace the LLM analysis. It shows that a plausible distributional source can jointly supply hierarchy and approximate metric information, and it sets up the mediation test below.

D.4 CO-OCCURRENCE MEDIATION PROTOCOL FOR HIDDEN-STATE GEOMETRY

The next mediation-style analysis asks whether a GPH-30K co-occurrence distance explains the same hidden-state geometry attributed to metric and tree distance. For each checkpoint, the protocol samples 1,500 GPH-30K locations, extracts final-layer last-token hidden states under the bare place prompt, samples 500,000 pairs, and fits rank-regression models of pairwise hidden-state distance:

$$D_{\text{emb}} \sim D_{\text{spatial}} + D_{\text{platial}}, \qquad D_{\text{emb}} \sim D_{\text{spatial}} + D_{\text{platial}} + D_{\text{cooc}},$$

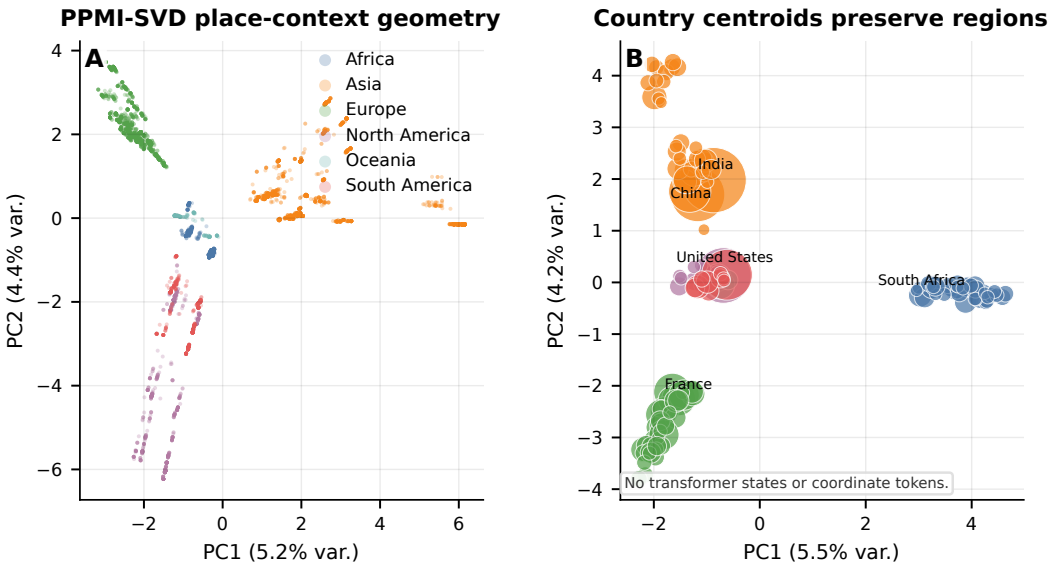

Figure 7: PCA view of the GPH-30K place-context PPMI–SVD baseline. The visualization is not transformer evidence: it uses only explicit hierarchy and auxiliary context tokens, with no coordinate tokens. It illustrates why a distributional baseline can recover both hierarchy regions and approximate metric structure at global scale.

where D_{cooc} is cosine distance in the Hierarchy+aux PPMI–SVD embedding from Appendix D.3. Unique terms are computed by comparing the full model to a model with that predictor removed. This section is a protocol specification and exploratory analysis rather than load-bearing evidence. The no-coordinate PPMI baselines above already establish the key distributional point: shallow context statistics can recover the hierarchy and approximate coordinates at GPH-30K scale. The mediation regression is useful for future model-specific origin claims, but we do not use it to support the paper's central coexistence claim because it is descriptive and its overall distance-regression R^2 values are small.

D.5 LAYER-WISE AND FAMILY ROBUSTNESS SUMMARIES

The main adjacent-edge depth summary is Table 3, and the main visual layer-wise result for non-adjacent ancestors is Figure 2. Together they show that hierarchy edges do not share a single trajectory. Across the Qwen/Llama series, country→subregion is usually available early, admin1→country is late and often final-layer strongest, and place→admin1 is weak even when the coarser hierarchy is recovered. The pattern is stable as the model set expands from small/mid checkpoints to 70B/72B models: place→admin1 remains the hardest adjacent edge, while non-adjacent coarse ancestors are recoverable through SAP. This supports the hierarchy-as-relations interpretation while bounding it: the relation is strongest for middle/coarse edges and weakest for fine local administrative assignment.

The final-layer geometry result in Table 7 adds a separate family/generation observation. Qwen-2.5 checkpoints have negative Δ throughout the tested 0.5B–72B scale, indicating a more spatial-leaning final-layer geometry. Llama-2-70B and Llama-3.1-70B are more platial/mixed, and Qwen3 changes from near-balanced at 4B/8B to mildly platial at 14B and mildly spatial at 32B. We do not claim this is a causal scaling law, because the panel mixes model families and training generations. We report it as a descriptive finding that broader model coverage made visible: linear coordinate and hierarchy availability can coexist with different dominant final-layer geometries.

External-family Pythia, OPT, and Mistral runs are reported in Appendix C.3. They are used only as robustness checks; the main claims are defined on the Llama/Qwen panel.

Table 14: Final-layer conflict-set representations on HGCD under bare place-name prompts for the Qwen/Llama series. Negative Δr indicates stronger alignment with metric distance than platial-tree distance after controlling for the other matrix. Llama-2-7B and Llama-2-13B are HGCD-only diagnostic references, not full GPH-series runs.

Model	Role	$r_{S\|P}\uparrow$	$r_{P\|S}\uparrow$	Δr platial $\uparrow$	Emphasis
Qwen-2.5-0.5B	GPH series	0.111	0.043	-0.068	spatial
Qwen-2.5-1.5B	GPH series	0.125	0.067	-0.058	spatial
Qwen-2.5-3B	GPH series	0.136	0.058	-0.078	spatial
Qwen-2.5-7B	GPH series	0.116	0.082	-0.034	spatial
Qwen-2.5-14B	GPH series	0.126	0.044	-0.082	spatial
Qwen-2.5-32B	GPH series	0.142	0.062	-0.079	spatial
Qwen-2.5-72B	GPH series	0.103	0.113	+0.010	mixed
Qwen3-4B	GPH series	0.096	-0.012	-0.108	spatial
Qwen3-8B	GPH series	0.057	0.018	-0.039	spatial
Qwen3-14B	GPH series	0.060	0.019	-0.041	spatial
Qwen3-32B	GPH series	0.055	0.001	-0.054	spatial
Llama-2-7B	HGCD ref.	0.174	0.045	-0.130	spatial
Llama-2-13B	HGCD ref.	0.057	0.132	+0.075	placial
Llama-2-70B	GPH series	0.084	0.128	+0.044	placial
Llama-3.2-3B	GPH series	0.155	0.073	-0.082	spatial
Llama-3.1-8B	GPH series	0.157	0.111	-0.047	spatial
Llama-3.1-70B	GPH series	0.135	0.144	+0.009	mixed

D.6 FINAL-LAYER CONFLICT-SET REPRESENTATIONS ON HGCD

D.7 PER-MODEL HIERARCHY PCA VIEWS

Figure 8–23 show the representative hierarchy-PCA views for the Qwen/Llama checkpoints with cached hierarchy3D coordinates, including the new 14B–72B/70B models. Each figure uses actual prompted hierarchy nodes, the same four-panel zoom sequence, and the same visual encoding as the main figure; only the model residuals and fitted hierarchy subspace differ.

E HIERARCHY ROBUSTNESS AND SCOPE

This final appendix block keeps only the robustness checks that directly bound the main hierarchy claim. It does not use decoder-token readouts or output-basis interventions.

E.1 QWEN/LLAMA SERIES PROTOCOL AND PARENT-HELD-OUT GENERALIZATION

The model-coverage experiments in this paper are deliberately centered on open-weight Qwen/Llama series checkpoints so that family, generation, and scale can be compared under one protocol. The main GPH-30K series spans Qwen-2.5 from 0.5B to 72B, Qwen3 from 4B to 32B, and Llama-3/Llama-2 up to 70B. External-family Pythia, OPT, and Mistral robustness runs are retained in Appendix C.3, but they are not used to define the main scale or generation claims.

Parent-disjoint parent recovery. The Cond.@1 numbers in Table 3 use stratified 5-fold cross-validation, which lets each parent label appear in both the train and the test fold. To bound how much of the recovery reflects relational generalization to unseen parents rather than associative memory between known children and known parents, the current protocol uses `GroupKFold` on the parent label for every edge in the Llama/Qwen panel. The held-out parent appears only as a retrieval candidate inside the relevant context. This is the verified parent-disjoint result reported in Table 15 and summarized in Table 3. The paper's claims use this parent-disjoint protocol.

Shuffled-label control for admin1→country. To check whether the surviving admin1→country transfer reflects a real learned relation rather than a candidate-set or fold-geometry artifact, we run a label-shuffle control on two strong small/mid models. For each shuffle seed (ten seeds), we replace each training admin1's parent-country embedding with a uniformly random country embedding drawn from the candidate pool, then refit the ridge map and evaluate against the genuine parent. If the

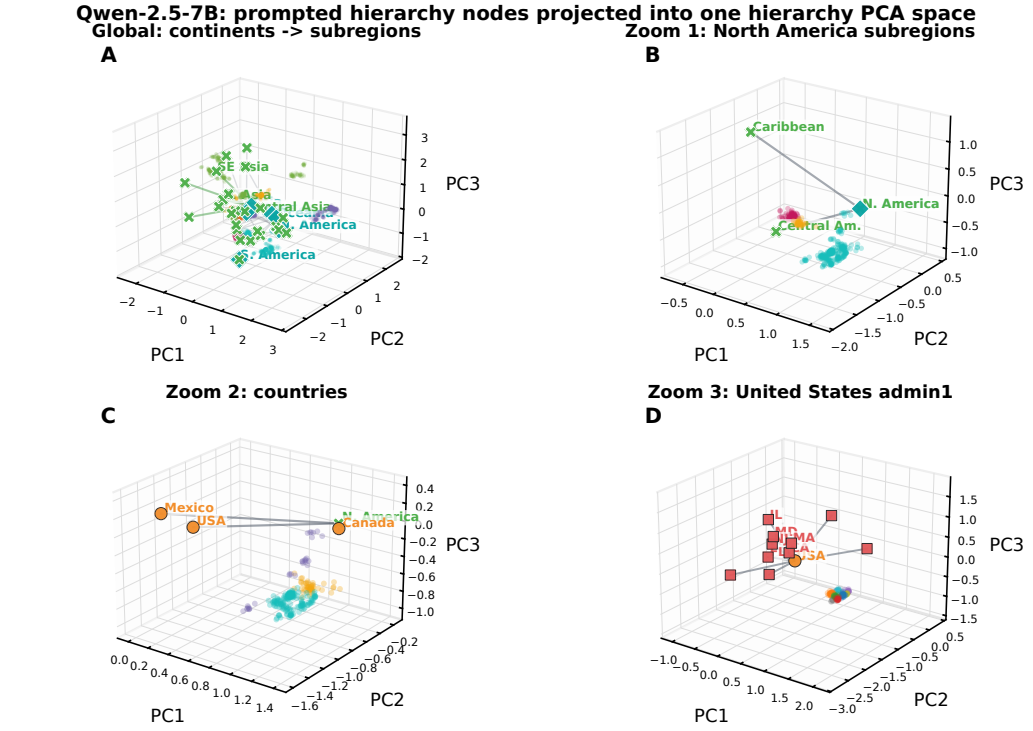

Figure 8: A three-dimensional view of the same hierarchical platial signal using actual prompted hierarchy nodes rather than place centroids. We fit one-vs-rest ridge classifiers for every available ancestor tier in the GPH-30K administrative tree, balance the classifier blocks for continent, subregion, country, and admin1, take the top-32 right singular vectors of the stacked classifier matrix, and PCA-reduce the resulting hierarchy subspace to three dimensions. Faint points are GPH-30K place residuals from Qwen-2.5-7B. Colored markers are hidden states from bare prompts for hierarchy labels themselves (e.g., *Asia*, *Central Asia*, *United States*, and *California, United States*), and line segments are parent-child edges in the GPH-30K tree. The panels progressively zoom from global continent-to-subregion structure to North American subregions, then to North American countries, and finally to US admin1 nodes.

linear map is learning a real admin1→country relation, this should drop to chance. For Qwen-2.5-7B the genuine accuracy (0.211 $[0.198, 0.225]$ across five `GroupKFold` country-partition seeds) is far above the shuffled-label distribution (mean 0.122, range across ten shuffle seeds $[0.090, 0.150]$); the genuine 95% lower bound exceeds the shuffled maximum by 0.05. For Llama-3.1-8B, the genuine accuracy (0.167 $[0.149, 0.185]$) is closer but still clears the shuffled distribution: shuffled mean 0.124, range $[0.093, 0.150]$, and the genuine 95% lower bound essentially matches the shuffled maximum. We therefore read this as clean separation for Qwen-2.5-7B and marginal separation for Llama-3.1-8B; the shuffled-label control collapses toward chance in both tested models, supporting the admin1→country relational claim while bounding how strong the support is for any individual model.

Parent-disjoint generalization on the Qwen/Llama series. The leave-one-continent-out test above is intentionally harsh because removing a continent removes a sizable fraction of the candidate context. To isolate parent-label leakage from regional context loss, we additionally evaluate parent-disjoint splits using `GroupKFold` on the parent label: for each edge we hold a group of parents out at training time and test on children whose true parent is in the held-out group, with the candidate pool conditioned on the relevant context (admin1→country candidates within the held-out admin1's true subregion; country→subregion candidates within the same continent; etc.). The held-out parent therefore appears only as a within-context retrieval candidate. Table 15 reports per-model accuracy

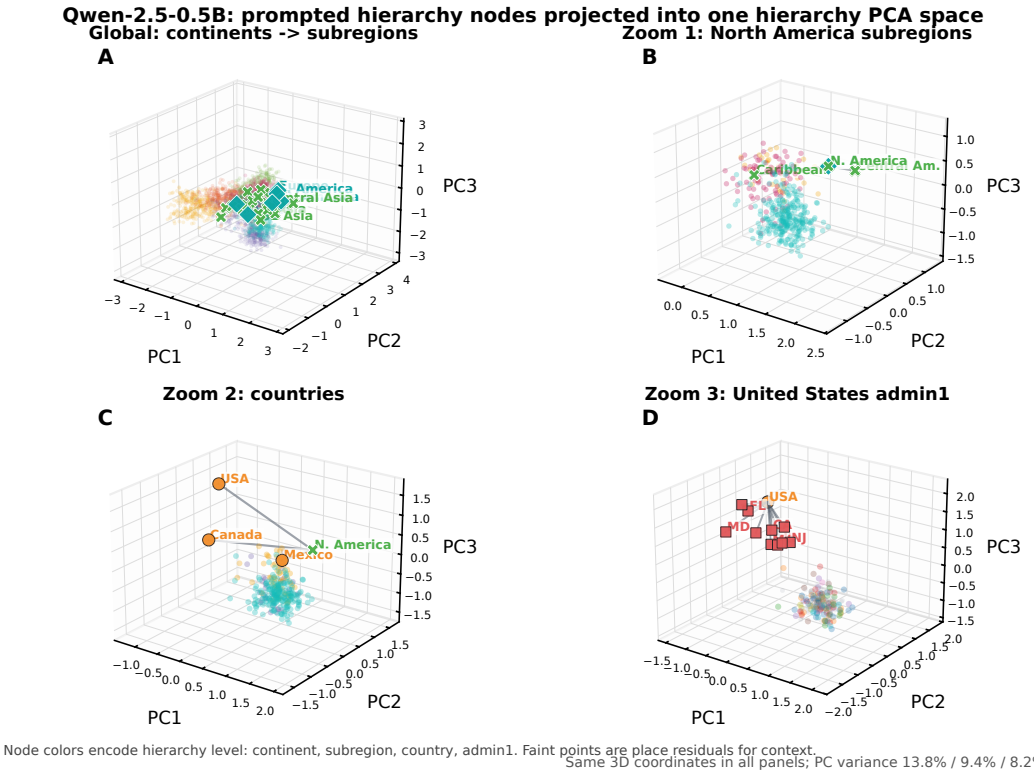

Figure 9: Hierarchy-PCA view for Qwen-2.5-0.5B. The construction is identical to Figure 8: prompted hierarchy nodes are projected into a model-specific hierarchy subspace and shown with progressively narrower geographic zooms.

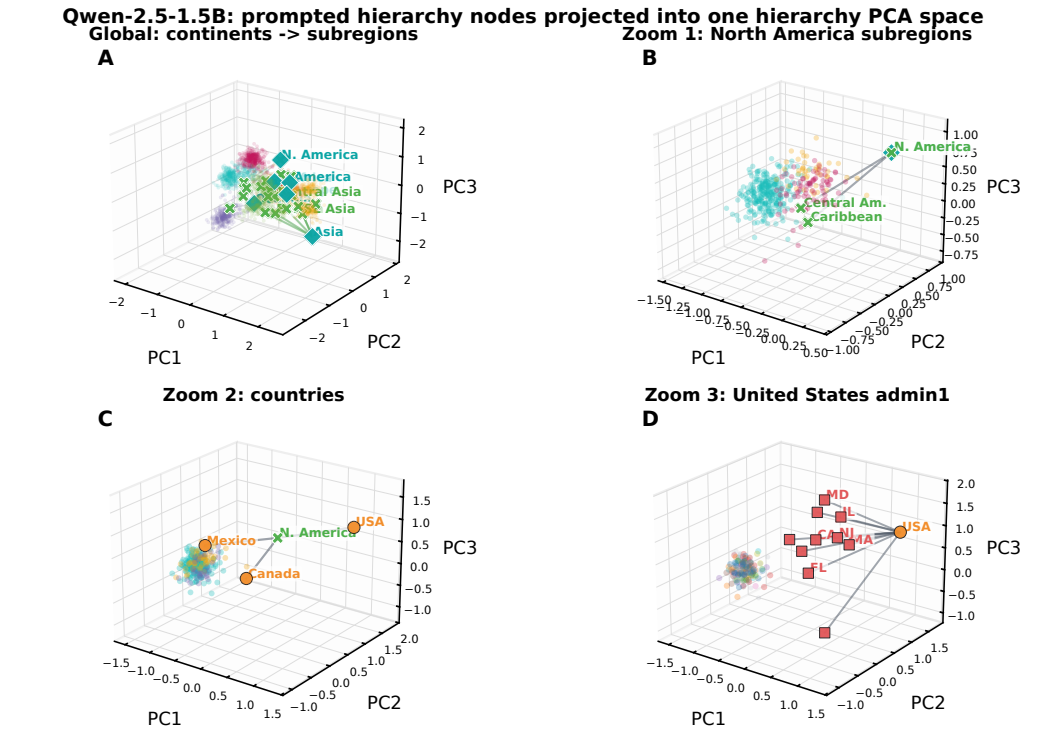

Figure 10: Hierarchy-PCA view for Qwen-2.5-1.5B.

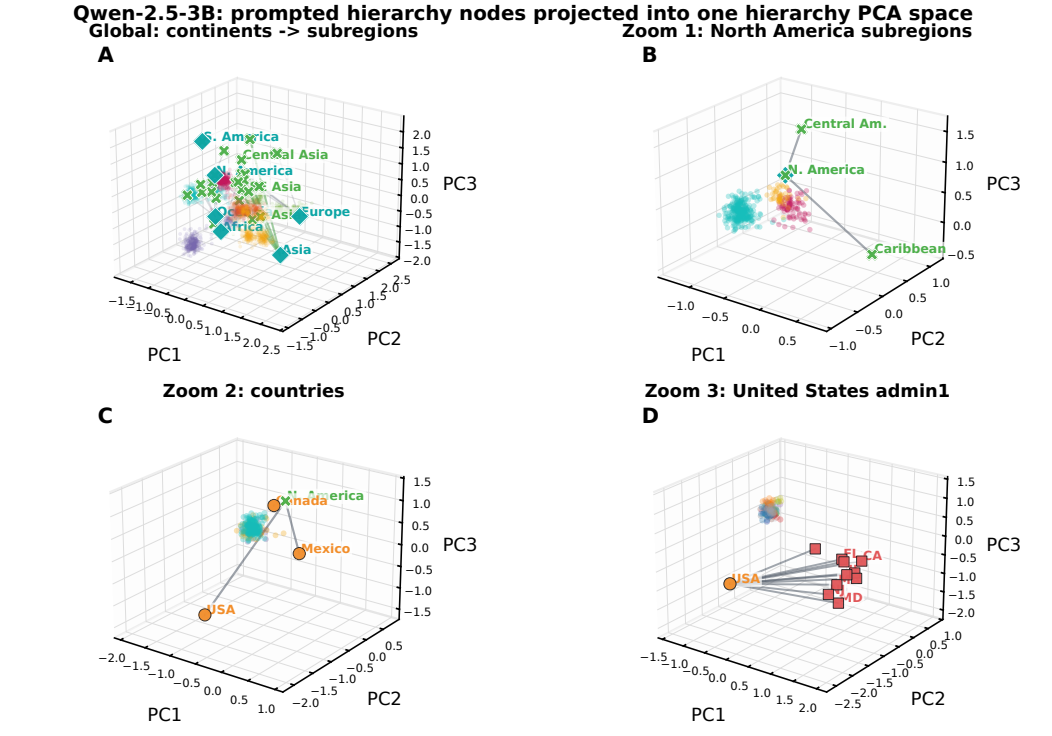

Figure 11: Hierarchy-PCA view for Qwen-2.5-3B.

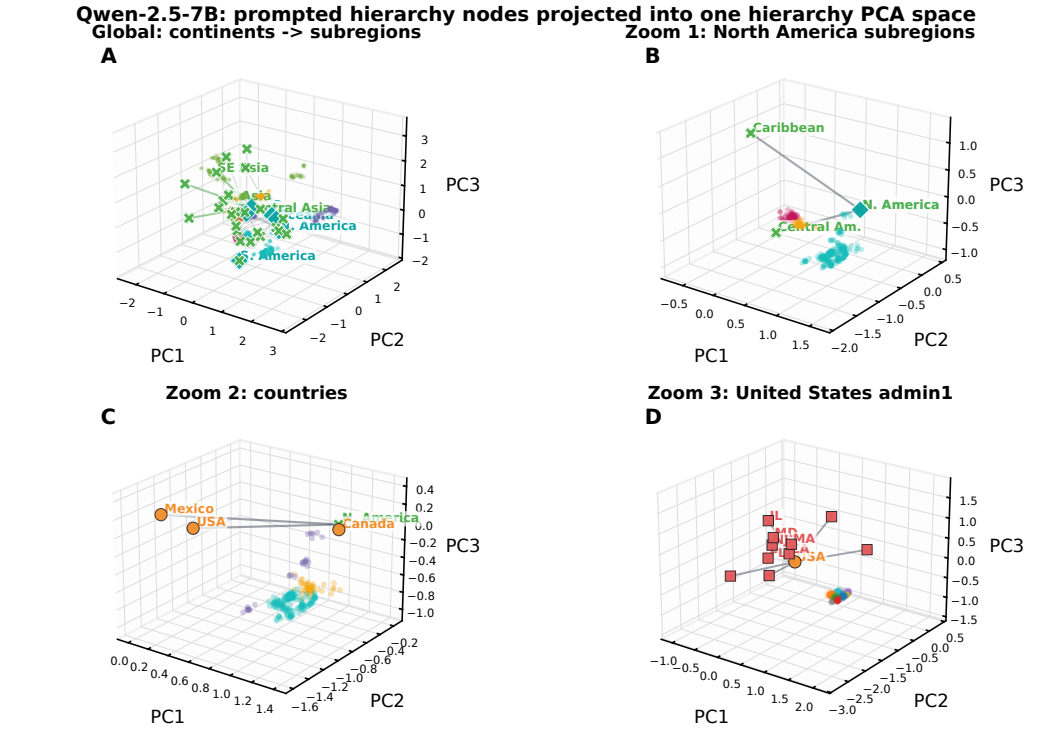

Figure 12: Hierarchy-PCA view for Qwen-2.5-7B.

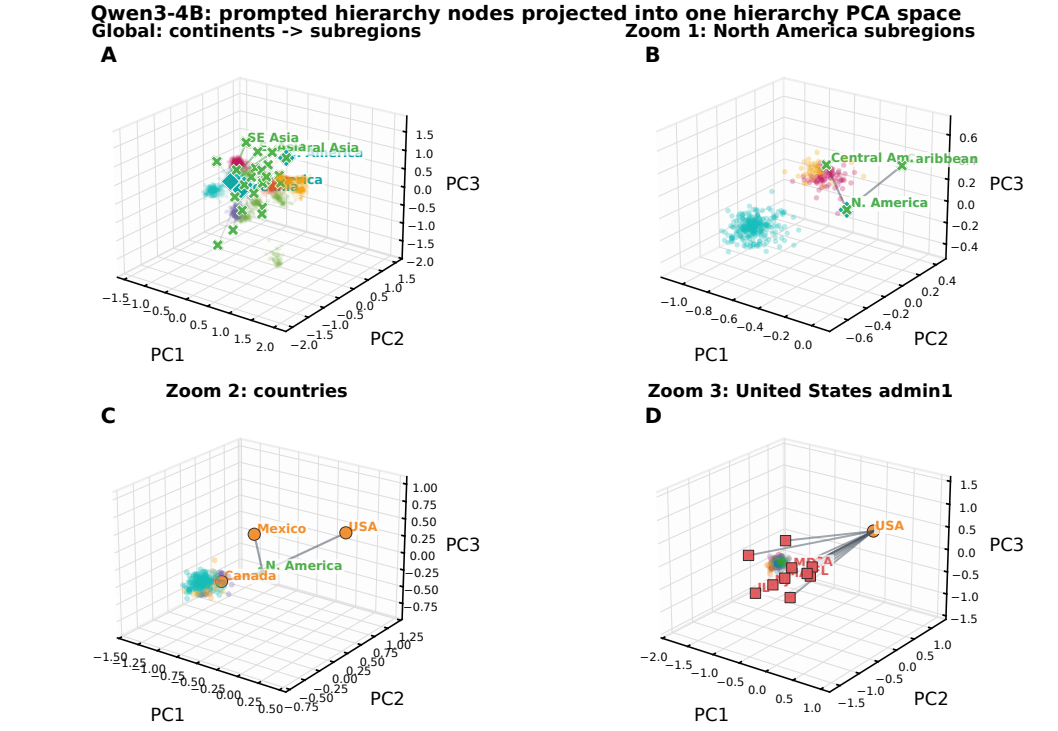

Figure 13: Hierarchy-PCA view for Qwen3-4B.

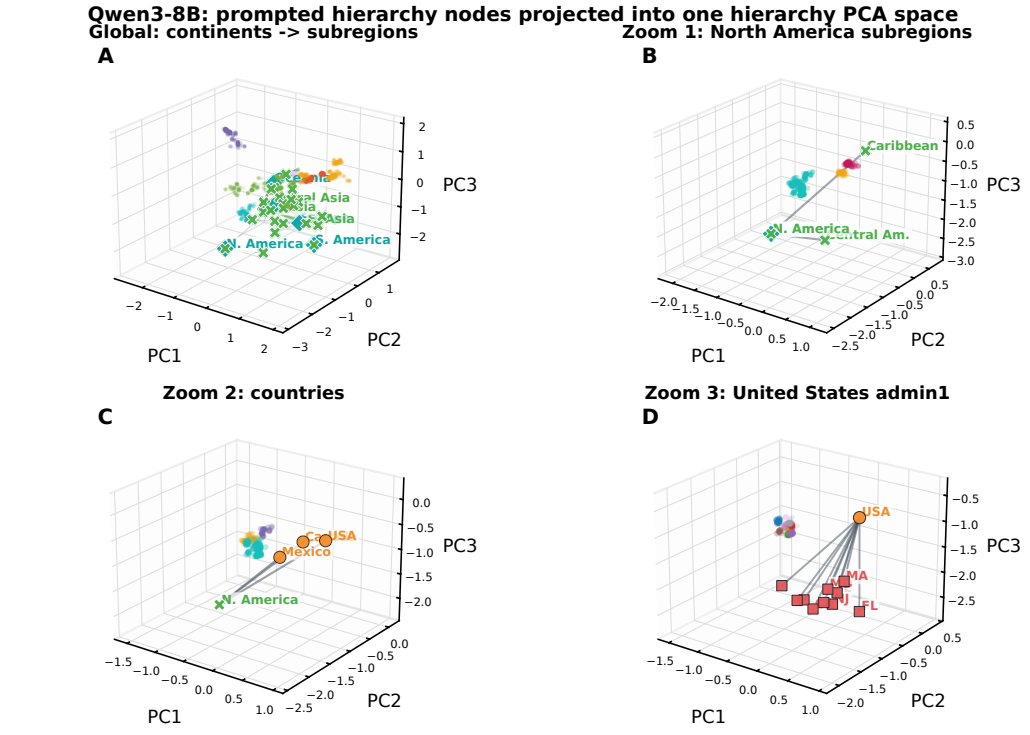

Figure 14: Hierarchy-PCA view for Qwen3-8B.

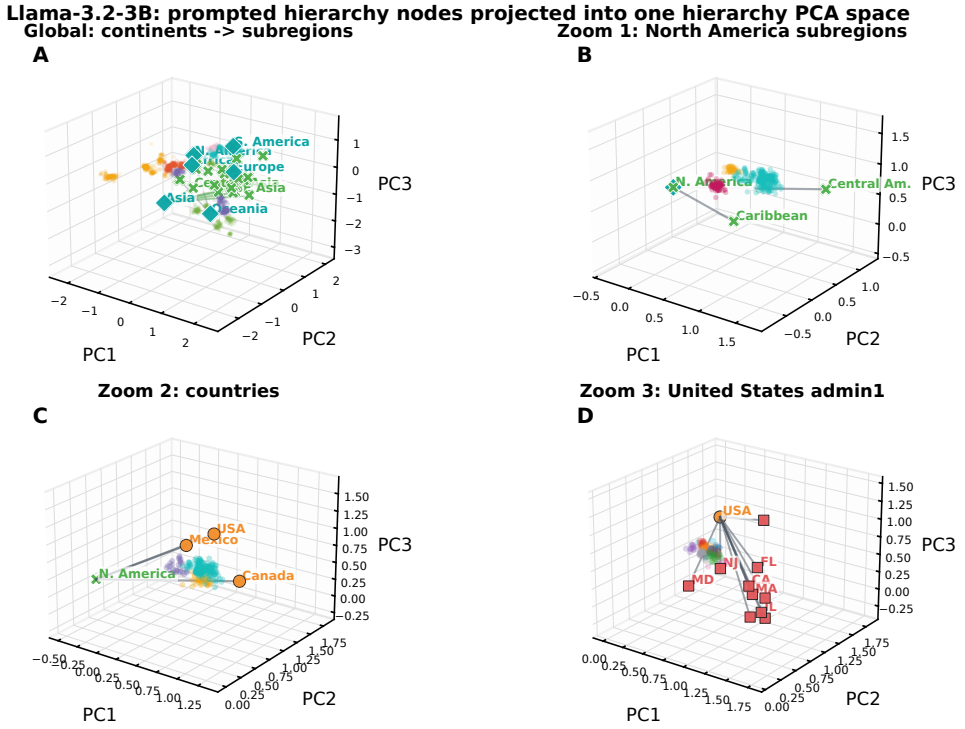

Figure 15: Hierarchy-PCA view for Llama-3.2-3B.

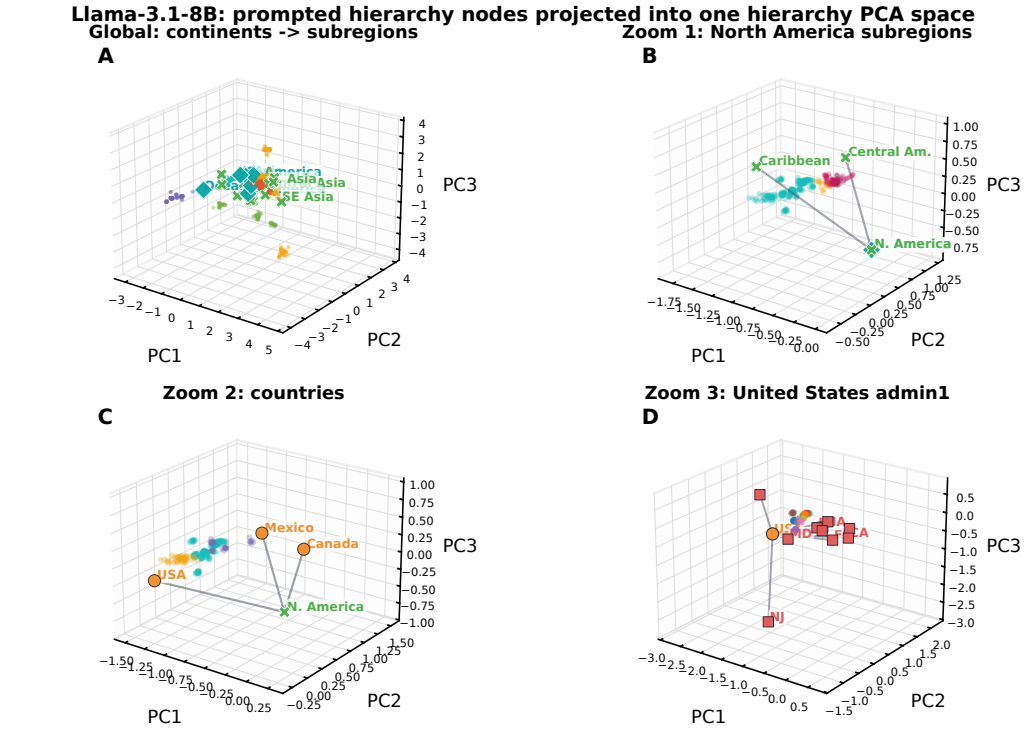

Figure 16: Hierarchy-PCA view for Llama-3.1-8B.

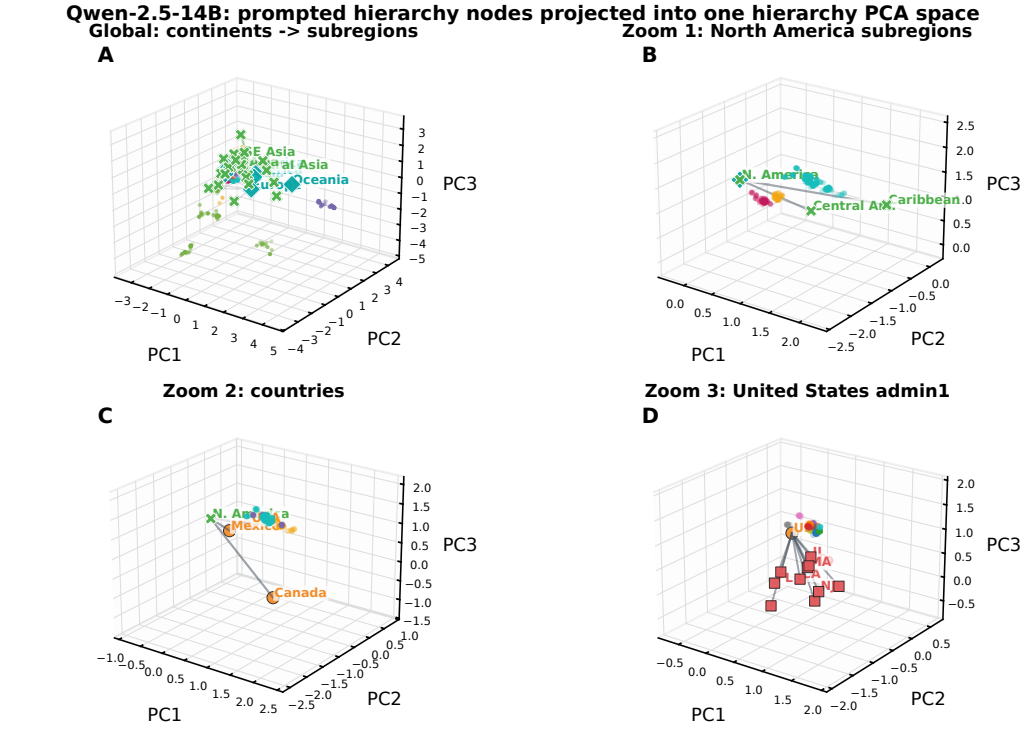

Figure 17: Hierarchy-PCA view for Qwen-2.5-14B.

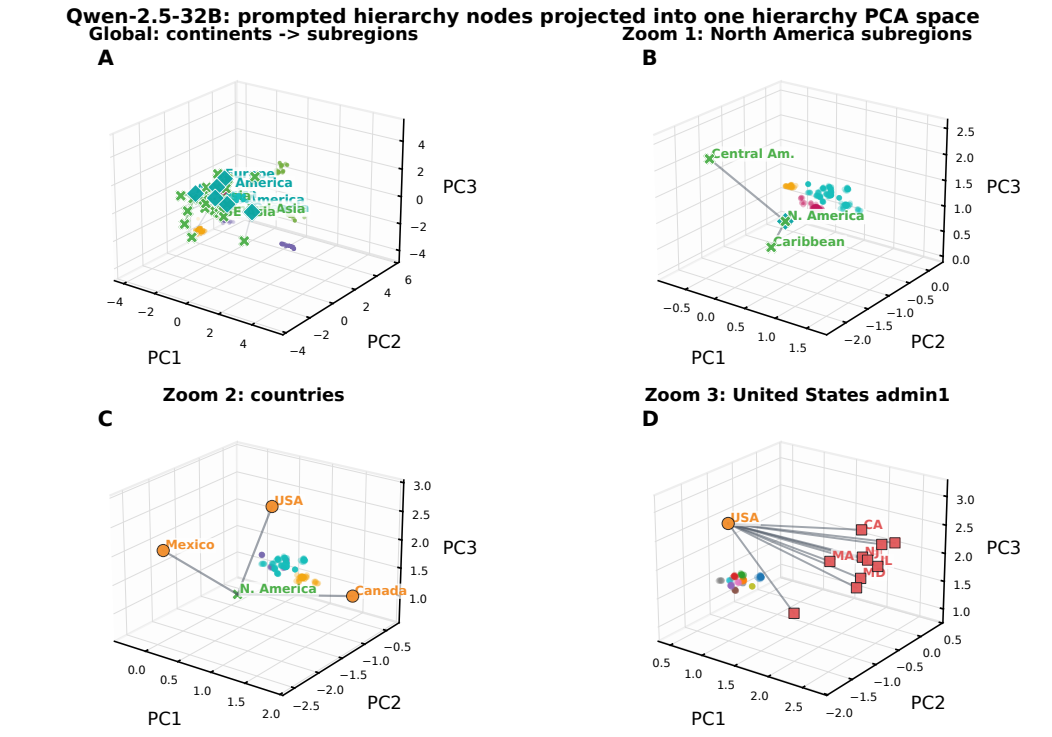

Figure 18: Hierarchy-PCA view for Qwen-2.5-32B.

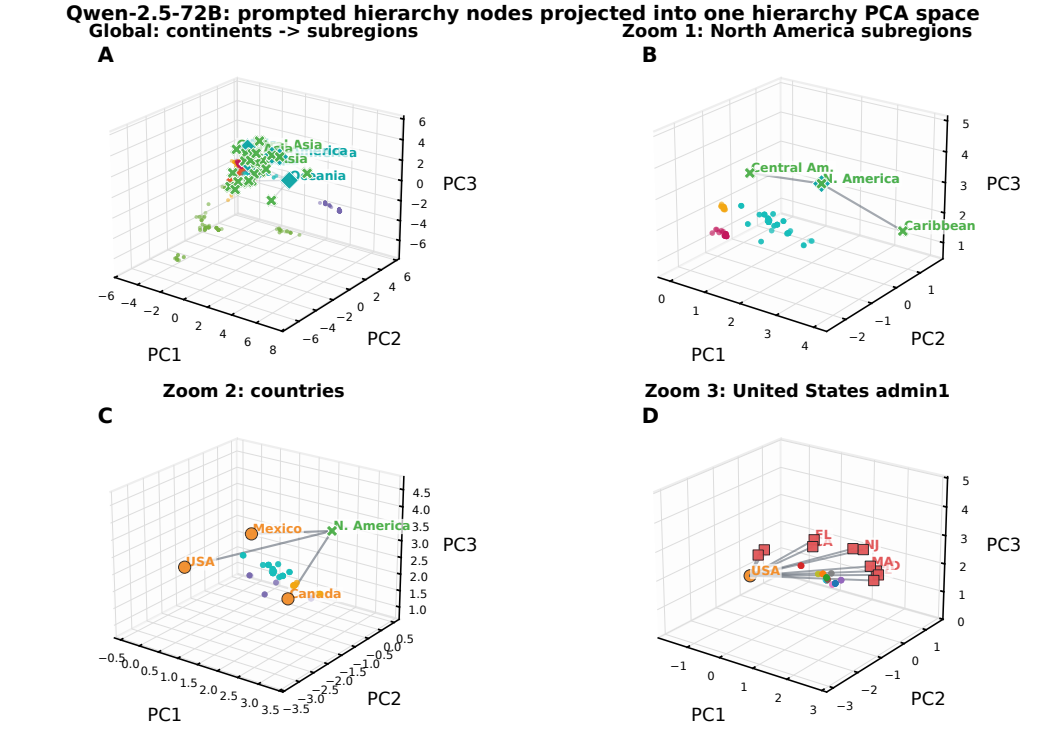

Figure 19: Hierarchy-PCA view for Qwen-2.5-72B.

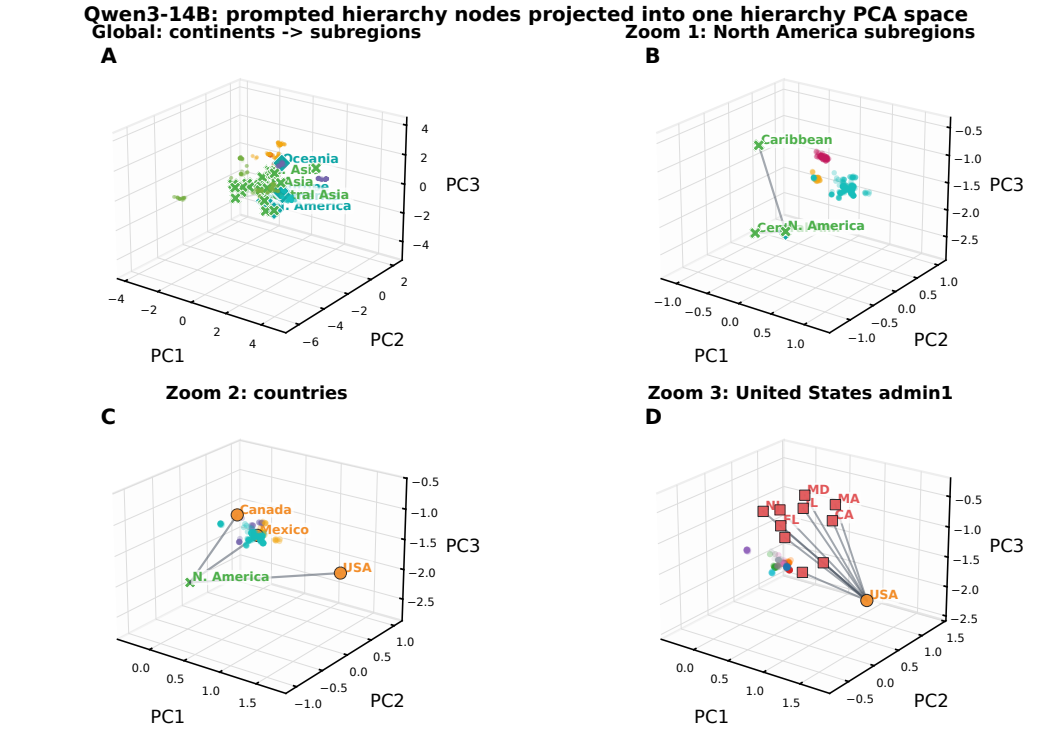

Figure 20: Hierarchy-PCA view for Qwen3-14B.

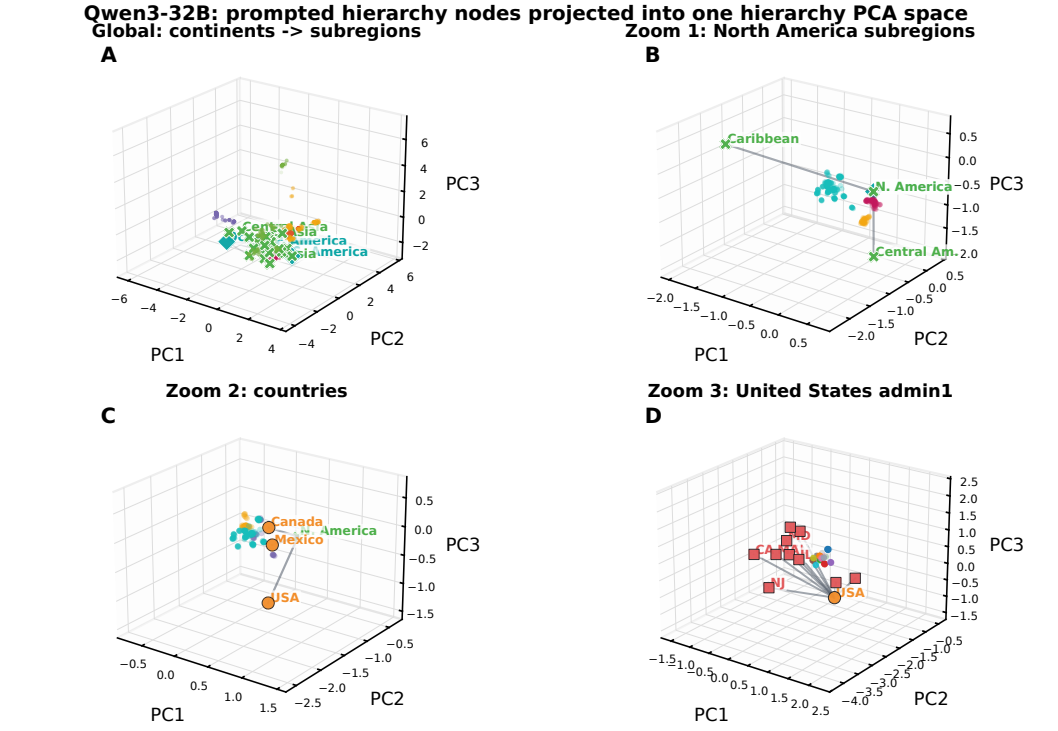

Figure 21: Hierarchy-PCA view for Qwen3-32B.

on the original small/mid Qwen/Llama checkpoints, while Table 16 reports the new large-model split-seed admin1→country summaries. Admin1→country generalizes above chance across the series, with Qwen-2.5-32B/72B strongest among the new large checkpoints. Place→admin1 sits at chance, country→subregion remains below chance under strict parent-out splits, and subregion→continent is constrained by the six-candidate parent set. The summary medians are reported in Table 3 as the Parent-disjoint @1 column.

Table 15: Parent-disjoint hierarchy-edge recovery on the original small/mid Qwen/Llama checkpoints. "IID" is stratified 5-fold cross-validation (=Cond.@1 in Table 3); "Parent-out" is `GroupKFold` on parent labels with within-context retrieval; "Region-out" is leave-continent-out. Chance for the conditional candidate set: place→admin1 0.071, admin1→country 0.112, country→subregion 0.263, subregion→continent 0.167. Bold entries mark cases where the parent-disjoint accuracy clears chance. For admin1→country, the bracketed numbers are split-seed 95% confidence intervals on the per-seed-mean accuracy across five different `GroupKFold` country-partition seeds ($\{42, 7, 13, 1234, 9001\}$); each seed produces a different partition of the 226 countries into five disjoint held-out folds. This estimates uncertainty over the choice of country split, not a true bootstrap over rows. All small/mid models have lower confidence bounds above chance (0.112); this is the surviving relational claim.

Edge	Model	IID ↑	Parent-out ↑	Region-out ↑
Place→admin1	Llama-3.1-8B	0.198	0.058	0.075
	Llama-3.2-3B	0.189	0.061	0.068
	Qwen-2.5-0.5B	0.098	**0.075**	0.075
	Qwen-2.5-1.5B	0.136	**0.075**	0.075
	Qwen-2.5-3B	0.136	0.065	0.067
	Qwen-2.5-7B	0.163	0.067	0.076
	Qwen3-4B	0.116	0.063	0.069
	Qwen3-8B	0.133	0.067	0.075
Admin1→country	Llama-3.1-8B	0.796	**0.167** [.149, .185]	0.256
	Llama-3.2-3B	0.759	**0.169** [.160, .178]	0.206
	Qwen-2.5-0.5B	0.448	**0.144** [.132, .156]	0.184
	Qwen-2.5-1.5B	0.575	**0.177** [.172, .181]	0.223
	Qwen-2.5-3B	0.619	**0.202** [.197, .206]	0.247
	Qwen-2.5-7B	0.675	**0.211** [.198, .225]	0.219
	Qwen3-4B	0.554	**0.171** [.165, .176]	0.203
	Qwen3-8B	0.605	**0.175** [.169, .181]	0.218
Country→subreg.	Llama-3.1-8B	0.841	0.142	0.296
	Llama-3.2-3B	0.845	0.137	0.327
	Qwen-2.5-0.5B	0.664	0.150	0.327
	Qwen-2.5-1.5B	0.726	0.155	0.270
	Qwen-2.5-3B	0.765	0.111	0.292
	Qwen-2.5-7B	0.673	0.119	0.261
	Qwen3-4B	0.748	0.142	0.323
	Qwen3-8B	0.708	0.133	0.208
Subreg.→continent	Llama-3.1-8B	0.750	0.000	0.000
	Llama-3.2-3B	0.792	0.000	0.000
	Qwen-2.5-0.5B	0.875	0.000	0.000
	Qwen-2.5-1.5B	0.792	0.083	0.083
	Qwen-2.5-3B	0.750	0.000	0.000
	Qwen-2.5-7B	0.792	**0.208**	**0.208**
	Qwen3-4B	0.750	**0.208**	**0.208**
	Qwen3-8B	0.750	0.000	0.000

Skipped versus chained ancestor recovery. The main text also tests whether coarse ancestors can be retrieved through skipped maps from child residuals, rather than only by composing adjacent maps. Table 17 gives the per-model version of the place→country comparison summarized in Table 6. SAP recovery is stronger than the CHP path in every Qwen/Llama series checkpoint, with positive gaps from 0.054 to 0.137. Large Llama checkpoints have the strongest SAP/CHP performance, while large Qwen3 remains weaker. This supports the interpretation that country-level information is available

Table 16: Large-model split-seed admin1→country parent-disjoint generalization. Values are per-seed means across five `GroupKFold` country-partition seeds for the large Qwen/Llama checkpoints. Fold-CI summarizes the distribution over all fold accuracies and is used as an uncertainty diagnostic for the choice of country split. Chance is 0.112.

Model	Mean ↑ ± seed s.d. ↓	Min ↑	Max ↑	Fold-CI ↑
Qwen-2.5-14B	0.147 ± 0.010	0.136	0.163	[0.073, 0.253]
Qwen-2.5-32B	0.198 ± 0.007	0.188	0.205	[0.082, 0.378]
Qwen-2.5-72B	0.199 ± 0.009	0.188	0.209	[0.073, 0.392]
Qwen3-14B	0.166 ± 0.009	0.152	0.175	[0.068, 0.312]
Qwen3-32B	0.180 ± 0.012	0.164	0.194	[0.072, 0.335]
Llama-3.1-70B	0.148 ± 0.005	0.140	0.153	[0.048, 0.246]
Llama-2-70B	0.135 ± 0.007	0.127	0.145	[0.053, 0.230]

in place states through skipped ancestor maps, while adjacent hierarchy composition remains only partial.

Table 17: Per-model comparison of SAP place→country recovery against a CHP place→admin1→country map on GPH-30K. Numbers are final-layer context-conditioned top-1 retrieval accuracy. The chance column is the mean conditional candidate-set chance.

Model	SAP Cond.@1 ↑	CHP Cond.@1 ↑	SAP–CHP gap ↑	Chance ↓
Qwen-2.5-0.5B	0.448	0.340	0.108	0.124
Qwen-2.5-1.5B	0.575	0.509	0.066	0.124
Qwen-2.5-3B	0.612	0.538	0.074	0.124
Qwen-2.5-7B	0.649	0.581	0.068	0.124
Qwen-2.5-14B	0.689	0.635	0.054	0.124
Qwen-2.5-32B	0.675	0.612	0.063	0.124
Qwen-2.5-72B	0.691	0.632	0.059	0.124
Qwen3-4B	0.512	0.427	0.086	0.124
Qwen3-8B	0.560	0.423	0.136	0.124
Qwen3-14B	0.522	0.386	0.136	0.124
Qwen3-32B	0.480	0.352	0.128	0.124
Llama-3.2-3B	0.722	0.662	0.061	0.124
Llama-3.1-8B	0.777	0.712	0.065	0.124
Llama-3.1-70B	0.792	0.724	0.068	0.124
Llama-2-70B	0.786	0.725	0.061	0.124

E.2 SELECTIVE SUBSPACE ERASURE FOR METRIC SPATIAL VS. HIERARCHICAL PLATIAL CODES

The parent-disjoint relational test in Table 3 shows the admin1→country edge generalizes to held-out parents within context. A natural follow-up asks whether this surviving relational claim is genuinely separable from the metric coordinate code or whether it is an artifact of how strongly geographic structure dominates the residual stream. We operationalize this with an interventional probe: fit low-rank *metric-spatial* and *hierarchical-platial* subspaces from training residuals, project place / admin1 / country residuals orthogonally to each subspace, and re-evaluate the same probes on the projected residuals. A random rank-matched subspace serves as a noise control.

Subspace fitting via iterative null-space projection. We use *iterative null-space projection* (Ravfogel et al., 2020) in two flavours. The metric INLP fits a held-out lat/lon ridge probe on 50% of place residuals, takes the latitude and longitude coefficient vectors as the next two directions to remove, projects all residuals orthogonal to the cumulative subspace, and repeats until the held-out coord R^2 drops below 0.05 or a model-specific cap (150–220 iterations) is reached. The resulting metric subspaces have ranks 406, 250, and 290 for Qwen-2.5-7B, Qwen3-8B, and Llama-3.1-8B. The hierarchy INLP fits a 226-class country one-vs-rest ridge classifier in scaled space and removes the top-4 right singular vectors of the weight matrix per iteration, up to 60 iterations or until country accuracy drops below 0.20. This yields rank-240 hierarchy subspaces, exceeding the linear dimension

needed to span a 226-class classifier. The random control is a Haar-uniform Gaussian subspace rank-matched to the larger of the metric and hierarchy subspaces for each model.

Selective dissociation. Table 18 reports the result on three models spanning two families (Qwen-2.5-7B, Qwen3-8B, Llama-3.1-8B). Three patterns are consistent across all three models: **(i)** INLP metric erasure eliminates linear coordinate recovery in the final 5-fold evaluation (mean R^2 falls from $+0.690$–$+0.800$ to -0.131–-0.061), but *preserves country, admin1, and admin1→country parent-disjoint recovery*. Country changes by at most 0.019 absolute, admin1 is unchanged at three decimals, and the parent-disjoint relational map never drops by more than 0.003 absolute. Coarse continent accuracy does drop, which is expected because continent is strongly collinear with latitude/longitude; the claim concerns the surviving country/admin1 and relational readouts. **(ii)** Rank-240 hierarchy erasure weakens country accuracy by about 10–12% relative but does not erase it, indicating that country information is not contained in a small easily nulled linear subspace. **(iii)** The rank-matched random subspace produces no detectable coordinate or hierarchy degradation. We read the clean metric-erasure result as causal evidence that the surviving admin1→country relational claim is not a downstream consequence of linear coordinate predictability.

Why the reverse direction is not equally clean. Even the deeper rank-240 hierarchy INLP leaves country accuracy far above chance (0.507–0.605) and leaves the admin1→country parent-disjoint map largely intact. The defensible claim is therefore narrow: these country directions are not sufficient to remove the hierarchical-platial information used by our probes and parent maps. We do not claim that hierarchy is unremovable in principle. A cleaner bidirectional double-dissociation would require stronger adversarial methods such as RLACE (Ravfogel et al., 2022) or nonlinear ablations. The reported half-dissociation is the strongest causal selectivity statement supported by low-rank linear erasure here.

Table 18: Selective subspace erasure on three representative Qwen/Llama checkpoints. "Coord R^2" is the mean of latitude and longitude held-out R^2 from a 5-fold ridge probe (negative values mean worse than predicting the mean). Cont./Subreg./Country/Admin1 are 5-fold probe accuracies on the same residuals. "A1→C PD" is admin1→country parent-disjoint Cond.@1 with within-context retrieval, mean across three `GroupKFold` country-partition seeds. Erasure is the orthogonal projection of place, admin1, and country residuals against the listed subspace. Bold entries highlight the cells that show the selective dissociation: metric erasure attenuates coordinate recovery while country, admin1, and the admin1→country relational map are largely preserved.

Model	Erasure (rank)	Coord R^2 ↑	Cont. ↑	Subreg. ↑	Country ↑	Admin1 ↑	A1→C PD ↑
Qwen-2.5-7B	none	$+0.735$	0.854	0.760	0.581	0.224	0.208
	INLP-metric (k=406)	-0.061	0.614	0.705	0.562	0.224	0.219
	INLP-hier (k=240)	$+0.649$	0.839	0.730	0.519	0.209	0.217
	random (k=406)	$+0.734$	0.853	0.757	0.576	0.221	0.211
Qwen3-8B	none	$+0.690$	0.838	0.740	0.574	0.216	0.174
	INLP-metric (k=250)	-0.131	0.623	0.706	0.563	0.216	0.171
	INLP-hier (k=240)	$+0.607$	0.823	0.704	0.507	0.209	0.179
	random (k=250)	$+0.689$	0.838	0.739	0.572	0.216	0.177
Llama-3.1-8B	none	$+0.800$	0.895	0.830	0.673	0.259	0.153
	INLP-metric (k=290)	-0.073	0.717	0.805	0.672	0.259	0.172
	INLP-hier (k=240)	$+0.754$	0.890	0.812	0.605	0.248	0.147
	random (k=290)	$+0.798$	0.895	0.829	0.673	0.260	0.153

E.3 RANK SEPARATES COARSE AND FINE HIERARCHY EDGES

The edge-recovery setup also lets us ask whether all hierarchy edges use the same amount of linear capacity. At the final layer, we repeat the parent-conditioned recovery with rank-constrained maps ($r \in \{8, 32, 128, 256\}$). Table 19 reports large-model chance-normalized conditional retrieval. The sweep separates coarse and fine structure: place→admin1 remains hard even at rank 256 ($0.004 \rightarrow 0.100$), while admin1→country improves strongly ($0.169 \rightarrow 0.669$), place→country improves ($0.272 \rightarrow 0.645$), and coarser continent/subregion ancestors become strong at moderate

rank. The subregion→continent edge is already high at rank 8 because it has only six continent candidates, so we treat it as a small-candidate-set ceiling rather than as the main rank-discriminative evidence. The rank profile is stable with scale: higher rank helps middle and coarse hierarchy edges strongly, but fine local administrative assignment remains difficult.

Table 19: Rank sweep for final-layer parent-conditioned hierarchy recovery on the large Qwen/Llama checkpoints. Entries are median chance-normalized conditional retrieval across the 14B–72B/70B subset.

Edge	$r = 8 \uparrow$	$r = 32 \uparrow$	$r = 128 \uparrow$	$r = 256 \uparrow$
Place→admin1	0.004	0.026	0.064	0.100
Admin1→country	0.169	0.336	0.579	0.669
Country→subreg.	0.099	0.369	0.610	0.640
Subreg.→cont.	0.650	0.700	0.700	0.700
Place→country	0.272	0.405	0.578	0.645
Place→subreg.	0.327	0.623	0.753	0.783
Place→cont.	0.629	0.765	0.838	0.849
Admin1→subreg.	0.166	0.543	0.792	0.826
Admin1→cont.	0.580	0.775	0.866	0.888
Country→cont.	0.538	0.798	0.851	0.857

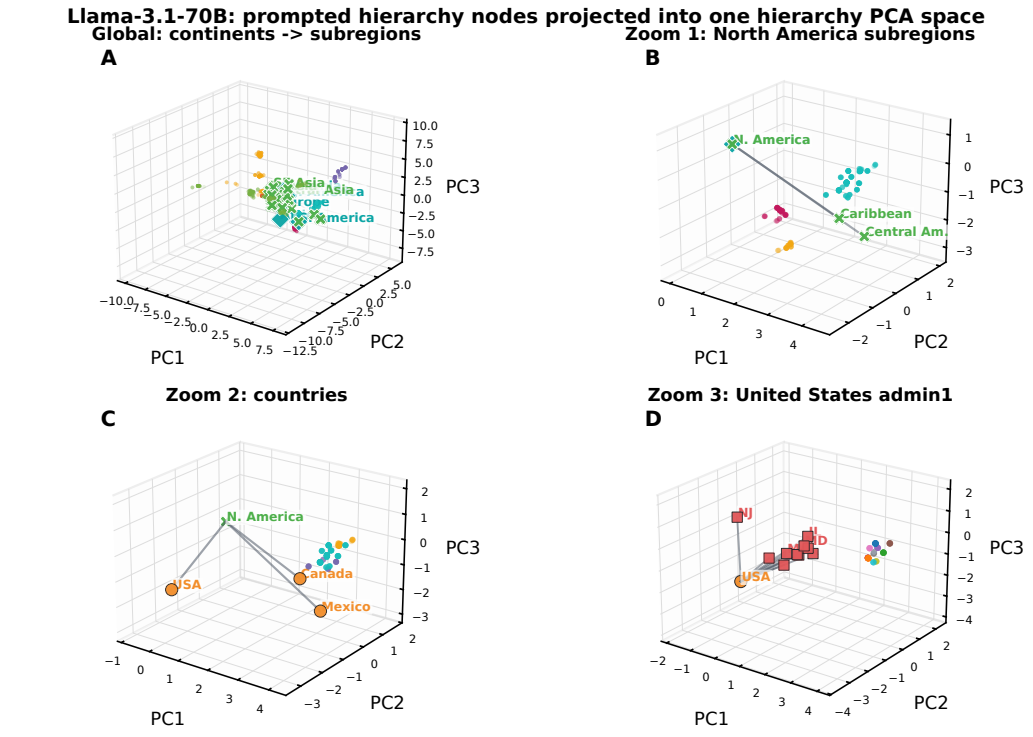

Figure 22: Hierarchy-PCA view for Llama-3.1-70B.

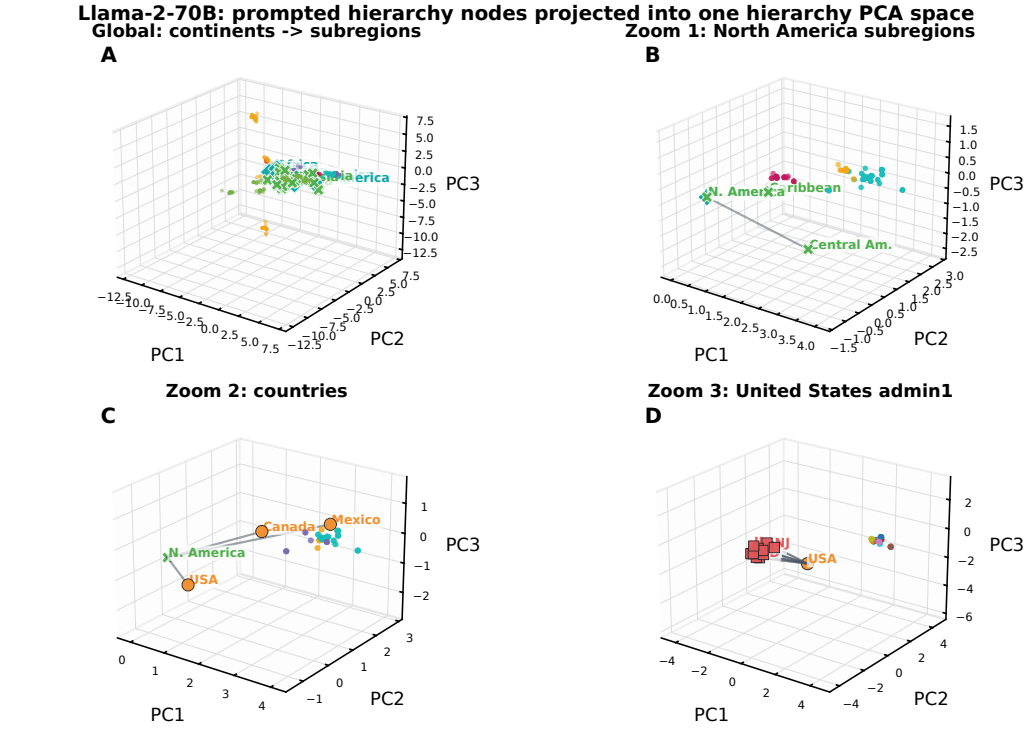

Figure 23: Hierarchy-PCA view for Llama-2-70B.