

# Cross-Model Dimensional Preservation in qwen2.5: 49 Hot Coordinates Survive Across 3B and 7B

Chen, Ho Yiing<sup>1</sup>

<sup>1</sup>Independent researcher, charenix.com, ORCID: 0009-0006-6816-9891  
abbychen981008@gmail.com

2026-05-21

## Abstract

Two members of the qwen2.5 family share an output-side grammar at the per-coordinate level. Aggregating activation heat at block  $L-1$  for qwen2.5:7B ( $D_7 = 3584$ ) and qwen2.5:3B ( $D_3 = 2048$ ) on ten multilingual prompts via Mercury [1], I take each network’s hottest hidden-state indices: 200 for the larger member, a proportionally scaled 114 for the smaller. The two top-K sets  $H_{7B}$  and  $H_{3B}$  share 49 indices inside the smaller variant’s range. Under a permutation null this is  $4.4\times$  chance. Eleven of those 49 also host 7B’s constant-firing set  $U$ , defined as cells with the all-ones signature [2]. A hypergeometric test against  $D_7 = 3584$  gives  $p < 1.2 \times 10^{-12}$  for this second overlap,  $20.6\times$  chance. The eleven triple-confirmed slots  $T = \{11, 12, 25, 279, 334, 382, 476, 481, 510, 715, 758\}$  are an addressable target list for merge recipes, distillation alignment, and per-axis interpretability probes. Code, raw binaries, and the underlying grid spec are open [3].

## 1 Introduction

The phrase “model family” usually means weights initialized from a shared base or trained on overlapping data. It rarely says anything quantitative about whether two members preserve the same internal addresses for the same kinds of inputs.

If they did, that would be useful.

It would say something about the strength of training-induced positional structure inside transformer state. Practically, it would give merge-kit recipes a target list to protect, distillation a guide for which channels to align, and interpretability work a between-variant anchor.

The cleanest test on consumer hardware is this. Two networks from one family. One observation pass each, identical prompts. For each member, aggregate block- $L-1$  heat per axis and rank. Take the hottest top-K, with K scaled so the top-K fraction matches. Intersect the two sets in the smaller range. Compare against a permutation null.

I do that for qwen2.5:7B ( $D_7 = 3584$ , top-200) and qwen2.5:3B ( $D_3 = 2048$ , top-114). The shared set has size 49.

Then a second test against 7B’s own internal structure. Take  $U$  to be the cells with the all-ones firing signature on the ten prompts [2]; 39 cells, on 39 distinct positions in 7B’s output block. Eleven of the 49 between-variant slots land on those 39.

Triple-confirmed.

Hot in 7B aggregate, hot in 3B aggregate, and constantly firing across the prompt panel in 7B.

This short paper reports those numbers, the procedure that produced them, the null distribution, the hypergeometric significance test, and a concrete coordinate list.

## 2 Method

### 2.1 Observation grids

Both networks are observed through a Mercury sensor grid [1]. The 7B grid has  $L = 28$  blocks, hidden  $D = 3584$ , and  $Q = 5$  quantile buckets per (block, axis) pair, totaling 501,760 cells. The 3B grid has  $L = 36$ ,  $D = 2048$ ,  $Q = 5$ , for 368,640 cells.

A cell fires when the corresponding activation exceeds a calibrated quantile threshold for its bucket. One observation pass through ten multilingual prompts produces, per prompt, a list of fired cell IDs written to a binary file. Cells that never fire are excluded.

### 2.2 Hot-axis extraction at each width

In 7B, read the slice of the heat array corresponding to block  $L-1$  (all  $D \cdot Q$  cells at the output). For each axis  $d \in \{0, \dots, D-1\}$ , sum heat across the  $Q$  buckets at  $d$  to get a scalar  $h_d$ . The 200 axes with largest  $h_d$  form the hot-axis set  $H_{7B} = \text{argsort}_d h_d [: 200]$ . Inside 3B, the analogous operation runs on its own block- $L-1$  slice with  $D = 2048$  and  $Q = 5$ , but with the cutoff scaled so the hot fraction matches: with 7B taking  $200/3584 = 5.58\%$ , the corresponding count for 3B is  $\lfloor 0.0558 \cdot 2048 \rfloor = 114$ . The shared set  $H_{7B} \cap H_{3B}$  is then computed in the smaller coordinate range  $\{0, \dots, 2047\}$ , since any 7B index beyond 2047 cannot match a 3B index by construction; the larger member’s hot list is effectively truncated to its first- $D_3$  portion before the comparison, and the smaller member’s hot list enters in full.

### 2.3 Set $U$ extraction in 7B

Inside 7B only,  $U$  follows the procedure of the discovery paper [2]. Each cell’s signature is a 10-bit string indicating which of the ten prompts caused it to fire. Cells with signature 1111111111 are the constant-firing members: they spike on every prompt regardless of topic or surface language. The discovery paper reports 39 such cells in 7B [2]. I rederive that count from the raw binary fired-cell files. The 39 cells live on 39 distinct hidden-state coordinates (no two share an axis here).

Let  $U \subseteq \{0, \dots, D_7 - 1\}$  be that 39-element set.

### 2.4 Triple-confirmed slots and statistical test

The set of triple-confirmed slots is

$$T = H_{7B} \cap H_{3B} \cap U,$$

the between-variant shared set further intersected with  $U$ . The two operations are nested in different ranges:  $H_{7B} \cap H_{3B}$  in  $[0, 2047]$ , and the further intersection with  $U$  in  $[0, 3583]$ , where every triple-confirmed address automatically falls below 2048 because it has already passed the first filter.

The first overlap is compared against a permutation null where both top-K sets are drawn uniformly at random in the smaller range:

$$\mathbb{E}[|H_{7B} \cap H_{3B}|]_{\text{null}} \approx \frac{200 \cdot 114}{2048} \approx 11.13.$$

The second overlap calls for a hypergeometric test. Of the 49 between-member indices, how many fall into  $U$  if the 49 were drawn uniformly at random from  $D_7 = 3584$  slots? The mass function is

$$P(X = k) = \frac{\binom{|U|}{k} \binom{D_7 - |U|}{49 - k}}{\binom{D_7}{49}},$$

and the upper-tail  $p$ -value is  $\sum_{k=k_{\text{obs}}}^{\min(|U|, 49)} P(X = k)$ .

### 3 Results

#### 3.1 Variant-to-variant overlap

| Quantity                       | Value                |
|--------------------------------|----------------------|
| $D_7$ (7B hidden width)        | 3,584                |
| $D_3$ (3B hidden width)        | 2,048                |
| $ H_{7B} $ (7B top-K hot axes) | 200                  |
| $ H_{3B} $ (3B top-K hot axes) | 114                  |
| $ H_{7B} \cap H_{3B} $         | <b>49</b>            |
| Permutation-null expectation   | 11.13                |
| Observed / expected ratio      | <b>4.40</b> $\times$ |

Forty-nine.

Two top-K sets, one drawn from a 7B network and one from a 3B network of one family, agree on 49 hidden-state indices inside a 2048-wide range. The permutation null sits at 11.13. Ratio 4.40 $\times$ .

#### 3.2 Triple-overlap with $U$

| Quantity                               | Value                 |
|----------------------------------------|-----------------------|
| $ U $ (7B constant-firing coordinates) | 39                    |
| $ T  =  H_{7B} \cap H_{3B} \cap U $    | <b>11</b>             |
| Hypergeometric null $\mathbb{E}[X]$    | 0.533                 |
| Observed / expected ratio              | <b>20.6</b> $\times$  |
| Upper-tail hypergeometric $p$ -value   | $1.2 \times 10^{-12}$ |

The eleven triple-confirmed slots are

$$T = \{11, 12, 25, 279, 334, 382, 476, 481, 510, 715, 758\}.$$

These are addressable integer coordinates satisfying all three filters: hot in 7B’s  $L-1$  aggregate heat, hot in 3B’s analogous aggregate (after proportional top-K scaling), and active on every one of the ten observation prompts when looked up in  $U$ .

#### 3.3 What 11 in 49 means

Eleven over 49 is  $\approx 22.4\%$ . Were  $U$  and the between-variant pair independent, the expected fraction would be  $39/3584 \approx 1.1\%$ . Observed enrichment is therefore not just statistically significant on its absolute count; it indicates that between-member sharing is preferentially concentrated in channels the larger network keeps constantly active. Put differently: hot channels that survive a scale change are disproportionately the same ones the larger network never lets go quiet, and the conditional probability of a between-member hot channel also landing in  $U$  is roughly  $22.4\%/1.1\% \approx 20\times$  what independence would predict —the same enrichment captured by the hypergeometric test but stated as a ratio of conditional probabilities, which is the form mergekit and distillation practitioners can use directly when prioritizing which channels to protect or align.

That ratio.

That is the number to remember.

### 4 Discussion

The first overlap (49 indices, 4.4 $\times$  chance) says that hot-index selection is not arbitrary at the per-position level. Two members of one family, separately trained, allocate disproportionate firing energy to one

shared handful of hidden-state slots. This is a structural fact about training, not about any particular prompt.

The second overlap (11 indices,  $20.6\times$  chance,  $p < 10^{-12}$ ) sharpens that picture. Between-variant sharing is preferentially located inside 7B’s constant-firing set  $U$ . The list  $T$  is a candidate target for any work that wants to identify the slots carrying universal compute in qwen2.5.

Three practical readings:

1. Mergekit-style [5] weight surgery between qwen2.5 members. The 49 hot indices shared between members, and especially the eleven triple-confirmed slots, are weights that should be protected by low merge temperature, since perturbations there will affect every input.
2. Distillation from 7B to 3B. Alignment loss applied along these channels should give larger learning signal per loss-byte than uniform alignment, the kind of efficiency gain that compute-optimal training studies [6] keep recommending but rarely operationalize at the per-coordinate level.
3. SAE-style interpretability [4]. The eleven triple-confirmed slots are an a-priori target subset for monosemantic feature probes.

## 5 Limitations

Ten prompts is small.  $U$  is defined by exact match of the all-ones signature on those ten inputs. A broader prompt panel will shrink  $|U|$  (more chances to break constant firing), and  $|T|$  will shrink with it, though the ratio against the chance null may stay informative.

One family. Replicating across at least one other transformer family (e.g. Llama 3.x or Gemma 2) at comparable parameter counts is the natural next step. Mercury observation is architecture-aware [1] and the analysis script accepts any pair of  $(L, D, Q)$  grids without code changes.

Not yet done.

Block  $L-1$  only. Aggregation here is restricted to block  $L-1$  (Tier-A surface in the method paper [1]). Equivalent analysis at earlier blocks will likely give a different and potentially richer picture; that is on the Tier-B roadmap and not in this paper.

Top-K choice. The result is robust at  $K_{7B} = 200$  but I have not run the full ablation across  $K$ . The  $4.4\times$  ratio is expected to hold for  $K$  in the range that captures most of the right-tail heat distribution, and to weaken as  $K$  approaches the hidden width (the trivial bound).

## 6 Conclusion

Two members of the qwen2.5 family, observed through one Mercury procedure on identical multilingual prompts, share 49 hot indices at block  $L-1$ ,  $4.4\times$  chance. Eleven of those 49 also host 7B’s constant-firing set  $U$ ,  $20.6\times$  chance,  $p < 10^{-12}$  on a hypergeometric test. The list is short, addressable by integer, and reproducible from the open binaries that ship with the Mercury-Viewer dataset [3].

If you are merging qwen2.5 weights, distilling between qwen2.5 members, or probing qwen2.5 with SAE-style methods, those eleven slots are where to look first.

## Acknowledgements

Drafted in continuation of the Mercury sprint as a paired-engineering session with an Anthropic Claude Code agent. Every numeric claim in this paper traces to a git commit hash and an on-disk binary in the companion dataset [3]. Relevant commits: 8ee6256 (between-variant extraction), 3fa2c96 (set  $U$  discovery), 97302ae (the 500K observation that defines  $U$ ).

## References

- [1] 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.
- [2] Chen, Ho Yiing. Unsupervised discovery of language and topic lanes in transformer models via multilingual co-firing signatures. Zenodo, 2026. DOI: 10.5281/zenodo.20313748.
- [3] 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.
- [4] 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>.
- [5] Goddard, C., Siriwardhana, S., Ehghaghi, M., Meyers, L., Karpukhin, V., Benedict, B., McQuade, M., and Solawetz, J. Arcee’s MergeKit: A Toolkit for Merging Large Language Models. arXiv preprint, 2024. DOI: 10.48550/arXiv.2403.13257.
- [6] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training Compute-Optimal Large Language Models. arXiv preprint, 2022. DOI: 10.48550/arXiv.2203.15556.