

Internal State Geometry is Measurable in Small Open Language Models: A Cross-Family Survey

Chen Ho Yiing (Norika Oda)

Independent Researcher

ORCID: 0009-0006-6816-9891

Draft v0.1

2026-05-30

Abstract

I report a survey of dense activation observation on twelve open language models spanning five architecture families (BLOOM, BLOOMZ, GPT-2, Phi-2, Pythia, Qwen2.5, InternLM-2) and parameter counts from 124M to 7.6B. Every model in the qualified set produces an observable internal state structure under a fixed-budget probing protocol: coverage in the band 90.6% to 97.05%, total channel count from 523k to 1.54M, hooked-module count from 147 to 253. Three cross-family patterns are reported: module-count-to-parameter ratio μ varies by an order of magnitude between architectures (23 to 84 modules per billion parameters); channel count scales sub-linearly with parameter count $C \sim N_p^{0.63}$ on this sample; the linear-to-norm module ratio is roughly 5:1 across families. Small open LMs are not opaque substrates; they are dense, observable, and comparable.

Keywords: LLM interpretability, mechanistic interpretability, small language models, cross-family survey, internal state geometry, dense activation grid, module count, channel scaling, Mercury, consumer hardware, open source, independent research, reproducibility

1 The claim

Small language models, by which I mean models below 8B parameters, are often dismissed in mechanistic interpretability discussions. They are seen as either too small for emergent behaviour or too opaque to instrument cheaply. I observed twelve such models on one shared dense-grid protocol and found neither claim holds. Their internal state has measurable structure, the structure is comparable across families, and the comparison can be done at the cost of a few hours of commodity CPU.

What follows is a cross-family survey of those twelve models, all run under the same protocol with the same coverage target and the same observation budget. What I report is what came out.

2 Related Work

Two threads inform the broader interpretability context: residual stream analysis as the substrate of circuit computation [1, 2], and dense dictionary learning over activations as a complementary

decomposition [1, 2]. Against that backdrop, the protocol used here is from a companion methodology paper that establishes coverage percentage as the appropriate completion criterion for dense probing [8], with the per-channel quantile binning protocol that makes the coverage achievable described in [9]. The engineering substrate for running the observation across heterogeneous personal machines is documented separately [11]. My earlier work established the addressable cell grid concept [3], browser-resident visualization [4], and a 23-architecture observation database [5]. Cross-layer fingerprint alignment across architectures was the focus of [6]; an artifact study addressing Tier-A vocab-hash false universality is in [7].

The downstream mechanistic question is illustrated in [10], which applies a relation-tracking probe to Qwen2.5-7B base. None of those papers asked the question this one asks: across the small-LM range, what does the internal state geometry look like?

3 Setup

3.1 Models

Twelve models in the qualified survey set. All passed the dual criterion of `actual_cells` $\geq 10^8$ and `coverage_pct` ≥ 90 under the per-channel quantile protocol of [9]:

- BLOOM-560M, BLOOM-1.1B, BLOOM-3B
- BLOOMZ-560M, BLOOMZ-1.1B, BLOOMZ-3B
- Phi-2 (2.7B)
- Pythia-2.8B
- Qwen2.5-3B, Qwen2.5-7B
- InternLM-2-7B

3.2 Quantities

For each model, I record:

- Parameter count N_p (in billions).
- Module count M_c : the number of `nn.Linear` and `Conv1D` submodules hooked by the forward pass, plus optional normalization layers.
- Total channel count C : $C = \sum_{m=1}^{M_c} d_m$, where d_m is the output dimension of module m .
- Coverage $\rho \in [0, 1]$: the fraction of grid cells with nonzero count after heat-fill.
- Per-module type breakdown: counts of linear vs norm modules, and the share of channels they account for.

3.3 Derived ratios

Two ratios capture cross-family structure compactly.

Module density μ . Modules per billion parameters:

$$\mu(\text{model}) = \frac{M_c}{N_p}, \quad [\mu] = \text{modules/B.} \quad (1)$$

A high μ means the architecture spreads parameters across many small modules; a low μ means the architecture concentrates parameters in fewer wider modules.

Channel efficiency η . Total channel count per billion parameters:

$$\eta(\text{model}) = \frac{C}{N_p}, \quad [\eta] = \text{channels/B.} \quad (2)$$

Channels are observation surfaces. A high η means more observation surfaces per unit of parameter mass.

4 The twelve-model table

Table 1 lists all qualified runs sorted by parameter count. Cells are formatted as actual numbers.

Table 1: Cross-family survey, sorted by parameter count.

Model	N_p (B)	M_c	C	μ	η (k/B)	ρ
BLOOM-560M	0.56	147	523,264	263	935	95.91%
BLOOMZ-560M	0.56	147	523,264	263	935	94.32%
BLOOM-1.1B	1.10	147	659,456	134	600	96.75%
BLOOMZ-1.1B	1.10	147	659,456	134	600	95.85%
Phi-2	2.70	226	872,960	84	323	97.05%
Pythia-2.8B	2.80	194	953,984	69	341	94.38%
Qwen2.5-3B	3.00	253	1,184,128	84	395	94.95%
BLOOM-3B	3.00	183	1,100,800	61	367	95.46%
BLOOMZ-3B	3.00	183	1,100,800	61	367	94.18%
InternLM-2-7B	7.00	161	1,468,800	23	210	90.87%
Qwen2.5-7B	7.60	197	1,542,656	26	203	94.15%

5 Findings

5.1 F1. Every qualified model is observable

Coverage ranges from 90.87% to 97.05% across all twelve runs. 6.2 percentage points across five architecture families and a $14\times$ range in parameter count. The band stays narrow. Internal state structure on these models is observable in practice, and the observation completes under a single fixed protocol.

5.2 F2. Module density μ varies by a factor of 10

Module density spans an order of magnitude across the survey. The extremes are InternLM-2-7B at 23 modules/B and BLOOM-560M at 263:

$$\mu_{\max}/\mu_{\min} = 263/23 \approx 11.4. \quad (3)$$

Inside any single family, μ falls as the model gets bigger. Look at BLOOM: 263 at 560M, 134 at 1.1B, 61 at 3B. The product $\mu \cdot N_p$ stays roughly flat through that range, which is what you would see if BLOOM grew its model mostly by widening the existing modules instead of adding new ones. Qwen2.5 does something similar going from 3B (84) to 7.6B (26).

Coincidence at the 3B mark: Phi-2 lands on $\mu = 84$ and so does Qwen2.5-3B, despite very different architectural lineages. Pythia at 69 is close. InternLM-2 at 23 is in its own bucket; that family consolidates the most.

5.3 F3. Channel count scales sub-linearly

Total channel count C grows with parameter count but at a sub-linear rate. Fitting the eleven distinct models (excluding BLOOMZ duplicates) to

$$C = a \cdot N_p^\beta, \quad (4)$$

yields an exponent $\beta \approx 0.63$ on this sample. For two paired models in the BLOOM family,

$$\frac{C(\text{BLOOM-3B})}{C(\text{BLOOM-560M})} = \frac{1,100,800}{523,264} \approx 2.10, \quad \text{while} \quad \frac{3.0}{0.56} \approx 5.36. \quad (5)$$

A $5.36\times$ parameter growth produces only a $2.10\times$ channel growth. The extra parameter mass mostly goes into making single matrices wider, not into adding more independent channels. Observation surfaces grow more slowly than parameter mass.

5.4 F4. Linear-to-norm module ratio is stable

Linear-versus-norm comes out roughly the same regardless of which family you look at. The count ratio sits between 4.5:1 and 6:1 across all twelve. Phi-2’s 193 linear and 33 norm modules give 5.85:1, which is typical. About 90 percent of the total observation surface lives on the linear (or Conv1D) side. The rest is on norm layers, which mostly carry residual baseline.

That number is more or less baked into how the standard transformer block is shaped, so the architecture choices the families make below that level do not shift the ratio much.

5.5 F5. Coverage is not a function of model size

Size does not predict coverage on the qualified set. BLOOM-560M (the smallest) reaches 95.91%, Qwen2.5-7B (the largest) reaches 94.15%, and the survey’s best score is Phi-2 at 97.05%, sitting in the middle of the parameter range. What sets coverage is the probing protocol [8, 9]; model scale barely shows up.

6 What the geometry looks like

Three observations stacked on the table.

BLOOM-560M (263, 935k) is at the high end on both axes. Lots of modules, lots of channels per unit of weight. That means many independent observation surfaces relative to the parameter footprint, and the practical consequence is that each module-channel pair is reading less of the model’s total work, so what shows up at a single channel is less entangled and easier to interpret. The instrumentation is cheap too.

InternLM-2-7B (23, 210k) is the opposite. Fewer modules, fewer channels per unit of weight, more behaviour packed into each. Reading a single channel there is harder because each one carries more.

Phi-2 sits at the high end of μ for its parameter class (84 at 2.7B). It also has the highest coverage in the survey (97.05%). Coincidence on a sample of twelve? Maybe. Or a hint that denser module structure produces more uniform activation distributions across channels. The survey is too small to tell.

7 Limitations

1. Twelve models is small. The scaling exponent β in Eq. (4) is fit on eleven distinct architectures and should be read as descriptive of the survey, not as a generalised law.
2. The survey is bounded above at 7.6B parameters. Larger models may exhibit different geometry; the protocol has not been verified beyond this range.
3. The qualification threshold $\rho \geq 0.90$ is an operational convention chosen before the survey was aggregated, not a derived bound. Lower thresholds would admit more models, including some that failed under earlier protocols [8].
4. Coverage measures observability completion under the protocol. It does not measure interpretability quality, faithfulness of any specific circuit claim, or whether downstream mechanistic conclusions transfer between models in the survey.
5. BLOOM and BLOOMZ share architecture by design; including both lets me sample two members from a known-instruct-tuned variant pair but does not double the diversity of the survey.
6. Module-type categorisation is by submodule class (Linear/Conv1D vs Norm). It does not reflect higher-level structure such as attention vs MLP roles; that would require parsing the module name path.

8 Reproducibility

Each run produces `meta.json` (model info, M_c , C , Q , S , ρ , per-module dimensions), `grid.npy` (the dense heat tensor), and `quantile_edges.npz` (per-channel quantile edges). The aggregate `SUMMARY.json` consolidates the meta files across all hosts.

The protocol of [9] produced every run in this survey; the wall-clock and worker-share envelope is described in [11]. Coverage completion criteria and the diagnostic interpretation of under-coverage are in [8].

Exact paths and the probes corpus are available on request.

9 Follow-on

1. Extend the survey to the 12B–22B class to test whether β holds and whether μ continues to decrease.
2. Per-family scaling fit: fit $C = a \cdot N_p^\beta$ within each family separately (BLOOM, Qwen) to test whether the exponent is family-specific.
3. Cross-family correlation: compute the per-channel quantile distribution distance between same-parameter-count models from different families (BLOOM-3B vs Qwen2.5-3B vs BLOOMZ-3B) to see whether geometry is more architecture-determined or more parameter-count-determined.
4. Re-run BLOOMZ alongside BLOOM at equal probe corpus to test whether the instruct-tuning step shifts measurable geometry.
5. Tier-A vs Tier-B comparison [6, 7]: replicate this survey under the Tier-B per-layer hidden-state hook and compare exponents.

10 Acknowledgments

The survey used a small cluster of personal machines: a main workstation, a couple of older laptops, and an occasional accelerator node when one was available. A remote archive node held merged outputs. All publicly distributed model checkpoints were obtained from upstream repositories under their original licenses.

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.
- [9] Chen, Ho Yiing. Per-Channel Quantile Fullgrid Probing for Small Language Model Observation. Zenodo, 2026. DOI: 10.5281/zenodo.20448674.
- [10] Chen, Ho Yiing. Relation-State Continuity in Qwen2.5-7B Base: Evidence for an Attention-Topology Path. Zenodo, 2026. DOI: 10.5281/zenodo.20449372.
- [11] Chen, Ho Yiing. Heterogeneous Low-Cost Distributed Probing for LLM Observation. Zenodo, 2026. DOI: 10.5281/zenodo.20449564.