

Coverage Audits for LLM Interpretability Claims

Chen, Ho Yiing (norika)¹

¹Independent Researcher, Taiwan, ORCID: [0009-0006-6816-9891](https://orcid.org/0009-0006-6816-9891), norika at charenix.com, Zenodo DOI: [10.5281/zenodo.20729124](https://doi.org/10.5281/zenodo.20729124)

June 17, 2026

Abstract

Mechanistic interpretability often asks whether a neuron, subspace, layer, or patch explains a behavior. I argue for an earlier gate: before such a claim is trusted, the observation grid itself should pass a coverage audit. Mercury’s current public atlas contains 38 opened language-model towers, but this paper does not pretend that every tower has the same evidence status. It audits the 19 saved fullgrid manifests that carry enough cell-level metadata for a strict coverage gate, spanning GPT-2, Gemma, Qwen, BLOOM/BLOOMZ, Pythia, Phi-2, and InternLM-family runs. In that auditable subset, older sparse or coarse protocols cleared a simple 100M-cell and 90%-touched audit in only 1 of 7 cases, while per-channel quantile dense runs cleared it in 12 of 12 cases. The point is not that coverage proves semantics. It does not. The point is sharper: sparse probing, activation patching, layer surgery, and pruning experiments can look decisive while being built on an unmeasured map. I propose a small audit object: cell count, touched share, channel balance, and protocol trace. Interpretability claims should report this object before claiming localization.

1 The Claim Before the Claim

The most tempting sentence in interpretability is: this component matters. Maybe it is a neuron, a head, a residual stream slice, a subspace, or a layer. That sentence may be true. It may also be a clean story drawn from a patchy map.

Sparse probing can find human-readable features [1]. Activation patching can localize causal pathways [2]. Subspace patching can also create an interpretability illusion when intervention and explanation come apart [3]. Layer-pruning work can show that some blocks are less needed than expected [4, 5]. Tooling and visualization systems make this work practical [7, 6]. None of these directions is the enemy. The risk is simpler: the method may start arguing about meaning before the observation surface has been measured.

This paper proposes a precondition:

$$\text{interpretability claim} \Rightarrow \text{coverage audit.}$$

The audit does not prove the explanation. It only says the measurement map is large enough and evenly enough touched that a localization claim is not built on obvious missing cells.

2 What Mercury Measures

Mercury fullgrid records a lattice of observed model coordinates. A cell is:

$$x = (m, c, b),$$

where m is a module or layer address, c is a channel, and b is a local bin or quantile bucket. A run materializes a set of cells:

$$\mathcal{G} = \{x_1, \dots, x_N\}.$$

During probing, a subset is touched:

$$\mathcal{T} \subseteq \mathcal{G}.$$

The first audit number is:

$$C = 100 \cdot \frac{|\mathcal{T}|}{|\mathcal{G}|}.$$

The Mercury release line used here is:

$$G = \mathbf{1}[|\mathcal{G}| \geq 10^8] \cdot \mathbf{1}[C \geq 90].$$

This gate is deliberately blunt. It does not ask whether a cell is important. It asks whether the map is present.

3 Why Channel-Local Bins Matter

Global thresholds are easy. They are also unfair. Let a_c be activations for channel c . If two channels have different scale or tail shape, a single threshold τ can make one channel look alive and another look empty. That is a measurement artifact.

For per-channel quantile fullgrid, each channel gets its own empirical CDF:

$$F_c(t) = P(a_c \leq t).$$

For a quantile level q , the threshold is:

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

A bin fires when:

$$z_{c,q} = \mathbf{1}[a_c \geq \tau_{c,q}].$$

The rule is not magic. It just prevents a loud channel from defining the ruler for a quiet one. That is the right default for a coverage audit.

4 Observed Failure

The broader Mercury corpus contains 38 model atlases. This audit intentionally uses the 19 saved Mercury fullgrid manifests under:

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

The other Mercury towers still matter for visualization, cross-model cartography, and qualitative atlas work. They are not counted in this table unless the corresponding run leaves the audit object needed here. The model families in the auditable subset include GPT-2, Gemma, Qwen, BLOOM/BLOOMZ, Pythia, Phi-2, and InternLM2. The gate is the same for every row: at least 10^8 cells and at least 90% touched.

Table 1: Coverage audit by protocol group.

Protocol group	Runs	Mean touched share	Passes	Pass rate
Older sparse/coarse or non-PCQ	7	79.89%	1	14.29%
Per-channel quantile dense	12	94.90%	12	100.00%

The gap is:

$$\Delta_C = 94.90 - 79.89 = 15.01$$

percentage points. The pass-rate gap is:

$$\Delta_G = \frac{12}{12} - \frac{1}{7} = 0.8571.$$

This is not a tiny formatting difference. The protocol changes whether the measurement earns the right to be treated as a completed map.

5 Manifest Rows

Table 2: The 19 auditable fullgrid manifests used in this coverage audit. They are a strict evidence subset of the 38-model Mercury atlas, not the size of the whole corpus. Names are shortened.

Run	Cells	Touched	Probes	Gate
Gemma 2B old	105,768,960	76.864%	30	fail
GPT-2 old	103,023,076	58.752%	30	fail
GPT-2 p180	103,023,076	85.686%	166	fail
GPT-2 medium old	104,203,908	70.452%	30	fail
GPT-2 medium p180	104,203,908	89.767%	166	fail
Qwen 3B p180	118,412,800	87.078%	166	fail
Qwen 3B q6s15 p180	106,571,520	90.639%	166	pass
Qwen2.5 3B PCQ	118,412,800	94.950%	1286	pass
Qwen 7B PCQ	154,265,600	94.150%	1286	pass
BLOOM 1.1B PCQ	102,875,136	96.750%	1286	pass
BLOOM 3B PCQ	110,080,000	95.462%	1286	pass
BLOOM 560M PCQ	102,559,744	95.912%	1286	pass
BLOOMZ 1.1B PCQ	102,875,136	95.846%	1286	pass
BLOOMZ 3B PCQ	110,080,000	94.180%	1286	pass
BLOOMZ 560M PCQ	102,559,744	94.318%	1286	pass
InternLM2 7B PCQ	146,880,000	90.875%	1286	pass
Phi-2 PCQ	105,628,160	97.046%	1286	pass
Pythia 2.8B PCQ	104,938,240	94.384%	1286	pass
Qwen 3B PCQ	118,412,800	94.878%	1286	pass

Two rows are especially useful. GPT-2 medium p180 touches 89.767% of its grid. That looks close. It still fails. Qwen 3B q6s15 p180 touches 90.639%, so it passes by the same rule. A gate should be allowed to be boring. That is how it prevents story-driven exceptions.

6 Coverage Audit Object

I propose reporting an audit object with every interpretability claim:

$$\mathcal{A} = (N, C, P, B, R, H).$$

Here:

- $N = |\mathcal{G}|$ is materialized cell count;
- C is touched share;
- P is probe count;
- B records whether bins are global, per-layer, or per-channel;

- R records route: sparse, sampled, dense, PCQ, or mixed;
- H is a hash or manifest path for reproduction.

The audit predicate is:

$$\text{Pass}(\mathcal{A}) = \mathbf{1}[N \geq N_0] \mathbf{1}[C \geq C_0] \mathbf{1}[B = \text{PCQ}]$$

for claims that depend on broad channel coverage. A different task may choose a different N_0 or C_0 . The important part is not my threshold. It is that the threshold is declared before the explanation.

7 Implications for Existing Methods

For sparse probing, the audit asks whether the search surface was broad enough before declaring that the found neurons are the relevant ones. The sparse probe may still be right. The audit prevents the weaker sentence: these are the neurons, because these are the only ones we found.

For activation patching, the audit asks whether the patched addresses live in a well-covered map. Zhang and Nanda show that metrics and corruption choices can change localization results [2]. A coverage audit adds another check: were enough candidate cells observed before the localization story was formed?

For subspace interventions, Makelov et al. show that a patch can move behavior without faithfully identifying the causal feature [3]. Coverage does not solve that faithfulness problem. It only removes a more basic failure: the claimed subspace may be drawn from a poorly touched grid.

For layer pruning, ShortGPT shows that layers can be redundant in some settings and fragile in others [4]. A Mercury audit would not replace perplexity or task tests. It would sit before them and ask whether the layer decision came from a map that was actually observed.

8 What This Does Not Prove

Coverage is not meaning. A 95%-touched grid can still support a bad interpretation. A PCQ run can still miss a rare behavior. A manifest can be large and still have weak prompts. The audit is not a truth machine.

The stronger claim is negative:

$$\neg \text{Pass}(\mathcal{A}) \Rightarrow \text{do not present broad localization as measured.}$$

That is enough. Many bad claims should die before they become diagrams.

9 Conclusion

The finding is not that Mercury sees everything. It does not. The finding is that many observation protocols do not even prove they saw enough. The full Mercury map is larger: 38 opened language-model atlases. The coverage result here is narrower on purpose. In the 19-run auditable fullgrid archive, older sparse or coarse protocols pass a basic audit in 1 of 7 cases. PCQ dense protocols pass in 12 of 12. That difference is large enough to change how interpretability work should report itself.

Before asking what a model means, ask whether the map was touched.

Data and Artifacts

The manifests used for this note are stored under:

```
observations/2026-05-29-fullgrid-sync/ai/fullgrid/*/meta.json
observations/2026-05-29-fullgrid-sync/gx10/fullgrid/*/meta.json
observations/2026-05-29-fullgrid-sync/john/fullgrid/*/meta.json
```

Acknowledgements

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

References

- [1] Gurnee, Wes; Nanda, Neel; Pauly, Matthew; Harvey, Katherine; Troitskii, Dmitrii; and Bertsimas, Dimitris. Finding Neurons in a Haystack: Case Studies with Sparse Probing. arXiv preprint arXiv:2305.01610, 2023. <https://arxiv.org/abs/2305.01610>.
- [2] Zhang, Fred; and Nanda, Neel. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. ICLR, 2024. <https://arxiv.org/abs/2309.16042>.
- [3] Makelov, Aleksandar; Lange, Georg; Geiger, Atticus; and Nanda, Neel. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. ICLR, 2024. <https://arxiv.org/abs/2311.17030>.
- [4] Men, Xin; Xu, Mingyu; Zhang, Qingyu; Wang, Bingning; Lin, Hongyu; Lu, Yaojie; Han, Xianpei; and Chen, Weipeng. ShortGPT: Layers in Large Language Models are More Redundant Than You Expect. Findings of ACL, 2025. <https://arxiv.org/abs/2403.03853>.
- [5] Ashkboos, Saleh; Croci, Maximilian; Nascimento, Marcelo Gennari do; Hoefler, Torsten; and Hensman, James. SliceGPT: Compress Large Language Models by Deleting Rows and Columns. ICLR, 2024. <https://arxiv.org/abs/2401.15024>.
- [6] Carter, Shan; Armstrong, Zan; Schubert, Ludwig; Johnson, Ian; and Olah, Chris. Activation Atlas. Distill, 2019. <https://distill.pub/2019/activation-atlas/>.
- [7] Nanda, Neel; and Bloom, Joseph. TransformerLens. Software repository, 2022. <https://github.com/TransformerLensOrg/TransformerLens>.