

Internal Specialization Scales with Model Size: Quantitative Evidence from qwen2.5 3B vs 7B

Chen, Ho Yiing¹

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

2026-05-21

Abstract

Folk wisdom in the LLM community holds that bigger transformers “specialize more.” I give that claim a cell-level test. Running identical ten-prompt observation passes through qwen2.5:3B and qwen2.5:7B on a Mercury sensor grid [1] produces directly comparable signature counts at each size. The 7B model has $2.08\times$ as many cross-lingual physics-topic cells, $1.66\times$ as many Chinese-only sensors, and $1.56\times$ as many English-only sensors. It also has $0.41\times$ as many universal-backbone cells; the smaller side dominates that single category. So topical and linguistic circuits expand with parameter count while the always-on backbone shrinks. This is a cell-level reading of a claim usually argued through perplexity curves and benchmark scores. Data and viewer are open [3]; the signature extraction is documented in the companion discovery preprint [2].

1 Introduction

Scaling laws for transformer LLMs are stated in terms of loss or downstream task score as a function of parameter count, dataset size, or compute [4, 5]. That tells us how outputs improve.

It does not tell us where inside the model the improvement lives.

A separate thread of work argues that capabilities “emerge” at certain sizes [6]. Whether such transitions are real or measurement artifacts is a live debate. They share an epistemic surface: scaling read off the output side.

The question I want to put on the table is internal. When 7B handles a query better than 3B handles an identical query, is there a corresponding shift in which sensors light up?

Mercury [1] places a fixed addressable lattice over a transformer’s hidden state. The lattice spec is architecture-aware: the recipe scales to any qwen2.5 variant by reading layer count and hidden size from the GGUF metadata. Two models under one recipe yield two lattices with shared coordinate semantics, even though absolute counts differ.

Same ten prompts. Two models. Six pre-declared bit-mask categories from the discovery paper [2]. Differences in category counts are differences in how 3B and 7B internally route an identical input set.

Three findings:

1. Topical and linguistic categories carry more cells in the larger model. Strict signature counts, no smoothing.
2. The always-on backbone carries fewer cells in the larger model. 3B spends more of its lattice on input-agnostic firing.
3. Both models show the cross-lingual physics-topic group. The phenomenon is not 7B-only.

The size-dependent change is not uniform across the firing distribution. Different regions move in different directions.

That is the picture this paper draws.

2 Method

2.1 Grid placement at two scales

Each Mercury recipe is L layers by D hidden dimensions by Q quantile buckets, giving $L \cdot D \cdot Q$ addressable cells. Reading GGUF metadata:

Model	L	D	Q	Total cells
qwen2.5:3B	36	2048	5	368,640
qwen2.5:7B	28	3584	10	1,003,520

The 7B grid has $2.72\times$ the cell tally of the 3B grid. That is the denominator against which the category ratios below should be read. Any group sitting at $2.72\times$ in 7B is just tracking grid growth. At $2.08\times$, a category is shrinking relative to growth. At $0.41\times$, it is shrinking in absolute terms.

2.2 Same prompts, same category definitions

Prompts are the ten queries from the discovery paper [2]: four topic-by-tongue pairs (casual chat, algorithms, machine learning, life reflection) plus two cross-lingual physics queries (量子糾纏 in Chinese, null geodesics in English). The six categories are bit masks specified before observation. Repeating them here:

Category	Bit mask (p1..p10)
Universal backbone	1111111111
Chinese-only	1010101010
English-only	0101010101
Physics cross-lingual	0000000011
Algorithms cross-lingual	0011000000
ML cross-lingual	0000110000

A sensor is counted in a category only if its signature exactly matches the mask. Loose matches (e.g. “physics plus one other prompt”) are excluded from the strict count and reported separately as a subset analysis.

3 Results

3.1 Active sensor counts

How much of each lattice fires at all, across the ten prompts:

Quantity	7B	3B	Ratio 7B/3B
Grid total cells	1,003,520	368,640	2.72
Cells that fired	14,912	6,782	2.20
Fired fraction	1.49%	1.84%	0.81
Unique signatures	944	612	1.54

7B recruits more absolute sensors but a smaller fraction of its lattice. Language-selective and topic-selective structure can grow even as average sparsity tightens.

3.2 Category counts

Strict-signature counts at each size:

Category	7B	3B	Ratio 7B/3B
Universal backbone	39	94	0.41
Chinese-only	73	44	1.66
English-only	56	36	1.56
Physics cross-lingual	50	24	2.08
Algorithms cross-lingual	66	31	2.13
ML cross-lingual	21	11	1.91

Five categories grow with size. One shrinks.
The shrinking category is the universal backbone.

3.3 The universal-backbone inversion

39 sensors fire on every prompt in 7B. 94 do so in 3B. The smaller transformer puts more than twice the absolute count on input-agnostic firing.

As a fraction of the lattice:

$$\frac{39}{1,003,520} \approx 3.9 \times 10^{-5} \quad \text{vs} \quad \frac{94}{368,640} \approx 2.6 \times 10^{-4}.$$

3B puts about $6.5\times$ more of its lattice into the always-on category. One reading: a smaller model leans harder on broadly tuned sensors because it cannot afford as many narrowly tuned ones. Another reading: 3B universals are partially redundant with what 7B has replaced by finer-grained linguistic and topical structure. These two stories make different ablation predictions; the data here does not pick between them.

3.4 Cross-language physics shows up at both sizes

50 sensors in 7B fire on exactly the two physics queries (one Chinese, one English) and on nothing else. 24 sensors in 3B meet the same strict mask.

This cross-lingual physics abstraction is not a 7B-only phenomenon. It is present at 3B and grows by roughly $2\times$ when parameter count roughly doubles.

That matters for the headline finding of the discovery paper [2], which used only 7B data. Had the 3B tally been zero, the cross-lingual topic story would be a feature that switches on somewhere between 3B and 7B, and the threshold-effect framing of emergent abilities [6] would apply. It is not zero. The category is present at both sizes and the change is gradual rather than a step.

3.5 Subset analysis: linguistic purity at relaxed thresholds

Strict single-language masks (1010101010 and 0101010101) miss sensors that fire on most but not all queries of one side. A relaxed definition: count cells firing on at least four of five Chinese prompts and on no English prompts, and the symmetric definition for the other direction.

Relaxed category	7B	3B	Ratio 7B/3B
Chinese-leaning (zh-pure subset)	2,673	1,554	1.72
English-leaning (en-pure subset)	2,799	1,789	1.56

Relaxed counts run an order of magnitude above the strict counts, yet 7B/3B ratios stay in the 1.56–1.72 range. The strict definition is conservative but not pathological.

4 Discussion

Summary across these tables. As the model grows, linguistic and topical share grows. Input-agnostic share drops.

That is a cell-level rendering of “the bigger one specializes more.” Not a perplexity claim. Not a benchmark claim. A count of how many addressable sensors on a fixed lattice satisfy each of six pre-declared bit masks. Nothing more.

Backbone inversion is the most surprising figure here, because it cuts against most scaling narratives. A common intuition holds that bigger transformers have “richer” representations everywhere. By this measurement, that intuition is not exactly right: bigger transformers do have richer specialization, but they also carry a thinner always-on layer, and cells freed at the input-agnostic layer plausibly fund the new specialized ones; one would predict, under the substitution reading, that ablating universal cells in a smaller model hurts performance more broadly than ablating the same proportion in a larger model, since the smaller model has fewer specialized backup pathways. That prediction has not been tested here, and the data presented in this paper cannot distinguish the substitution reading from a simpler “larger models just allocate more cells everywhere except where saturation already exists” reading.

Suggestive. Not proven.

The matched ablation that would confirm causal substitution has not been run. Future work.

One more note on the cross-lingual physics result. Discovery [2] reported 50 such cells in 7B and treated that as evidence of tongue-neutral topic abstraction. The 3B tally of 24 strengthens that reading in one way: the abstraction does not appear out of nowhere at 7B. It weakens that reading in another way: a smaller cluster has less statistical room above the null at 3B. The relevant test is a wider physics prompt set in both directions across both sizes. Not yet done.

5 Limitations

Two sizes from one family is a small N. A serious cross-family study would need at least one other family (e.g. Llama 3.x or Gemma 2) at comparable parameter counts, run through the Mercury recipe. The method paper [1] notes the lattice procedure is architecture-aware and generalizes to other transformer families; the empirical test of that has not been done.

The strict bit-mask definitions are conservative. Many “almost-monolingual” or “almost-topic” signatures land outside the strict categories. Subset analysis suggests trends survive when definitions relax, though exact ratios will move.

Ten prompts is small. Obviously.

Three or four physics topics per side, rather than one each, would tell us whether the 50 (and 24) cross-lingual physics cells are a stable group or an artifact of which Chinese and English queries were chosen. I expect a stable group. I have not run the wider sweep.

Quantile-bucket counts differ across grids ($Q = 10$ for 7B, $Q = 5$ for 3B). Signature definitions are quantile-independent (they look at whether a sensor fires, not how strongly), so the comparison is well-defined. But finer-grained quantile binning at 7B can promote slightly weaker signals into the active set, which can inflate 7B counts uniformly. I do not think this drives the headline ratios, because they vary widely across categories (0.41 to 2.13), which a uniform inflation cannot produce.

A 14B observation pass is the obvious next datapoint. It fits a single consumer-grade GPU [1] and the bench script in the open dataset [3] accepts a third row of this table without code changes. Not yet run.

6 Conclusion

Two transformers from one family, observed on identical ten-prompt inputs through Mercury lattices scaled to each architecture, show internal category structure that differs in a directional way. Topical and linguistic categories carry more cells in the larger model. The always-on backbone carries fewer.

Cross-lingual physics is present at both sizes.

Both.

Five of six pre-declared categories grow with parameter count; one shrinks. This is a cell-level reading of internal specialization that can be compared across transformers, runs in minutes on a laptop, and accepts further model sizes without code changes.

Acknowledgements

Drafted in continuation of the Mercury sprint as a paired-engineering session with an Anthropic Claude Code agent. Every numeric claim traces to a git commit hash and an on-disk binary in the companion dataset [3]. The commits relevant to the 3B-vs-7B comparison are 8ee6256 (scaling extraction) and 3fa2c96 (category discovery on 7B); the 3B observation lives in fd43cd8 alongside the 1M-grid build.

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] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models. arXiv preprint, 2020. DOI: 10.48550/arXiv.2001.08361.
- [5] 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.
- [6] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent Abilities of Large Language Models. arXiv preprint, 2022. DOI: 10.48550/arXiv.2206.07682.