

Coverage Failure in Sparse LLM Observation and Per-Channel Quantile Fullgrid Recovery

Chen, Ho Yiing (norika)¹

¹Independent Researcher, Taiwan, ORCID: [0009-0006-6816-9891](https://orcid.org/0009-0006-6816-9891), [norika at charenix.com](mailto:norika@charenix.com), Zenodo DOI: [10.5281/zenodo.20727860](https://doi.org/10.5281/zenodo.20727860)

June 17, 2026

Abstract

Some LLM observation runs look large enough to trust and are still missing too much of the grid. I report a 19-run Mercury study over GPT-2, Gemma, Qwen, BLOOM/BLOOMZ, Pythia, Phi-2, and InternLM-family systems. The release rule is plain: at least 100 million allocated entries and at least 90% observed entries. Older sparse or coarse settings clear that line in only 1 of 7 trials. Several allocate more than 100 million entries but touch only 58.75–89.77%. A dense per-channel quantile (PCQ) protocol clears the same bar in 12 of 12 trials, with mean touched share 94.90%. This is not a semantic claim. It says something more basic: do not interpret a map until the map itself has been measured.

1 Introduction

Interpretability work often starts too late. Hook libraries make model states available [1]; attention viewers and activation atlases make some of those states visible [2, 3]; dictionary-learning work asks whether better units than individual neurons can be found [4]. Those are useful directions. I want to stop one step earlier: did we actually measure the grid we are talking about?

Here *fullgrid* (FG) means an explicitly indexed activation lattice over modules and channels and bins. Each trial records how many entries were materialized and how many were touched at least once. I use a deliberately blunt release line:

$$\text{complete} \iff \text{actual_cells} \geq 100,000,000 \quad \wedge \quad \text{coverage_pct} \geq 90.$$

This is not a law of nature. It is a stop sign. If a run does not cross it, I can still learn from the failure, but I should not sell it as a completed FG result.

The key failure is simple and nasty. A trial can allocate more than 100 million entries and still touch only 58–89% of them. More probes help, but not always enough. PCQ plus dense probing crosses the line across twelve observed trials.

What is new here.

1. A release rule for FG work: actual units plus observed share.
2. A negative result: older sparse/coarse settings clear the bar in only 1 of 7 cases.
3. A practical fix: PCQ dense trials clear it in 12 of 12 cases in this archive.
4. A boundary: this validates measurement reach, not semantic truth.

2 Problem

LLM internals are not polite: one direction may fire all the time, another may barely move, another may hide in a fat tail, and another may stay quiet until the prompt enters the right state. A global threshold makes that mess look cleaner than it is. It can over-read loud directions and under-read quiet ones.

Mercury makes the failure visible by forcing every unit to have an address. The lattice is:

$$\mathcal{G} = \{(m, c, q) : m \in \mathcal{M}, c \in C_m, q \in Q_c\},$$

where m is a module, c is a channel, and q is a local bin. The important move is that bins are calibrated per channel. The grid is not allowed to pretend that every coordinate has the same scale.

3 Method

3.1 Completion Gate

For each trial, Mercury records:

- `actual_cells`: number of cells materialized in the observation grid;
- `nonzero_cells`: number of cells observed as nonzero at least once;
- `coverage_pct`: nonzero coverage percentage;
- `probe_count`: number of probes applied;
- `per_channel_quantiles`: whether calibration thresholds are channel-local.

For trial i , let A_i be `actual_cells` and Z_i be `nonzero_cells`. The observed share is

$$C_i = 100 \cdot \frac{Z_i}{A_i}.$$

This is the reported `coverage_pct`. It is deliberately plain. If a entry was never touched, it should not quietly count as measured.

The completion gate is:

$$G_i = \mathbf{1}[A_i \geq 10^8] \cdot \mathbf{1}[C_i \geq 90].$$

Runs below this line remain useful as diagnostics. They do not get to carry a completed-map conclusion.

3.2 PCQ Rule

For a channel c , let F_c be its empirical activation CDF on a calibration probe set. A quantile threshold for bin q is:

$$\tau_{c,q} = F_c^{-1}(q).$$

A entry (m, c, q) fires when the observed activation $a_{m,c}$ crosses its channel-local threshold:

$$\mathbf{1}[a_{m,c} \geq \tau_{c,q}].$$

The reason is not aesthetic. It is damage control. A global cutoff can reward large-scale channels and erase small-scale ones. A channel-wise cutoff gives each coordinate a fair chance to be seen before anyone argues about meaning.

3.3 Dataset of Runs

The analysis uses 19 synchronized Mercury manifests. Host machines do not matter for the claim; only the saved fields do. Families include GPT-2, Gemma, Qwen, BLOOM/BLOOMZ, Pythia, Phi-2, and InternLM2.

For any set of trials S , I compute:

$$\bar{C}(S) = \frac{1}{|S|} \sum_{i \in S} C_i, \quad R(S) = \frac{1}{|S|} \sum_{i \in S} G_i.$$

Thus Table 1 can be recalculated directly from the CSV summary without trusting my prose.

4 Results

4.1 Aggregate Result

Table 1: Touched-cell share by protocol group.

Protocol group	Runs	Mean touched share	Clears line	Rate
Older sparse/coarse or non-PCQ trials	7	79.89%	1	14.29%
PCQ dense trials	12	94.90%	12	100.00%

The older group contains multiple 100-million-entry runs that still miss the 90% bar. This group clears it every time in this archive.

The measured gap is:

$$\bar{C}(S_{\text{PCQ}}) - \bar{C}(S_{\text{old}}) = 94.90 - 79.89 = 15.01$$

percentage points. The gate-rate gap is:

$$R(S_{\text{PCQ}}) - R(S_{\text{old}}) = \frac{12}{12} - \frac{1}{7} = 0.8571.$$

So the claim is not that the picture “looks better.” The claim is that the manifest-level pass rate changes by a computable amount.

4.2 Manifest-Level Results

Table 2: Run manifests used in this study. Names are shortened for readability.

Trial	Actual entries	Touched	Probes	Line
Phi-2 PCQ	105,628,160	97.05%	1286	yes
BLOOM 1.1B PCQ	102,875,136	96.75%	1286	yes
BLOOM 560M PCQ	102,559,744	95.91%	1286	yes
BLOOMZ 1.1B PCQ	102,875,136	95.85%	1286	yes
BLOOM 3B PCQ	110,080,000	95.46%	1286	yes
Qwen2.5 3B PCQ	118,412,800	94.95%	1286	yes
Qwen 3B PCQ	118,412,800	94.88%	1286	yes
Pythia 2.8B PCQ	104,938,240	94.38%	1286	yes
BLOOMZ 560M PCQ	102,559,744	94.32%	1286	yes

Trial	Actual entries	Touched	Probes	Line
BLOOMZ 3B PCQ	110,080,000	94.18%	1286	yes
Qwen 7B PCQ	154,265,600	94.15%	1286	yes
InternLM2 7B PCQ	146,880,000	90.87%	1286	yes
Qwen 3B q6s15	106,571,520	90.64%	166	yes
GPT-2 medium p180	104,203,908	89.77%	166	no
Qwen 3B p180	118,412,800	87.08%	166	no
GPT-2 p180	103,023,076	85.69%	166	no
Gemma 2B old	105,768,960	76.86%	30	no
GPT-2 medium old	104,203,908	70.45%	30	no
GPT-2 old	103,023,076	58.75%	30	no

4.3 Probe Count Alone Is Not Enough

The GPT-2 and GPT-2-medium pairs show that simply raising probe count improves the touched share but does not necessarily produce a completed map. GPT-2 rises from 58.75% to 85.69% when moving from 30 to 166 probes. GPT-2-medium rises from 70.45% to 89.77%. Both remain below the 90% bar. This matters because the failure mode would be easy to miss if a report only stated that more than 100 million entries were allocated.

For those two direct pairs:

$$\Delta C_{\text{GPT2}} = 85.69 - 58.75 = 26.94, \quad \Delta C_{\text{GPT2m}} = 89.77 - 70.45 = 19.32.$$

Both still have $G_i = 0$ because $C_i < 90$. More probes helped. They did not make the result releasable.

4.4 PCQ Is Not Merely More Probes

This group uses a larger probe count (1286) and channel-wise calibration. This paper does not claim to isolate those two factors completely; the observed intervention is the combined protocol. The practical conclusion is still useful: dense probing with channel-wise bins gives an auditable map here, while the older variants often do not. Future ablations should separately vary probe count, bin count, calibration corpus, and channel-wise versus global thresholds.

The ablation target is:

$$C = f(P, Q, B, T),$$

where P is probe count, Q is bin count, B is the calibration corpus, and T is threshold type. This paper measures a combined move in (P, T) , not a single-coordinate causal effect.

5 Interpretation Boundaries

The main risk is overclaiming. An observed entry is not automatically a concept, feature, neuron, circuit, or causal mechanism. The release rule only says that the declared lattice was reached. It is a prerequisite for interpretation, not an interpretation by itself.

Likewise, a trial below the bar is not useless. Such runs are useful as negative controls and method comparisons. They should not, however, be used as the basis for claims such as “this model lacks structure in region X” unless the missing area is explicitly accounted for.

6 Implications

For interpretability tooling, the result suggests a simple reporting norm: include actual grid size, observed share, probe count, and calibration method with every activation-map result. Without these four fields, a reader cannot tell whether an apparent structure is a model property or a measurement artifact.

For local analysis, the result also matters operationally. PCQ dense trials are slower than sparse probes, but they prevent expensive false conclusions. The cost of a long run is lower than the cost of building a paper or demo or surgery plan on a partial map.

7 Limitations

This is an empirical methods note, not a universal theorem. The archive is heterogeneous: machines and families; wall times also differ. This group also uses more probes than the older group, so the current data does not cleanly estimate the independent contribution of channel-wise calibration alone. The 100-million-entry and 90% bar is a Mercury release criterion, not a universal law.

The study also does not evaluate downstream semantic interpretability quality. That requires separate validation: causal interventions, held-out prompts, human annotation, or agreement with independently trained feature dictionaries.

8 Conclusion

The first failure mode may not be semantic. It may be geometric. Sparse or coarse observation can allocate a large lattice and still leave too much of it untouched. In the Mercury archive studied here, older settings clear the 100-million-entry/90% bar in only 1 of 7 trials. PCQ dense trials clear it in 12 of 12. Before interpreting an activation map, report whether the map was actually measured.

Data and Reproducibility

The run manifests used for the tables are stored in the Mercury observation archive under:

```
observations/2026-05-29-fullgrid-sync/
```

and summarized in:

```
var/mercury-fullgrid-plot-index/20260605_145727/runs.csv
```

The paper package includes machine-readable metadata for Zenodo and citation.

Acknowledgements

This work was carried out as independent research using personally maintained local machines and opportunistic off-hours compute access. Draft preparation and code editing were assisted by AI coding tools under the author’s direction.

References

- [1] Nanda, Neel and Bloom, Joseph. TransformerLens. 2022. <https://transformerlensorg.github.io/TransformerLens/content/citation.html>.
- [2] Vig, Jesse. A Multiscale Visualization of Attention in the Transformer Model. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019. <https://aclanthology.org/P19-3007/>.
- [3] Carter, Shan; Armstrong, Zan; Schubert, Ludwig; Johnson, Ian; and Olah, Chris. Exploring Neural Networks with Activation Atlases. Distill, 2019. doi: [10.23915/distill.00015](https://doi.org/10.23915/distill.00015). <https://distill.pub/2019/activation-atlas/>.

- [4] Bricken, Trenton; Templeton, Adly; Batson, Joshua; Chen, Brian; Jermyn, Adam; Conerly, Tom; Turner, Nick; Anil, Cem; Denison, Carson; Askeell, Amanda; Lasenby, Robert; Wu, Yifan; Kravec, Shauna; Schiefer, Nicholas; Maxwell, Tim; Joseph, Nicholas; Hatfield-Dodds, Zac; Tamkin, Alex; Nguyen, Karina; McLean, Brayden; Burke, Josiah E.; Hume, Tristan; Carter, Shan; Henighan, Tom; and Olah, Chris. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Transformer Circuits Thread, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/>.