

Relation-State Continuity in Qwen2.5-7B Base: Evidence for an Attention-Topology Path

Chen Ho Yiing (Norika Oda)
Independent Researcher
ORCID: 0009-0006-6816-9891

Draft v0.2
2026-05-30

Abstract

Qwen2.5-7B base, eight relation-tracking cases, 29 layers. Residual carry AUC 0.55, attention anchor mass AUC 0.089, formation peak 0.94, cross-layer circuit score 0.39. The formation profile has two peaks (layer 1 and layer 28) and no single intermediate layer dominates. The work of routing anchor information to query position is spread across layers 19 through 27, where residual carry rises from 0.60 to 0.81 while attention to anchors stays in the 0.12 to 0.23 range. I label this pattern *attention topology path*. The classification is observational; causal verification by activation patching is the next experiment.

Keywords: mechanistic interpretability, relation tracking, Qwen2.5-7B, residual stream, attention anchors, topology path, layer attribution

1 The question

A model reads “Alice gave Bob the key.” Later it answers “Who has the key?”. The relation came in early; the answer comes out late. By what path does it travel?

Two candidate stories. *Localized layer formation*: a single transformer layer integrates the relation, and the rest of the network propagates the answer. *Attention topology path*: multiple layers and attention edges route information from anchor to query, with residual stream carrying state between them. The two predict different things and should produce different observation profiles.

I probed Qwen2.5-7B base with eight short relation cases. The data says: not localized.

2 Related Work

Three threads. First is residual stream analysis as the substrate of circuit computation [1, 2]. Second is attention pattern analysis. Third is the dense activation probing line in my prior work [3, 4, 5], where per-channel observation tooling supports cross-layer hypotheses. The coverage metric used here was established in [8], and per-layer fingerprint comparison in [6, 7].

This paper turns the dense observation toolkit on one question, on one model. Eight cases. No cross-architecture sweep.

3 Method

3.1 Notation

A model M has L layers, $\ell \in \{0, 1, \dots, L\}$. The residual stream at layer ℓ at token position t is $h_\ell^{(t)} \in \mathbb{R}^d$. Qwen2.5-7B: $L = 28$, $d = 3584$.

A *case* is a triple $\mathcal{C} = (P, A, Q)$. P is the prompt of length T , $A \subset \{1, \dots, T\}$ is the anchor region (relation introduced), $Q \subset \{1, \dots, T\}$ is the query region (relation applied). I enforce $\max A < \min Q$.

3.2 Per-layer metrics

Four scalars per case per layer, each with a one-line definition (detail in the equations).

Residual carry ρ_ℓ .

$$\rho_\ell = \cos(\bar{h}_\ell^{(A)}, \bar{h}_L^{(Q)}), \quad \bar{h}_\ell^{(R)} := \frac{1}{|R|} \sum_{t \in R} h_\ell^{(t)}. \quad (1)$$

Cosine alignment between layer- ℓ anchor representation and layer- L query representation. High ρ_ℓ means the anchor already points where the model ends up.

Formation delta Δ_ℓ .

$$\Delta_\ell = \rho_\ell - \rho_{\ell-1}, \quad \Delta_0 := 0. \quad (2)$$

Per-layer increment in residual carry. Under localized formation a single Δ_{ℓ^*} dominates, while under a topology path the Δ_ℓ values spread across many layers.

Attention anchor mass m_ℓ .

$$m_\ell = \frac{1}{H \cdot |Q|} \sum_{h=1}^H \sum_{q \in Q} \sum_{a \in A} \alpha_{q,a}^{(h,\ell)}, \quad (3)$$

where $\alpha_{q,a}^{(h,\ell)}$ is the softmax attention weight from query position q to anchor position a in head h at layer ℓ . H is heads per layer. Layers that route query to anchor have high m_ℓ .

Attention entropy E_ℓ .

$$E_\ell = - \frac{1}{H \cdot |Q|} \sum_{h=1}^H \sum_{q \in Q} \sum_{t=1}^T \alpha_{q,t}^{(h,\ell)} \log \alpha_{q,t}^{(h,\ell)}. \quad (4)$$

Sanity check: high m_ℓ with low E_ℓ means the attention is decisively routing to anchors, not diffusing.

3.3 Aggregate scores

Average per-case curves to get mean curves $\bar{\rho}_\ell$, $\bar{\Delta}_\ell$, $\bar{m}_\ell$. From them four summaries.

$$\text{AUC}_\rho = \frac{1}{L} \sum_{\ell=0}^{L-1} \bar{\rho}_\ell, \quad (5)$$

$$\text{AUC}_m = \frac{1}{L} \sum_{\ell=1}^L \bar{m}_\ell, \quad (6)$$

$$\Pi = \max_{\ell \in \{0, \dots, L\}} \bar{\Delta}_\ell, \quad (7)$$

$$\text{CLC} = \text{corr}\left(\{\bar{\rho}_\ell\}_{\ell=2}^{L-1}, \{\bar{m}_\ell\}_{\ell=2}^{L-1}\right). \quad (8)$$

The interesting one is CLC. It correlates residual carry with attention anchor mass across non-boundary layers. Localized formation: low CLC (residual jump independent of attention). Topology path: high CLC (residual rise tracks attention activity).

3.4 Verdict rule

$$\text{Verdict} = \begin{cases} \text{localized layer} & \text{if } \Pi \geq 0.95 \text{ and } \arg \max_\ell \bar{\Delta}_\ell \in \{2, \dots, L-2\}, \\ \text{attention topology path} & \text{if } \bar{\Delta}_\ell \text{ is bimodal and } \text{CLC} > 0.3, \\ \text{undetermined} & \text{otherwise.} \end{cases} \quad (9)$$

“Bimodal” here means the second-largest $\bar{\Delta}_\ell$ is at least half the largest, and the two peaks are separated by at least three layers. This rule labels the observation profile. It does not claim a causal mechanism.

3.5 Cases

Eight short English prompts. Each one introduces a relation early and queries it late. Prompt lengths 33–44 tokens, anchor 8–17 tokens, query 7–11 tokens. Each case returns a one-token answer. I record the model’s answer logit margin (the gap between the correct token and the top incorrect token) as a behavioural sanity check.

Table 1: The eight cases.

Case	Prompt	Anchor	Query	Margin
owner_transfer_key	44	17	11	4.75
promise_delegate	41	15	9	3.97
permission_revoked	39	13	7	5.96
state_location	39	12	7	4.69
role_successor	38	15	8	5.81
debt_assignment	34	11	8	6.49
trust_boundary	35	8	7	5.14
thread_state	33	9	8	5.00

Every case has positive answer logit margin (range 3.97–6.49). The model can answer all eight above chance.

4 Results

4.1 Four numbers

$$\text{AUC}_\rho = 0.5499, \quad \text{AUC}_m = 0.0890, \quad \Pi = 0.9385, \quad \text{CLC} = 0.3912.$$

4.2 Top layers

Each metric’s top eight layers (Table 2).

Table 2: Top eight layers by each metric.					
Residual carry $\bar{\rho}_\ell$		Anchor mass $\bar{m}_\ell$		Formation $\bar{\Delta}_\ell$	
Layer	Value	Layer	Value	Layer	Value
27	0.8091	19	0.2278	1	0.9385
26	0.8053	22	0.2173	28	0.8165
25	0.7758	20	0.2159	4	0.1935
24	0.7183	27	0.1976	9	0.1595
28	0.6737	18	0.1824	8	0.1419
23	0.6587	26	0.1488	10	0.1358
22	0.6336	23	0.1275	2	0.1239
20	0.6003	24	0.1246	6	0.1117

Three patterns.

Residual carry stacks in the late layers: 0.81 at layer 27, dropping smoothly through 0.72 at layer 24, still 0.60 at layer 20.

Attention anchor mass concentrates in middle layers 19–22, with a secondary spike at layer 27. The 19–22 block is where the model is actively reading anchor positions.

Formation delta is bimodal, with layer 1 jumping 0.94 and layer 28 jumping 0.82. The third-highest delta is only 0.19 at layer 4, an order of magnitude below either peak. No single intermediate layer is doing the formation work.

4.3 The verdict

Applying Eq. (9) to these numbers:

- $\Pi = 0.9385 < 0.95$, so the strict localized threshold is not met.
- Formation is bimodal. Layer 1 (0.94) and layer 28 (0.82) are separated by 27 layers. The second peak is 0.87 of the first.
- $\text{CLC} = 0.39 > 0.30$.

Verdict: attention topology path.

5 What’s actually happening

Layer 1 looks dramatic but does not compute the relation. The $\bar{\Delta}_1 = 0.94$ jump comes from the embedding-plus-positional encoding settling into the model’s residual coordinate system. Anchor and query already share positional information immediately after embedding lookup, so cosine

alignment rises sharply at the first layer. This jump is a baseline artifact of how ρ_ℓ is defined at the boundaries.

Layer 28 plays a similar boundary role on the other side, contributing the final read-out into the unembedding subspace.

The actual work happens in layers 19 through 27. In that band, residual carry rises steadily from 0.60 to 0.81. In the same band, attention anchor mass stays in the 0.12–0.23 range, with a particular concentration in 19–22. No single layer in this band does the formation; the formation is distributed.

CLC = 0.39 matches the picture. The residual carry rises track attention anchor activity but not one-for-one; other factors contribute, and the topology path is one driver among several.

6 What this evidence does and does not support

It supports: on Qwen2.5-7B, on these eight cases, relation continuity is not concentrated in a single mid-network layer.

It does not support a causal claim. The profile is observational only. Ablating any single layer in band $\{19, \dots, 27\}$ should leave the relation answerable; ablating the band as a whole should break it. That experiment is the next one.

It also does not generalise beyond this setup. One model, eight cases, hand-designed prompts: a larger procedurally generated case set and replication on other base models are required before this can be called a property of large language models in general.

7 Limitations

1. Only Qwen2.5-7B base was probed in this study. No cross-architecture or cross-scale replication has been performed.
2. The case set contains eight hand-designed prompts, which is small enough that selection bias cannot be ruled out.
3. All results are observational; causal direction between the layer-band activity and answer correctness has not been established. Activation patching is the natural next experiment.
4. Boundary artifacts dominate $\bar{\Delta}_1$ and $\bar{\Delta}_{28}$. A future revision of the verdict rule should subtract a layer-0 baseline before peak detection, so the bimodal pattern reflects interior layer behaviour and not embedding or unembedding effects.
5. The cosine metric in Eq. (1) captures direction not magnitude, so two layers with identical ρ_ℓ may differ in how the relation is stored in residual norm.
6. Results are from a single decoding seed; averaging across multiple seeds is left to follow-on work.

8 Reproducibility

- Probe tool: relation-state continuity tool, available with the project codebase.
- Output schema: `relation-state-continuity-probe-v1`.
- Per-case raw curves: 29-layer arrays for each of ρ , Δ , m , E , retained per case.
- Aggregate scores: per Eqs. (5)–(8).

- Hardware: single x86 CPU host, model held in bfloat16. Wall time approximately 36 minutes.

Exact paths and case prompt files available on request.

9 Follow-on

Five experiments are queued for the next iteration.

1. Activation patching as the causal test: ablate the layer subsets $\{1\}$, $\{19, \dots, 22\}$, $\{23, \dots, 27\}$, and $\{28\}$ separately, and measure the resulting drop in answer logit margin for each ablation.
2. Cross-architecture: replicate on Llama-3-8B, Mistral-7B, Phi-2.
3. Boundary-corrected verdict: subtract layer-0 baseline from all $\bar{\Delta}_\ell$.
4. Scale-up cases: 100+ procedurally generated relation prompts.
5. Per-head attribution within layers 19–22: which individual heads carry the anchor mass.

10 Acknowledgments

A single CPU host. Qwen2.5-7B base from the publicly distributed checkpoint. The relation cases were written for this study.

References

- [1] Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread, Anthropic, October 2023. <https://transformer-circuits.pub/2023/monosemantic-features>.
- [2] Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy Models of Superposition. Transformer Circuits Thread, Anthropic, September 2022. https://transformer-circuits.pub/2022/toy_model/.
- [3] Chen, Ho Yiing. Mercury: An addressable million-cell sensor grid for browser-accessible LLM observability on consumer hardware. Zenodo, 2026. DOI: 10.5281/zenodo.20313154.
- [4] Chen, Ho Yiing. Mercury-Viewer: A self-contained browser-based interactive 3D visualization of LLM internal sensor maps. Zenodo, 2026. DOI: 10.5281/zenodo.20313150.
- [5] Chen, Ho Yiing. Mercury MCP v0.1.0 — Cross-Architecture LLM Internal Observation Database for 23 LLMs. Zenodo, 2026. DOI: 10.5281/zenodo.20352085.
- [6] Chen, Ho Yiing. Cross-Architecture Per-Layer Residual Stream Alignment via Tier-B Fingerprints: 84 Model-Pair Matrices Across 9 Open LLMs Show a Universal Middle-Depth Slot. Zenodo, 2026. DOI: 10.5281/zenodo.20366010.

- [7] Chen, Ho Yiing. Tier-A Vocab-Hash Artifact in Cross-Architecture Hot-Dim Studies: Why Output-Layer Token-ID Projection Manufactures False Universality, and an Artifact-Aware Screening Protocol. Zenodo, 2026. DOI: 10.5281/zenodo.20366134.
- [8] Chen, Ho Yiing. Coverage Failure as a Diagnostic Signal in Neural Model Probing. Zenodo, 2026. DOI: 10.5281/zenodo.20448669.