

Coverage Failure as a Diagnostic Signal in Neural Model Probing

Chen Ho Yiing (Norika Oda)
Independent Researcher
ORCID: 0009-0006-6816-9891

Draft v0.3
2026-05-29 (internal)

Abstract

We report a methodological observation from a multi-machine dense-grid activation probing study of nineteen small to mid-sized language models (124M to 7.6B parameters). Every experiment satisfied a nominal threshold of one hundred million observation slots, yet observed fill rate (fraction of those slots with nonzero counts) varied from 58.75 percent to 97.05 percent across experiments. The observed gap is strongly associated with protocol choices, in particular the prompt corpus size and the use of per-channel quantile binning. We cannot yet rule out interaction with the network under study, the input set, hook discovery completeness, or quantile edge quality; the comparisons we report are observational rather than controlled. We argue that total slot count alone is an invalid completion criterion for activation probing, that fill percentage should be reported as a first-class result, and that under-fill should be treated as a diagnostic signal about the protocol before being attributed to the system.

Keywords: model probing, mechanistic interpretability, activation observation, methodology, reproducibility

1 Introduction

A dense activation probe records, for each (module, output dimension) pair in a language network, how often the per-dimension activation across a corpus of input prompts falls into one of Q quantile buckets at each of S evenly spaced sequence positions. The resulting four-dimensional tensor of counts (modules $\times$ channels $\times$ quantiles $\times$ sequence positions) is the heat tensor.

In our study we used a target of one hundred million observation bins as a practical operating scale: it is large enough to expose per-channel structure in networks up to a few billion parameters, and small enough to fit in commodity RAM and complete in commodity wall time. Studies that report tensors of this kind of size typically report the total bin count as a measure of scale.

A natural question follows. If two researchers both report ~ 100 M bins, can the results be compared? Or more practically: if a researcher has produced a 100M-bin tensor for one network, is the observation complete?

The answer in our study is no. Total bin count is a property of the tensor shape; it says nothing about how many bins actually received any signal. A tensor in which only 58 percent of entries were ever incremented contains qualitatively less information than one in which 97 percent were incremented, even when both are nominally the same size.

We therefore propose that *coverage percentage*, defined as the fraction of entries with strictly positive count after the heat fill is complete, be treated as the primary completion signal for dense activation probing. Total bins alone is necessary but not sufficient.

2 Setup

2.1 Models

Nineteen single-model runs across the following families:

- BLOOM (560M, 1.1B, 3B) and BLOOMZ (560M, 1.1B, 3B)
- GPT-2 (124M) and GPT-2-medium (355M)
- Gemma-2 (2B instruct)
- Phi-2 (2.7B)
- Pythia (2.8B)
- Qwen2.5 (3B and 7B)
- InternLM-2 (7B)

Models were chosen for parameter range, family diversity, and the availability of open weights compatible with the HuggingFace transformers loader.

2.2 Probe protocol

For each network we hook every `nn.Linear` and (where applicable) `Conv1D` submodule that participates in the forward pass, plus optional normalization layers when the `include_norms` flag is set. For each input prompt we perform one forward pass, sample S evenly spaced token positions, and for each hooked module record per-output-dimension activation values. Activations are binned into Q quantiles. Quantile edges are computed in a separate calibration pass over the identical prompt corpus, then re-used across all heat-fill passes for that network.

The heat tensor is dense and uses `uint32` counts so that shards from multiple workers can be merged additively.

2.3 Methodology axes that varied across runs

Three axes varied across the nineteen runs:

1. Prompt corpus size P . Three settings appeared in our experiments: a default of 30 prompts, an expanded 180 prompts, and a final 1200 prompts.
2. Per-channel quantile binning. Initial experiments used a single set of quantile edges per module (pooled across channels). Later experiments computed per-channel edges, that is one quantile edge vector per (module, output channel) pair.
3. Budget allocation. With Q and S fixed by the target, larger P does not increase `actual_cells` but does increase the count of heat-fill operations per bin.

All other settings (`dtype = bfloat16`, `target_cells = 1e8`, `include_norms = True`, calibration corpus identical to heat corpus) were held constant across the experiments we report.

2.4 Completion criteria

Two thresholds were applied:

- `actual_cells` must reach or exceed `target_cells` (1e8).

- `coverage_pct` (fraction of bins with strictly positive count) must reach or exceed 90 percent.

Runs meeting both are reported as *qualified*. Those meeting only the first are reported as *not qualified*.

3 Results

Table 1 shows the nineteen runs sorted by `coverage_pct` ascending.

Table 1: Nineteen single-model runs sorted by coverage ascending.

Run	Cells	Coverage	Per-channel	Probes	Qual.
gpt2	103,023,076	58.75%	no	30	no
gpt2-medium	104,203,908	70.45%	no	30	no
gemma-2-2b-it	105,768,960	76.86%	no	30	no
gpt2	103,023,076	85.69%	no	180	no
qwen-3b	118,412,800	87.08%	no	180	no
gpt2-medium	104,203,908	89.77%	no	180	no
qwen-3b (Q=6,S=15)	106,571,520	90.64%	no	180	yes
internlm2-7b	146,880,000	90.87%	yes	1200	yes
qwen2.5-7b	154,265,600	94.15%	yes	1200	yes
bloomz-3b	110,080,000	94.18%	yes	1200	yes
bloomz-560m	102,559,744	94.32%	yes	1200	yes
pythia-2.8b	104,938,240	94.38%	yes	1200	yes
qwen-3b	118,412,800	94.88%	yes	1200	yes
qwen2.5-3b	118,412,800	94.95%	yes	1200	yes
bloom-3b	110,080,000	95.46%	yes	1200	yes
bloomz-1b1	102,875,136	95.85%	yes	1200	yes
bloom-560m	102,559,744	95.91%	yes	1200	yes
bloom-1b1	102,875,136	96.75%	yes	1200	yes
phi-2	105,628,160	97.05%	yes	1200	yes

Three regimes are visible:

- **Regime A** ($P = 30$, pooled quantiles): three runs, coverage 58.75% to 76.86%. All not qualified.
- **Regime B** ($P = 180$, pooled quantiles): three runs, coverage 85.69% to 89.77%. All not qualified despite expanded probe corpus.
- **Regime C** ($P = 1200$ or pooled with $Q = 6, S = 15$, with per-channel quantiles in most cases): thirteen runs, coverage 90.64% to 97.05%. All qualified.

Holding the network fixed, GPT-2 (124M) traverses two regimes: 58.75% under regime A, 85.69% under regime B. Neither reaches the 90% qualification threshold. GPT-2-medium (355M) goes from 70.45% (A) to 89.77% (B). Qwen-3B goes from 87.08% (B) to 94.88% (C).

Counts in column 2 vary modestly (within roughly 50 percent of each other) because Q and S are recalculated per network to fit the budget. Size therefore does not predict coverage: one

entry with 103M bins (`gpt2` under regime A) sits at 58.75% coverage, while another with 102M bins (`bloom-1b1` under regime C) reaches 96.75%. Tensor dimensions are nominally identical; the protocols are different.

4 Discussion

4.1 Why cell count alone is invalid

The heat tensor is parameterized by four axes (modules $\times$ channels $\times$ quantiles $\times$ sequence positions). Its total size is the product of the four dimensions. That is a property of the container, not of the data inside the container.

Any experiment can satisfy the size target by enlarging Q or S without actually filling those bins. Under regime A the calibration corpus was small (30 prompts), so the per-module per-channel quantile edges were computed from a sample distribution that did not span the realistic activation range. The heat-fill pass then concentrated counts in only a few quantile buckets, leaving most buckets empty. The tensor was 100M+ bins in size, but most entries were dead.

This pattern can be present in any dense probing study where:

1. The calibration corpus is small relative to the tensor size.
2. Quantile edges are pooled across many channels with different scales.
3. The prompt corpus does not exercise the network across the activation distributions seen at inference time.

4.2 Coverage as the right completion axis

`coverage_pct` is more directly informative about whether the probe is observing the network. An entry that is never incremented contributes zero information regardless of how many entries exist. Coverage reports the fraction of entries contributing.

A useful intuition: total bin count tells you how many bins you built. Coverage tells you how many bins you actually hit.

If two experiments on one network report different coverage under an identical budget, those experiments are not comparable as observations even when tensor dimensions match. The lower-coverage experiment observes a smaller effective subspace.

4.3 Under-coverage as diagnostic signal

We propose treating under-coverage as a diagnostic signal about the protocol before attributing it to the network. Across our data, every observed gap was reduced or closed when the protocol changed (more prompts, per-channel quantiles, smaller Q). Because these comparisons are across-experiment rather than controlled, we cannot prove the network was held strictly constant in its observed behaviour; we can only show that, where two protocols ran on a fixed network, the protocol change accounted for a large fraction of the difference.

This has a practical consequence. If a probing study reports low coverage and concludes that the network has unusual structure (sparse channels, narrow activation range, dead channels), that conclusion is not safe until the protocol has been ruled out by a controlled rerun. The signal could be the protocol.

4.4 What changed coverage in our runs

Within our nineteen-experiment sample, the strongest predictor of coverage was per-channel quantile binning combined with $P = 1200$. Every experiment in regime C exceeded 90 percent. Every experiment in regimes A and B fell short except for one $P = 180$ trial on Qwen-3B at $Q = 6, S = 15$ that just passed.

We do not separate the contribution of P from the contribution of per-channel binning in this draft. Both Qwen-family work on the GPU node and BLOOM / Phi-2 / Pythia work on the Apple Silicon nodes were configured together with identical P and per-channel settings, so the two protocol axes co-vary in our sample. An ablation isolating them is left to follow-on work.

5 Recommendations

We recommend that activation probing studies report:

1. `actual_cells` and `target_cells`, as before.
2. `coverage_pct = nonzero_cells / actual_cells`.
3. The calibration corpus size and the heat-fill corpus size (separately).
4. Whether quantile edges are per-channel or pooled.
5. A pre-registered qualification threshold (e.g., `coverage_pct` ≥ 90).
6. The exact per-module Q and S used.

Studies that report only cell counts without coverage cannot be compared cleanly across runs, and their reproducibility is weakened.

6 Limitations

1. We do not control for hardware variability across runs. Some runs were performed on x86 CPU hosts, some on consumer Apple Silicon nodes, and one model (Qwen-7B) was observed on an NVIDIA ARM-GPU workstation. Determinism settings (torch deterministic algorithms, math kernel) were not pinned.
2. $P = 180$ and $P = 1200$ use different probe text. We did not test the same $P = 1200$ probe set under pooled quantiles to isolate the per-channel factor.
3. We do not yet have qualified runs above 8B parameters in this sample. The coverage behavior at 12B+ may differ.
4. The qualification threshold of 90 percent is a working operational convention chosen before the runs were aggregated. We do not claim it is optimal; we claim only that it is a more honest completion signal than cell count alone. A sensitivity sweep across alternative thresholds (e.g., 80, 85, 90, 95) is listed under follow-on work.
5. Calibration corpus and heat-fill corpus share the P probes within each run. This is by design (the quantile edges should reflect the activation distribution that is later binned), but it does mean coverage reported here is in-sample with respect to the corpus. Out-of-sample coverage on a held-out probe set is not measured.
6. The comparisons we report are observational, not controlled. The 2x2 ablation (pooled vs per-channel, $P = 180$ vs $P = 1200$) holding model and all other axes constant is the experiment that would settle the per-channel-vs-corpus-size question and is also listed under follow-on work.
7. `coverage_pct` measures whether grid cells received any count. It is an observability com-

pletion metric. It does not measure interpretability quality, faithfulness of any downstream circuit claim, or whether the activation distribution was well sampled within each occupied cell.

7 Reproducibility

All artifacts referenced in this paper are organized as follows on the research host filesystem:

- Fullgrid observation tooling: probe runner, calibration pass, heat-fill pass, and merge script.
- Distributed and single-GPU runbooks.
- Per-run output directory containing `meta.json`, `grid.npy`, and `quantile_edges.npz`.
- Aggregate summary file consolidating per-run metadata across all hosts.
- Probes corpus file used for both calibration and heat-fill passes.

Exact paths and run identifiers are available on request. A regression suite that re-runs the qualification check across all `meta.json` files in the aggregate summary is the recommended next addition.

8 Acknowledgments

This work used a heterogeneous distributed compute substrate spanning Linux x86 hosts, Apple Silicon nodes, and one NVIDIA ARM-GPU workstation, coordinated over a private mesh VPN with a remote archive node serving as merge host. Additional Apple Silicon nodes contributed probe shard execution. The probe corpus consists of synthetic, public-domain, and internally generated prompts spanning English and Chinese, code, mathematics, structured data, and short instruction-following queries.

9 Notes for future work

The natural follow-on questions, in order of urgency:

1. Ablation: isolate the contribution of per-channel quantiles vs $P = 1200$ probes vs both.
2. Models beyond 8B: extend the coverage protocol to 12B, 22B, and 70B classes.
3. Sensitivity: vary Q , S , and `dtype` to map the coverage surface.
4. Statistical floor: derive a `coverage_pct` lower bound below which observed structure cannot be trusted.
5. Cross-run merger: when multiple shards are merged additively, propagate per-shard coverage statistics so the aggregate coverage reflects per-shard variance.