CALIBRATED PROBES FOR LANGUAGE MODEL DECISIONS

Jai Pal

30 October 2026

ABSTRACT

A probe is a small fitted predictor that reads a language model's internal state. Separate tutor probes estimate visible understanding, comprehension of an explanation, and success on a later problem. Probes predict current evidence well in these cases, but future forecasts remain weak. A fitted probability adjustment improves some held-out measures, but the fixed analysis lacks tests that isolate calibration. A second study selects 45 of 54 agent policy errors for review, versus 29 for adjusted actor confidence at the same budget. Both studies use synthetic cases; the agent study assumes perfect repair.

1 Tutor targets and test design

The tutor study fits separate probes for three distinct targets:

1. **Demonstrated understanding:** Visible work gives a valid explanation or an independent, justified use of the focal concept.
2. **Explanation comprehension:** The student correctly interprets or uses the tutor's latest substantive explanation.
3. **Next-problem success:** The student answers a later independent problem correctly before further help.

Each target has its own probability. The probabilities do not sum to one or measure a percentage of mastery. Absent or ambiguous evidence receives an unknown label, not a negative label. The input ends before the future answer.

The six Qwen models range from 4 to 72 billion (B) parameters, the fitted numbers in a model. The model weights stay fixed while a linear probe fits each target. A token is a unit of model input. The probe reads the mean or last internal state over the final student tokens, before the tutor responds.

Development data select the state location and fit penalty. The study also tests Cloudflare's released Clef 27B and Clef-Flash 9B with fixed questions. The probes receive task-specific fits; Clef retains its released weights. This unequal adaptation limits comparisons of the two methods.

Table 1: Conversation data. Whole related pairs stay within one split.

	Training	Development	Calibration	Final test
Conversation prefixes	720	120	240	240
Related pairs	360	60	120	120

Model agents author the conversations and future outcomes. Separate blind model reviews judge current evidence and future answers; a third review resolves disagreement. These labels do not provide independent human ground truth. Known final labels cover 200 understanding cases, 196 comprehension cases, and 239 future outcomes. Each source group receives equal total weight in the reported metrics. The study fixes all fits, probability adjustments, and cutoffs before final evaluation.

2 Calibration, probability error, and decisions

Calibration means that predicted probabilities match observed outcome frequencies in the relevant population. For example, about 80% of cases assigned probability 0.8 should satisfy the target. Accuracy measures correct decisions at a cutoff. Discrimination measures whether scores place positive outcomes above negative outcomes. AUROC measures this separation across cutoffs; 0.5 indicates chance and 1.0 indicates perfect order.

Brier score is mean squared probability error. Log loss penalizes confident mistakes more strongly. Neither measure isolates calibration.

The tutor study fits a separate probability adjustment on calibration groups:

$$p_{\text{adj}} = \frac{1}{1+\exp[-(az+b)]}, \qquad 0 \le a \le 10.$$

Here, z is the raw score before conversion to a probability. The fit minimizes group-weighted log loss with penalty $\frac{1}{2}[(a-1)^2+b^2]$. The known calibration sets contain 188, 205, and 240 rows for the three targets, respectively. They contain 108, 111, and 120 source groups. The final test remains separate.

A fitted adjustment does not itself prove calibration. The fixed analysis contains no reliability curve, which compares predicted probabilities with observed frequencies. It also contains no separate calibration-error statistic. It does not establish reliable probability levels for real student traffic.

The study selects decision cutoffs on the same calibration split. This is a separate operation from the probability adjustment. For math errors, one cutoff targets a false flag in at most 5% of negative calibration groups. This empirical target does not guarantee the same rate on new groups.

3 Tutor results

Current-evidence probes separate the study labels well (Figure 1). Within Qwen3.5, raw understanding Brier falls from 0.0535 at 4B to 0.0101 at 27B. Comprehension Brier falls from 0.0409 to 0.0112. The paired 95% intervals for these reductions are 0.0165–0.0730 and 0.0105–0.0526. The intervals resample whole conversation groups. Model family, architecture, and prior training also differ; these results do not isolate parameter count as the cause.

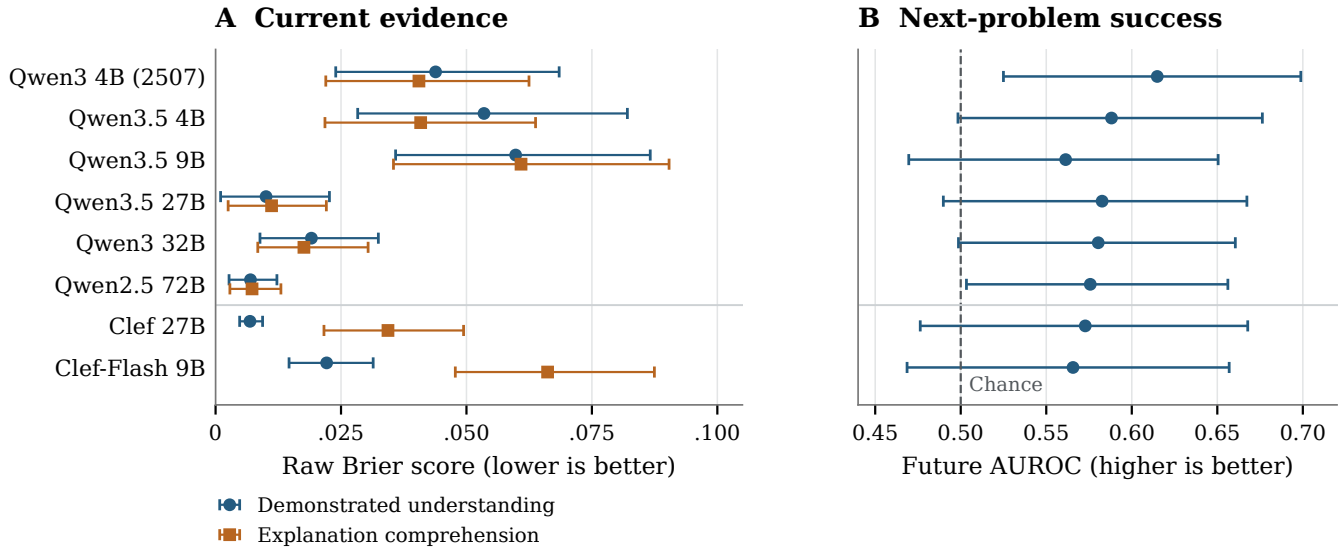


Figure 1: Full-conversation results before the probability adjustment. Current targets use 200 and 196 known labels; the future target uses 239. Bars show archived 95% intervals from 1,000 group resamples. Different targets have different labels and coverage; compare models within a target.

Future-outcome AUROC ranges from 0.561 to 0.615 across the six probes. The 27B interval, 0.490–0.667, includes chance. Table 2 shows that the adjustment can reduce probability error without better discrimination. Clef's future Brier falls from 0.3605 to 0.2463, but its AUROC stays at 0.573. The 27B probe's future Brier falls from 0.2602 to 0.2504, while its AUROC stays at 0.583. Both remain near the text baseline's raw Brier of 0.2479.

The fitted maps have positive slopes, so they preserve score order. The adjustment raises 27B understanding Brier from 0.0101 to 0.0105 and also raises its log loss.

Table 2: Final probability error before and after the fitted adjustment. Lower values are better. These measures do not isolate calibration.

		Brier score		Log loss	
System	Target	Raw	Adjusted	Raw	Adjusted
Qwen3.5 27B probe	Demonstrated understanding	0.0101	0.0105	0.0421	0.0432
	Explanation comprehension	0.0112	0.0106	0.0343	0.0370
	Next-problem success	0.2602	0.2504	0.7230	0.6950
Clef 27B	Demonstrated understanding	0.0069	0.0011	0.0646	0.0090
	Explanation comprehension	0.0344	0.0143	0.1556	0.0518
	Next-problem success	0.3605	0.2463	1.0713	0.6875

The related math test contains 2,304 responses from 96 final problems: 576 correct responses and 1,728 errors. Exact arithmetic checks these labels. The Qwen3.5 27B probe detects 1,412 errors (81.7%) with three false flags at the low-false-alarm cutoff. At the raw 0.5 cutoff, it detects 1,597 errors (92.4%) with 61 false flags. The high-recall cutoff detects 1,718 errors (99.4%) with 201 false flags. These are decision tradeoffs; they do not prove probability calibration.

4 Limits of student-state estimates

Earlier turns supply relevant evidence. The 72B understanding Brier rises from 0.0070 to 0.0746 with only the latest exchange. Its comprehension Brier rises from 0.0073 to 0.0522. Both views use full-conversation labels, so the reduced view lacks some required evidence by design.

Separate evidence probes allow an unknown output. The 27B understanding check accepts all 200 known cases and one of 40 unknown cases. Its comprehension check accepts 192 of 196 known cases and eight of 44 unknown cases. A low target score and absent evidence therefore need distinct interpretations.

Format changes expose failures. The 27B math probe detects 55 of 192 alternate-notation errors (28.6%), versus 506 of 576 core errors (87.8%). These variants change step notation, not the solution method. The 72B probe flags 4 of 384 correct responses under familiar references, versus 25 of 192 under table references. The table format changes several features together, so this test does not isolate one cause.

The current targets do not establish durable mastery or a strong forecast of future student behavior. The future answers are authored outcomes, not observed student results. Synthetic outcome rates also limit the interpretation of adjusted probabilities.

5 A second use case: agent review selection

A separate pilot uses Qwen3-4B-Instruct-2507 on policy and evidence tasks with fixed A/B choices. The actor is the model that chooses an action. A probe estimates the probability that each original actor choice is correct. The pilot selects the highest estimated error scores at a nominal 20% review budget. Its confidence control fits a scalar probability map on training data, with development selection. This control differs from the tutor study's separate calibration split.

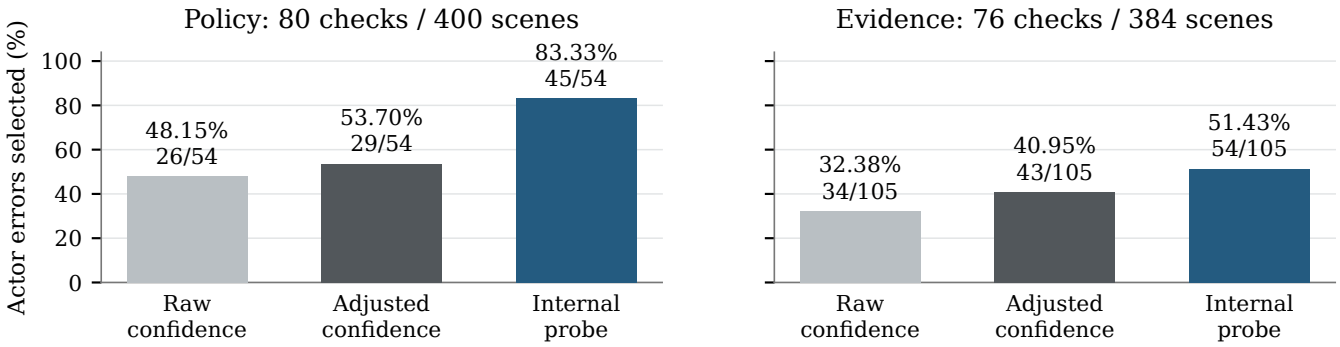


Figure 2: Original actor errors within selected cases. Labels show percentages and error counts. Adjusted confidence is the pilot's fitted confidence control. The evidence budget rounds down to 76 checks.

The probe selects 45/54 policy errors (83.33%), versus 29/54 (53.70%) for adjusted confidence. For evidence, the selected cases contain 54/105 errors (51.43%) versus 43/105 (40.95%). Figure 2 gives the equal case and check counts.

These counts measure review selection; no real external checker ran. Perfect repair would give 97.75% policy accuracy with probe selection, versus 93.75% with adjusted confidence. Those totals are simulated, not measured deployment gains.

The pilot also exposes limits. Injected instructions reduce policy-probe accuracy from 92.71% to 55.21% on 96 fixed scenes. Exact text parsers with the full rule specification solve all original scenes. Useful review selection therefore does not establish a reliable security barrier or superiority to available exact rules.

A stronger tutor test needs real student records, independent outcomes, and reliability checks across score ranges. Agent tests need real checkers and measures of repair success under a fixed cost budget.