What error rate does the review's activation line imply? A conformal line for 784 VERIFIED poles

Will Sackfield, Open Interpretability · https://open-interpretability.com/research/conformal-line

Abstract

The review counts a probe as firing a pole when its peak reaches a quarter of the pole's maximum, the same share in every model. We asked what error rate that line implies, and what a line with a stated error rate would change. On held-out corpus text, the fixed line fires a median 23.1% of the time in Pythia-70m and 0.7% in Qwen2.5-1.5B. A split-conformal line set for a 5% error rate fires about 5% of the time in all four models, as its guarantee says. With that line in place, 451 of the 784 explanations would still pass their probe gates. The guarantee assumes that new text is like the calibration text, and the reviewer's probes are not. In Qwen2.5-1.5B, 26.2% of the probes predicted quiet fire at that line, five times the bound. In Pythia-70m it is 3.1%. A line that is fair to corpus text is therefore not fair to adversarial probes, and the gap differs by model.

Status: Preprint. Work left: #6551 (choosing the probe gates to bound false VERIFIED verdicts), #6552 (a line from only the windows the explanation doesn't cover)

1 Introduction

The registry decides whether an explanation of a neuron pole holds by testing it on probes. A reviewer writes hard negatives, which the explanation predicts leave the pole quiet, and holdouts, which it predicts fire the pole. A probe fires the pole when its peak strength reaches `activation_threshold` times the pole's recorded maximum, which is 0.25 (`ReviewPolicy`, `core/schemas.py`). The explanation passes when at least 80% of hard negatives stay quiet and at least 60% of holdouts fire.

That line is one share of the maximum for every pole and every model. It states no error rate. A pole whose maximum is far above its usual activity gets a line that ordinary text almost never reaches. A pole whose maximum is close to its bulk gets a line that ordinary text reaches often. So the same 0.80 specificity gate may be easy for one pole and hard for another.

Conformal prediction sets a threshold from held-out data, with a guarantee on its error rate that needs no model of the data (Angelopoulos & Bates, 2021). This paper gives each VERIFIED pole a conformal line. Its contributions are the answers to three questions:

1. **What error rate does the fixed line imply?** On held-out corpus text, the median pole fires on 23.1% of windows in Pythia-70m, 11.1% in GPT-2, 1.6% in Qwen2.5-0.5B and 0.7% in Qwen2.5-1.5B (Table 1). *Limit:* one corpus, and windows rather than probes.

2. **Does the conformal line keep its promise, and what verdicts would it change?** At $\alpha = 0.05$ the median held-out coverage is 0.950, and the lowest of the 784 poles is at 0.913 (Figure 1). With that line, 451 of 784 explanations still pass both probe gates. 174 fail on specificity and 159 on sensitivity, and none fail on both (Table 2). *Limit:* only the two probe gates are read again, because the other gates don't depend on the line.
3. **Does the guarantee hold on the reviewer's probes?** At $\alpha = 0.05$, the probes predicted quiet fire at 3.1% in Pythia-70m and 6.2% in GPT-2, near the bound. In Qwen2.5-0.5B the rate is 16.7%, and in Qwen2.5-1.5B 26.2% (Figure 2, Table 3). *Limit:* this compares probes with 20-token windows, and probes vary in length.

These results do not show that any explanation is wrong. They show that the activation line is a choice, and that its error rate depends on the model and on the text.

The paper bears on research question 2, *How reliable is one verdict?* (research questions).

2 Related work

Split conformal prediction holds out a calibration set, scores each of its points, and takes the $\lceil (n+1)(1-\alpha) \rceil$-th smallest score as a threshold. If the calibration points and a new point are independent and identically distributed, the new point's score is at most that threshold with probability at least $1-\alpha$ (Angelopoulos & Bates, 2021, Section 1.1, Theorem 1). The guarantee holds for any score and any distribution, but only for new points from the calibration distribution. The same text names distribution shift as the case where a test point comes from another distribution, and points to weighted conformal methods for it (Angelopoulos & Bates, 2021, Section 7).

In the registry, the reliability of a verdict has been studied by repeating reviews, swapping hypotheses and reproducing records (`docs/research-questions.md`, question 2). None of those studies sets the activation line itself. This paper keeps every record's probes and changes only the line.

3 Method

Poles. We took each MLP neuron pole whose newest record in the registry is VERIFIED, on 2026-11-24. There are 784: 154 in Pythia-70m, 206 in GPT-2, 350 in Qwen2.5-0.5B and 74 in Qwen2.5-1.5B.

The record's line. The review measures each probe's peak strength over its tokens. A probe fires when its peak is at least `activation_threshold` times `max_activation` (`agents/review_measurements.py`). The record stores every probe's peak.

Windows. For each model we cut `corpus.jsonl`, the pinned corpus, into windows of 20 tokens. That is about a probe's length: the probes' median is 14 words. A seeded shuffle (seed 0) splits them into 1,000 calibration windows and 1,000 held-out windows. One forward pass per window reads each pole's peak strength, skipping the first token where the record has a first-token floor. On 25 Pythia-70m poles, 14 of them with a floor, these peaks equal the review's own `harness.units.peak_strength` exactly.

The conformal line. The score is a window's peak, and $n = 1000$. The quantile is the $\lceil (n+1)(1-\alpha) \rceil$-th smallest calibration peak, for $\alpha = 0.05$ and $\alpha = 0.01$. The theorem

bounds the chance that a new peak is *above* the quantile, but a probe fires when its peak *reaches* the line. So we put the line one floating-point step above the quantile, which keeps the bound for the review's rule when peaks tie. Layer-0 poles often tie, because their activation depends on the token alone. The source notes that ties need a tie-breaking step for the upper bound on coverage (Angelopoulos & Bates, 2021, Section 1). This step is our own, for the lower bound under the review's rule.

Re-reading the verdicts. With a new line, each stored probe passes or fails again, as the review decides: it passes when "reaches the line" matches its prediction. The rates weight near-duplicate probes as the review does (`core.stats.effective_counts`), and group probes by kind. With the record's own line, this reproduces the stored specificity and sensitivity of all 784 records exactly. A pole *passes its probe gates* when both rates reach the record's own gates, 0.80 and 0.60.

Probes predicted quiet. For the guarantee, the relevant probes are those predicted quiet: hard negatives, and the holdouts of an inverse claim. We count how many of them reach the line, pooled over each model's poles.

4 The fixed line on corpus text

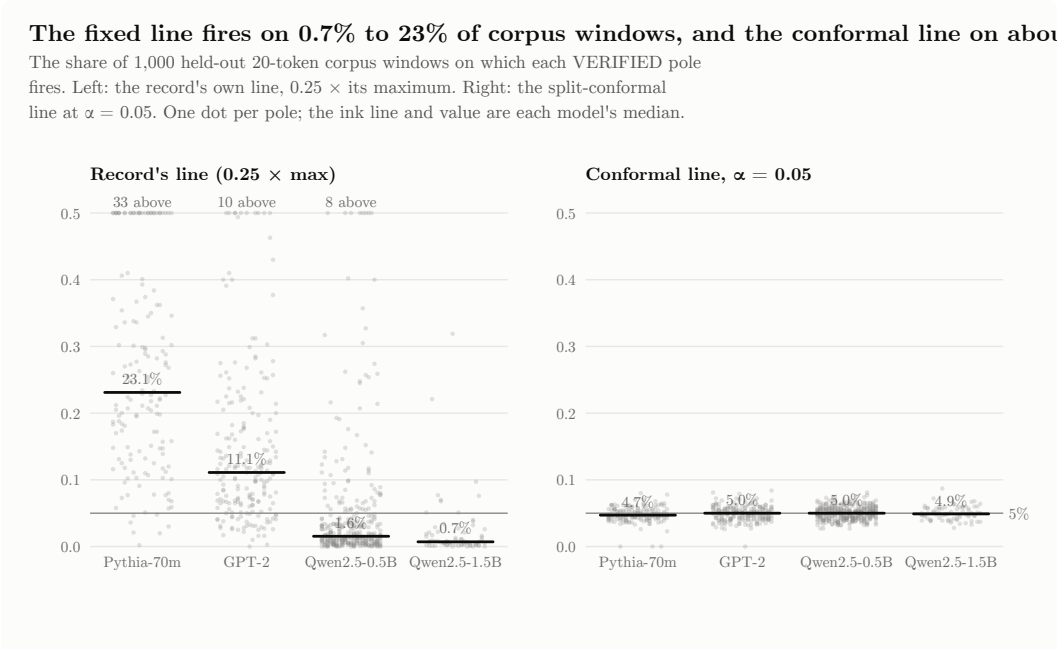

Figure 1. The share of 1,000 held-out corpus windows on which each pole fires. Left: the record's own line. Right: the conformal line at $\alpha = 0.05$. Dots above 50% sit on the top edge and are counted above it.

Model	Poles	Fires at the record's line	Fires at the 5% line	Coverage at the 5% line: median, 5th percentile	Coverage at the 1% line: median, 5th percentile
Pythia-70m	154	23.1%	4.7%	0.953, 0.936	0.991, 0.982
GPT-2	206	11.1%	5.0%	0.950, 0.933	0.991, 0.982
Qwen2.5-0.5B	350	1.6%	5.0%	0.950, 0.934	0.991, 0.984
Qwen2.5-1.5B	74	0.7%	4.9%	0.951, 0.933	0.991, 0.983

Table 1. The median share of held-out windows on which a pole fires, and the conformal line's coverage. The 5% line is at $\alpha = 0.05$ and the 1% line at $\alpha = 0.01$. Coverage is the share of held-out windows under the line.

The record's line is a share of the pole's maximum, and that maximum comes from the pole's own scan. In Qwen2.5 the maxima sit far above the bulk of activity, so the line is high and fires rarely. In Pythia-70m and GPT-2 the maxima are closer to the bulk, so the same share fires often. Under the record's line, 33 Pythia-70m poles fire on more than half of all windows. The conformal line instead fixes the rate, and its coverage stays within sampling noise of $1 - \alpha$ in every model. With 1,000 held-out windows, the standard error of a coverage near 0.95 is about 0.007.

5 Verdicts under a conformal line

Model	Poles	Median ratio of the 5% line to the record's	Pass at the 5% line	Fail specificity	Fail sensitivity	Pass at the 1% line
Pythia-70m	154	1.90	92	7	55	37
GPT-2	206	1.51	131	10	65	67
Qwen2.5-0.5B	350	0.64	203	112	35	208
Qwen2.5-1.5B	74	0.32	25	45	4	42
All	784	1.02	451	174	159	354

Table 2. The probe gates under the conformal line. Every one of the 784 passes both gates with the record's own line. No pole fails both gates at the 5% line.

The two kinds of failure follow the line's direction. Where the conformal line is higher than the record's, as in Pythia-70m and GPT-2, weak holdouts stop firing and sensitivity fails. Where it is lower, as in Qwen2.5, hard negatives start to fire and specificity fails. At $\alpha = 0.01$ every line rises. So more Pythia-70m and GPT-2 explanations fail on sensitivity, and fewer Qwen2.5 ones fail on specificity.

6 Probes predicted quiet

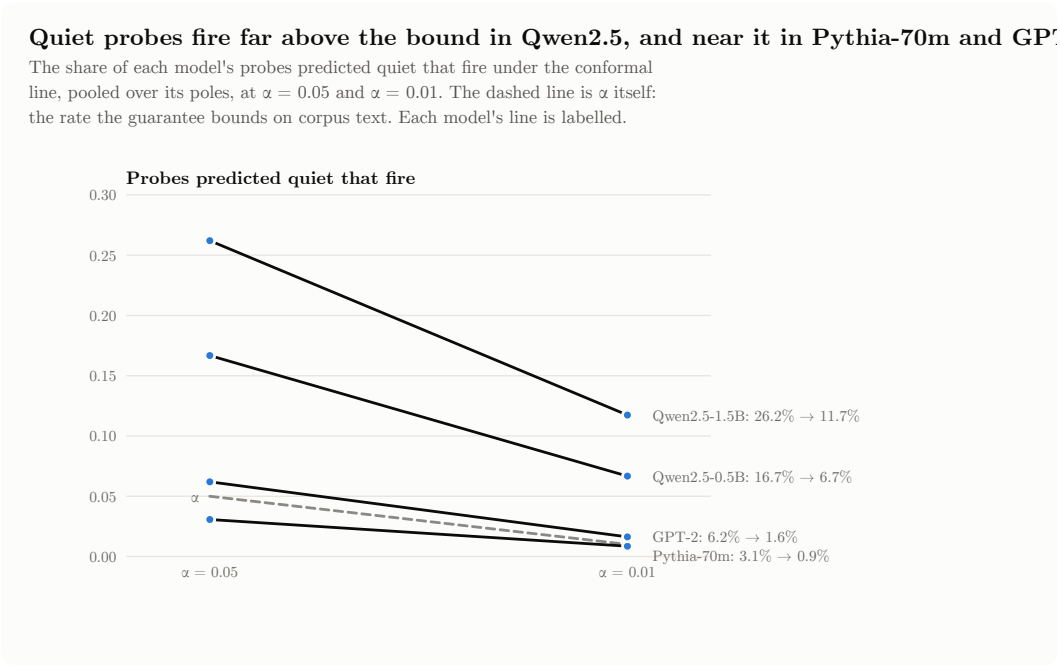

Figure 2. The share of each model's probes predicted quiet that reach the conformal line. The dashed line is α, the rate that the guarantee bounds on corpus text.

Model	Probes predicted quiet	Fire at the record's line	Fire at the 5% line	Fire at the 1% line
Pythia-70m	3,030	9.4%	3.1%	0.9%
GPT-2	2,887	9.4%	6.2%	1.6%
Qwen2.5-0.5B	4,469	7.4%	16.7%	6.7%
Qwen2.5-1.5B	1,244	8.9%	26.2%	11.7%
All	11,630	8.6%	11.5%	4.4%

Table 3. Probes predicted quiet that fire, pooled over each model's poles.

The guarantee needs new text from the calibration distribution. The reviewer writes hard negatives to break an explanation, so they are not drawn from the corpus. In Pythia-70m the probes predicted quiet fire less often than corpus windows. In Qwen2.5 they fire three to five times as often as α.

7 Discussion

The fixed line was chosen as a share of the maximum so that one rule fits every pole. These results show that it does not fit every model in the same way. The same rule lets 23.1% of Pythia-70m's corpus windows fire the median pole, and 0.7% of Qwen2.5-1.5B's. A specificity gate of 0.80 therefore asks less of a Qwen2.5 explanation's hard negatives than the corpus would suggest, and more of a Pythia-70m one.

A conformal line fixes the rate on corpus text, and keeps that promise in every model. It does not fix the rate on the reviewer's probes, which are adversarial by design. We did not test why Qwen2.5's probes cross the line so often. One possibility is that the conformal line in Qwen2.5 sits low, in the bulk of the activations, where near-miss text reaches it easily. Either way, no single α gives the same false-fire rate on probes in every model.

So a conformal line alone would not make the verdicts comparable across models. It would change which explanations pass, in opposite directions by model, without a guarantee on the probes that decide the verdict. A rule with a stated error rate on false VERIFIED verdicts would need labelled examples of wrong explanations (#6551).

8 Limitations

- **One corpus for every pole.** Each record pins its own scan corpus, but all lines here come from `corpus.jsonl`. A pole's line would differ on its own corpus.
- **Every window calibrates, not only the uncovered ones.** The issue asked for the windows that the explanation does not cover. Labelling those needs a judge per window, so this paper uses all windows. Covered windows raise the line, so these lines are, if anything, too high. #6552 tests how much this matters.
- **Windows of one length.** A probe's peak is over its own tokens, and probes vary in length. A longer text has more chances to cross the line. So the comparison in Table 3 is approximate.
- **Windows are not independent.** Neighbouring windows come from the same document, and the split is by window, not by document. The guarantee assumes independent, identically distributed scores.
- **Only the probe gates.** A VERIFIED verdict also needs counterfactuals, a lookahead check, an independent review and a confound control. The line doesn't enter those, so a pole that passes its probe gates here still passes the rest.

9 Conclusion

The review's fixed activation line implies a false-fire rate on corpus text that differs about thirtyfold between models. A split-conformal line removes that difference on corpus text. But it changes 333 of 784 probe verdicts at $\alpha = 0.05$, and it gives no guarantee on the reviewer's adversarial probes. Three steps would take this further:

- Choose the probe gates to bound the rate of false VERIFIED verdicts, with Learn then Test or conformal risk control (#6551).
- Calibrate the line only on windows that the explanation does not cover, and see whether the verdicts move (#6552).
- Decide whether the review should adopt a conformal line. That is a change to the decision rule, so it needs its own issue and a `PROTOCOL_VERSION` bump (`docs/PROTOCOL.md` §6).

References

Angelopoulos, A. N., & Bates, S. (2021). *A gentle introduction to conformal prediction and distribution-free uncertainty quantification.* Preprint, version 6 of 7 December 2022. arXiv:2107.07511.

Appendices

Reproduce

```
uv run python scripts/conformal_line.py --out papers/conformal-line/results.jsonl
uv run python scripts/paper_figures.py
```

The script reads the newest record of each pole from `wiki/` on the day it runs. It ran on an M1 Max in about 15 minutes, with the models already downloaded. The window shuffle uses seed 0.