# RANKINGS SURVIVE, SCALES DRIFT: HIDDEN-STATE ERROR ESTIMATION ACROSS POST-TRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Hidden-state probes offer an efficient way to estimate whether a vision-language model (VLM) will answer correctly. Yet post-training changes both the representations a probe reads and the errors it must predict. We ask whether a probe remains useful as the model evolves, separating error ranking from probability calibration. Our central finding is that a frozen probe can retain useful rankings even when its probability estimates become unreliable. We characterize ranking stability in terms of score perturbations and establish that updated error probabilities are not identifiable from unlabeled scores alone. Under a common score shift that preserves the conditional error relation, we derive a finite-sample calibration guarantee for re-centering a calibrated base probe using only unlabeled inputs. Changes in this relation call for labeled recalibration; deteriorating rankings motivate refitting. We further find that sampled disagreement can supervise transferable rankings without correctness labels, and that hidden-state probes remain informative when reasoning precedes the answer and first-token uncertainty is less predictive. Experiments across VLM post-training methods and answer formats support these findings. These results separate the cost of recalibrating a probe from that of relearning its error signal.

## 1 INTRODUCTION

Knowing when a vision-language model (VLM) is likely to be wrong helps determine when to trust its answers or seek further verification. Prior work on hidden-state probing provides a starting point: a lightweight predictor reads an internal representation and scores answer correctness (Kadavath et al., 2022; Azaria & Mitchell, 2023; Orgad et al., 2025) or uncertainty over possible answers (Kossen et al., 2024). At inference time, this provides an error signal from a single forward pass, avoiding the repeated generations used by sampling-based uncertainty estimators (Kuhn et al., 2023; Farquhar et al., 2024).

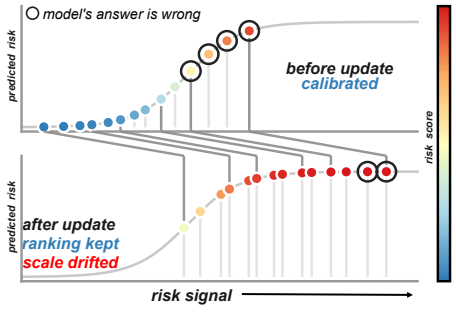



Figure 1: **Ranking survives, calibration drifts.** Black circles indicate incorrect answers.

Yet a VLM does not remain fixed after a probe is trained. Supervised fine-tuning (SFT), reinforcement learning with verifiable rewards such as GRPO (Shao et al., 2024; Guo et al., 2025), and on-policy distillation (OPD) (Agarwal et al., 2024) continually adapt its behavior. These updates can change both sides of the relation learned by the probe: the hidden states it reads and the errors it predicts. An update may improve accuracy, resolve previously uncertain answers, or change which inputs remain difficult. Even on the same questions, the probe therefore faces changing representations and changing prediction targets. Improved VLM accuracy does not guarantee a reliable error monitor.

This raises a question about the lifetime of an error estimator: **can a probe trained on one VLM checkpoint still tell us when its updated version is wrong?** Success on one checkpoint does not establish this compatibility. Refitting after every update requires new supervision, while unchecked



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

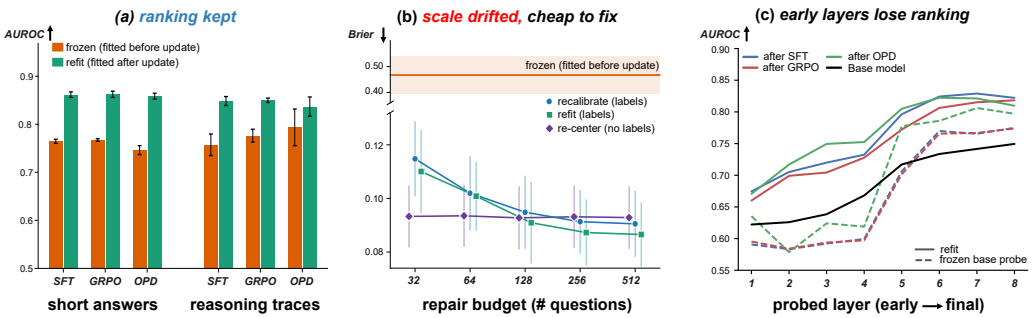



Figure 2: **Probe transfer and repair after post-training.** **(a)** Error AUROC of frozen and refitted probes on short answers and reasoning traces. **(b)** Brier score across repair budgets, comparing labeled recalibration and refitting with unlabeled re-centering; the orange line marks the frozen probe, and the vertical axis is broken. **(c)** Error AUROC across probed layers: solid curves denote probes fitted in the evaluated state, and dashed curves denote transferred base probes.

reuse risks relying on an outdated relation between scores and errors. Identifying what can be reused and what must be relearned is central to reliable monitoring as VLMs evolve.

Our starting point is that probe reuse has two meanings. A probe can *rank* answers by risk, identifying which deserve further checking, or estimate *error probabilities*, expressing how often answers with a given score are wrong. These abilities need not survive together. Figure 1 illustrates this separation: after an update, incorrect answers can remain among the highest-scored cases even as the frozen probe's probability estimates become miscalibrated. The Brier-score decomposition separates miscalibration—the mismatch between predicted probabilities and conditional error rates—from the error uncertainty remaining given a prediction (Brier, 1950; Murphy, 1973; DeGroot & Fienberg, 1983). An increasing transformation preserves score order but can change calibration. We therefore examine ranking and probability calibration separately.

This separation explains why changing hidden states need not invalidate the entire probe. Under a local smoothness assumption, we bound how a fixed readout changes with the model and relate ranking stability to score perturbations relative to the gaps between examples, isolating this effect from changes in error labels. A shift shared by all scores leaves their ordering intact. Error probabilities depend on the relation between scores and the updated model's errors. Unlabeled scores alone cannot identify that relation: the same score distribution is compatible with different error probabilities. This limits unlabeled repair even when an update improves accuracy.

Within this limit, the structure of the drift offers a simple repair. If an update shifts the scores while preserving their relation to errors after translation, we can estimate the shift from unlabeled inputs and *re-center* the frozen probe. This retains its ranking and restores the probability map under the stated assumptions; our sample-cost bound shows how estimation error decreases with the number of inputs. An unchanged overall error rate alone is insufficient. When the conditional error relation changes, labels are needed to recalibrate the map; when ranking deteriorates, the probe itself may need refitting. The resulting approach separates the cost of maintaining a probability scale from the cost of relearning an error signal.

We study this distinction across five VLMs, eight short-answer tasks, and a long-form reasoning protocol. Frozen probes retain useful error rankings while their probabilities drift, and re-centering repairs much of the mismatch where a common shift describes the update (Figure 2). Accuracy changes expose the remaining need for labeled recalibration, supporting the analysis within the evaluated post-training regimes.

The value of maintaining a probe also depends on how its signal is learned. Disagreement among sampled answers can supervise a transferable error ranking without correctness labels. First-token uncertainty is competitive for immediate answers but becomes less informative when reasoning precedes the answer, where hidden-state probes retain a useful error signal. Together, these findings motivate maintaining monitors alongside the VLMs they assess: preserve useful readouts, repair their probability interpretation, and focus supervision on what each update changes.

## 2 PRELIMINARIES

**Prediction target and observations.** A model $\theta$ produces a greedy response $a_\theta(x)$ to input $x$. We predict its error event $E_\theta(x) = 1 - \mathrm{Correct}(a_\theta(x), x) \in \{0, 1\}$, determined by a task-specific external scorer. We check the final-layer state $H_\theta(x) \in \mathbb{R}^d$ at the last prompt token (*prefill*), after 32 generated tokens (*Prefix-32*), or after the answer but before confidence content (*answer end*). Token statistics $L_\theta(x)$ capture entropy, maximum probability, margin, and realized surprisal from the same prefix. Both use only information available at the specified observation time. Verbalized confidence $V_\theta(x) \in [0, 1]$ is evaluated after generation as the error score $1 - V_\theta(x)$.

**Hidden-state scores and targets.** A probe maps the hidden state to a scalar score,

$$s_\theta(x) = w^\top \nu(H_\theta(x)) + b_h,$$

where $\nu$ standardizes the features and $(w, b_h)$ are fitted parameters. The score lies in $\mathbb{R}$. Training targets are either $E_\theta$ or sampled disagreement, which requires no correctness labels. For $K$ responses sampled at temperature one, with answer-class counts $N_c$, disagreement is $Y_K(x) = 1 - \max_c N_c/K$ (Wang et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). It measures concentration, not correctness: a model can be consistently wrong.

**Probe fitting and probability outputs.** The main short-answer $Y_8$ probe uses ridge regression with squared loss; its output is not an error probability. Short-answer error probes use logistic regression, fitting $q_\theta = \sigma(s_\theta)$ to $E_\theta$ by cross-entropy, where $\sigma$ denotes the sigmoid. Long-trajectory error probes first fit $s_\theta$ to $E_\theta$ by ridge regression, then freeze the score and fit a logistic map $g(s) = \sigma(\alpha s + \beta)$ by cross-entropy on a disjoint calibration set (Platt, 1999). The scalar score's calibration-stage standardization is absorbed into $(\alpha, \beta)$. Both error probes thus output $q_\theta = g(s_\theta)$; the short-answer probe initially has $(\alpha, \beta) = (1, 0)$ and no separate map-fitting stage. Protocol details and matched-target comparisons are in Sections B.2 and D.

**Transfer across model updates.** An update $\theta_0 \to \theta_1$ can change both $H_\theta(x)$ and the model's answer, hence its error label. We evaluate transferred probes against the updated model's errors $E_{\theta_1}$. Freezing the base probe gives

$$s_{0\to1}(x) = w_0^\top \nu_0(H_{\theta_1}(x)) + b_{h,0}, \qquad q_{\mathrm{F}}(x) = g_0(s_{0\to1}(x)).$$

For error-probability estimation, we compare three maintenance policies:

- FROZEN: retain the normalization, score parameters, and probability map; use no target-model labels.
- RECALIBRATE: retain the normalization and score parameters; refit the probability map using a target-model calibration budget of $n$ labels.
- REFIT: refit the normalization and score on target-model fit data, with the same calibration budget for the map in the two-stage protocol.

For the short-answer disagreement probe, RECALIBRATE instead fits a scalar affine map to $Y_8$, and REFIT refits the disagreement predictor; these fits use sampled disagreement targets rather than correctness labels.

**Ranking and calibration.** We measure error ranking with AUROC, AURC, and selective risk at fixed coverage (Geifman & El-Yaniv, 2017), and probability quality with Brier score and expected calibration error (ECE, ten bins) (Guo et al., 2017). For a probability estimate $q$, let $\pi_q(t) = \Pr(E = 1 \mid q = t)$ be its calibration function. The Brier score decomposes as (Murphy, 1973; DeGroot & Fienberg, 1983)

$$\mathrm{Brier}(q) = \mathbb{E}\big[(q - E)^2\big] = \underbrace{\mathbb{E}\big[(q - \pi_q(q))^2\big]}_{\text{calibration}} + \underbrace{\mathbb{E}\big[\pi_q(q)(1 - \pi_q(q))\big]}_{\text{refinement}}. \tag{1}$$

The first term measures mismatch between predicted probabilities and conditional error rates; the second is the remaining error uncertainty given $q$. For disagreement probes, we report error AUROC and MSE to $Y_8$ separately; the latter measures prediction of the proxy, not calibration of error probabilities.

# 3 Understanding Probe Transfer Across Post-Training

A probe learns to estimate a model's internal state, but post-training can change both that state and the model's accuracy or uncertainty. Can the old probe still be used? Following Figure 3, we separate ranking from probability estimation, examine when ranking survives and how it can be learned, then turn to probability drift and repair.

We fix the base probe, writing $h_\theta = \nu_0(H_\theta)$ and $s_\theta = w_0^\top h_\theta + b_{h,0}$. Let $s_j = s_{\theta_j}$ and $E_j = E_{\theta_j}$, so $s_1$ is the transferred score $s_{0\to1}$ in Section 2. The frozen probability is $g_0(s_1)$, whereas the true conditional error rate is $\pi_j(t) = \Pr(E_j = 1 \mid s_j = t)$. Proofs and supporting identities are in Section A.

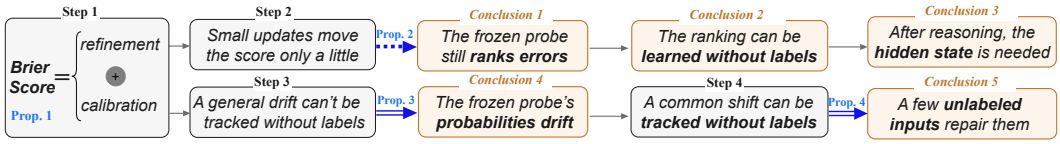



Figure 3: **Why rankings can survive while calibration drifts.** The Brier decomposition motivates separate analyses of ranking (top) and calibration (bottom). The upper branch examines ranking stability, disagreement supervision, and observation time; the lower branch examines probability drift and repair. Steps 1–4 correspond to Propositions 1–4 (proofs in Section A). Blue arrows distinguish theoretical support: solid arrows ($\Longrightarrow$) denote exact results on non-identifiability and calibration recovery under a common shift preserving the score–error relation; the dashed arrow ($\dashrightarrow$) links the local score-drift bound to ranking stability under fixed error labels and sufficiently separated scores, with ranking retention tested empirically.

## 3.1 Ranking and Probability Estimation

Reusing a probe requires distinguishing error ranking from probability estimation. The Brier decomposition in eq. (1) provides the starting point (Step 1 in Figure 3): it separates calibration from the uncertainty remaining given a prediction. A strictly increasing transformation preserves score ordering, but need not preserve calibration.

**Proposition 1** (Order versus scale)**.** *Let $s$ be a real-valued score and $\phi$ strictly increasing. Then $\mathrm{AUROC}(\phi \circ s) = \mathrm{AUROC}(s)$, and AURC and the risk at any fixed accepted coverage are also unchanged. Among all maps of $s$, the Brier score is minimized by $\pi_s(s) = \Pr(E = 1 \mid s)$, whose calibration term in eq. (1) is zero, and every other map $\phi$ pays exactly $\mathbb{E}[(\phi(s) - \pi_s(s))^2]$ in excess Brier score.*

An old probe can thus remain useful for selecting safer answers even when its probability estimates need repair. We first follow the ranking branch of Figure 3, then examine the probability scale.

## 3.2 Score Drift and Ranking Stability

Changing hidden states need not invalidate their linear readout. Step 2 bounds how far an update moves the frozen score. Fix the input and text prefix, and write $\theta_1 = \theta_0 + Uu$, with $U^\top U = I$. Here $U$ spans the update directions, including adapter-induced changes (Hu et al., 2021), and $u$ denotes update coordinates rather than calibration parameters.

**Proposition 2** (Local score drift)**.** *Fix the input distribution $P$ and suppose the gradient of the frozen score $s_\theta$ is $\Lambda$-Lipschitz along the segment from $\theta_0$ to $\theta_1$. Then*

$$\|s_{\theta_1} - s_{\theta_0}\|_{L_2(P)} \le \kappa_U \|u\|_2 + \tfrac{\Lambda}{2}\|u\|_2^2, \qquad \kappa_U^2 = \mathbb{E}_P\left[\|U^\top \nabla_\theta s_{\theta_0}(x)\|_2^2\right]. \tag{2}$$

Score drift depends on sensitivity to the update directions; ranking stability additionally depends on the separation between erroneous and correct examples.

**Corollary 1** (Rank stability)**.** *Let $X_1, X_0$ be independent draws of an erroneous and a correct input, with no cross-class ties in the original score. If $|s_{\theta_1} - s_{\theta_0}| \le \delta$ except on a set of probability at most $\eta_e$ under each class $e$, then $\mathrm{AUROC}(s_{\theta_1}) \ge \mathrm{AUROC}(s_{\theta_0}) - \Pr\left(0 < s_{\theta_0}(X_1) - s_{\theta_0}(X_0) \le 2\delta\right) - \eta_1 - \eta_0$, where both AUROCs are computed against the same error labels.*

The corollary isolates score drift. Error labels and generated prefixes can also change; we examine these effects in Sections C and F. Decompose the drift as

$$s_1(x) - s_0(x) = \Delta + r(x), \qquad \Delta = \mathbb{E}[s_1(X) - s_0(X)], \quad \mathbb{E}[r(X)] = 0.$$

The common shift $\Delta$ cancels in pairwise comparisons; only $r(x) - r(x')$ can reorder inputs.

**Conclusion 1** (The frozen probe still ranks errors)**.** The frozen probe can retain ranking when residual score changes are small relative to score margins, subject to changes in error labels. This motivates testing probe reuse under our low-rank, anchored updates. Ranking retention and the predominance of a common shift are evaluated empirically; neither follows from update size alone.

### 3.3 Supervision targets and observation times

If a ranking remains useful across updates, two further questions determine the value of reusing it: must it be learned from correctness labels, and when does it offer more than token uncertainty? These extend the upper branch of Figure 3.

For $m_Z(H) = \mathbb{E}[Z \mid H]$, Equation (3) gives the excess squared error of $m_Y$ over $m_E$ as $\mathbb{E}[(m_E(H) - m_Y(H))^2]$. Disagreement is useful insofar as it reflects the conditional error signal. Under a zero-mean sampling-noise model, regression on $H$ can also average away finite-bank noise (Section A). These identities motivate disagreement supervision, while its error ranking must be tested empirically.

**Conclusion 2** (Rankings can be learned without correctness labels)**.** A probe trained on sampled disagreement can learn a transferable error ranking without correctness labels. Our short-answer experiments support this claim and show that the probe outperforms its sampled target in error AUROC; the squared-error identities alone do not guarantee that improvement.

The observation time also matters. For immediate answers, first-token uncertainty reflects the initial answer choice; with a reasoning preamble, it describes the start of reasoning (Nakkiran et al., 2026). The oracle identity in Section A shows that adding $H$ to $L$ cannot worsen optimal Brier prediction, without guaranteeing gains for fitted estimators.

**Conclusion 3** (Hidden states provide the error signal after reasoning)**.** In our evaluated protocols, first-token uncertainty matches the hidden-state probe for immediate answers but loses predictive power when reasoning precedes the answer, where the probe remains informative. The need for hidden states in Figure 3 refers to this comparison, rather than a universal limitation of token-based estimators.

### 3.4 Identifying probabilities after an update

We now turn to the calibration branch of Figure 3. A useful ranking does not specify error probabilities: improved accuracy changes the prediction target, and even at an unchanged error rate, errors may move between inputs. Step 3 establishes why unlabeled scores alone cannot identify the resulting conditional error rates.

**Proposition 3** (Non-identifiability of the scale)**.** *For any distribution of scores $\mathcal{M}$ and any two calibration functions $\pi_1 \neq \pi_2$ on a set of positive $\mathcal{M}$-measure, there are joint distributions of $(s, E)$ with the same score marginal $\mathcal{M}$ and calibration functions $\pi_1$ and $\pi_2$. Hence no map computed from unlabeled scores of the updated model can be calibrated for every update.*

The frozen map incurs excess Brier score $\mathbb{E}[(g_0(s_1) - \pi_1(s_1))^2]$ relative to the optimal map of $s_1$, without determining its AUROC.

**Conclusion 4** (The frozen probe's probabilities drift)**.** The frozen probability map can become miscalibrated while the score remains useful for ranking, as observed in our post-training experiments. Non-identifiability rules out a universal unlabeled repair; it does not imply that every update causes drift. Repair without labels therefore requires additional structure.

### 3.5 Calibration repair under a common shift

A common shift supplies such structure (Step 4 in Figure 3). Unlike a general change in conditional error rates, a shift of the score distribution can be estimated without labels. Estimate the difference of

score means on unlabeled inputs from the same input distribution and retain the old probability map:

$$\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0, \qquad q_{\mathrm{RC}}(x) = g_0(s_1(x) - \hat{\Delta}).$$

This preserves score order. Calibration additionally requires the score–error relation to be preserved after translation.

**Proposition 4** (Re-centering). *Suppose $s_1(x) = s_0(x) + \Delta$ and $\pi_1(t) = \pi_0(t - \Delta)$, and the frozen map $g_0 = \pi_0$ is continuous and strictly increasing. Then $\Delta = \mathbb{E}[s_1(X)] - \mathbb{E}[s_0(X)]$ is identified from unlabeled inputs, and $g_0(s_1 - \hat{\Delta})$ is calibrated for the updated model as $\hat{\Delta} \to \Delta$; its AUROC equals the frozen AUROC.*

For the sample-cost analysis, write the frozen logistic map in standardized coordinates as $g_0(s) = \sigma(a(s - \hat{\mu}_0)/\sigma_0 + b)$, where $\sigma_0 > 0$ is the base score standard deviation and $a > 0$. This is the same map as in Section 2, with $\alpha_0 = a/\sigma_0$ and $\beta_0 = b - a\hat{\mu}_0/\sigma_0$. Re-centering replaces $\hat{\mu}_0$ by the target score mean $\hat{\mu}_1$.

**Corollary 2** (Sample cost of re-centering). *Under Proposition 4, let the base and target means be estimated from independent samples of $n_0$ and $n$ inputs. For the logistic map above, the expected calibration term after re-centering is at most $\frac{a^2}{16}\big(\sigma_1^2/(\sigma_0^2 n) + 1/n_0\big)$, where $\sigma_1^2$ is the target score variance, whatever the error rate $p$. A labeled sample of size $m$ contains no error at all with probability $(1-p)^m$.*

**Conclusion 5** (A few unlabeled inputs repair the probability scale). Under these conditions, probability repair requires only estimating a shift from unlabeled inputs. The target-mean contribution decreases as $1/n$ without requiring observed errors. As with normalization adaptation and prior-shift correction (Li et al., 2016; Saerens et al., 2002; Lipton et al., 2018), the guarantee depends on an invariance assumption: an unchanged error rate alone is insufficient. Changes in conditional error rates motivate labeled recalibration; loss of ranking motivates refitting.

# 4 Experiments

## 4.1 Experimental Setup

**Models and post-training operators.** We use five vision-language models from three families: Qwen3-VL-2B and -8B (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025), and Qwen3.5-4B and -9B (Qwen Team, 2026). From each base model we derive updated states with three operators, all with rank-16 LoRA adapters (Hu et al., 2021). **SFT** minimizes token cross-entropy on gold answers for one epoch over 5,629 training examples (three training seeds). **GRPO** continues from the SFT state with group-relative, clipped policy optimization (Shao et al., 2024), a correctness reward, groups of four on 512 prompts, and a KL anchor to the SFT state ($\beta = 0.04$). **OPD** continues from the SFT state by minimizing the full-vocabulary KL divergence from a stop-gradient teacher on the student's own samples (Agarwal et al., 2024; Gu et al., 2024), with Qwen3-VL-8B teaching Qwen3-VL-2B and Qwen3.5-9B teaching Qwen3.5-4B. All operators train on image-disjoint training splits of the visual question answering tasks below, with three training seeds each; Section B.1 lists every hyperparameter.

**Two measurement protocols.** The short-answer protocol evaluates all states on eight tasks—SLAKE (English and Chinese) (Liu et al., 2021), VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2020), ScienceQA (Lu et al., 2022), A-OKVQA (Schwenk et al., 2022), POPE (Li et al., 2023), and VizWiz (Gurari et al., 2018)—with an immediate-answer prompt, so that 80.8% of answers have at most two tokens. The probe uses the final-layer hidden state at the last prompt token. This gives 736 model–state–task cells spanning three training seeds per operator and, for GRPO and OPD, fresh and archived rollouts and two KL anchors. The long-trajectory protocol asks Qwen3-VL-2B and -8B and their SFT, GRPO, and OPD states (OPD for the 2B student) to explain their reasoning in ordinary text and then give an answer and a confidence in $[0, 1]$, on ScienceQA and the text-only GSM8K (Cobbe et al., 2021) (greedy decoding, up to 2,048 new tokens, no forced format). Each task contributes 512 fit, 512 calibration, and 512 evaluation problems with disjoint groups, and the probe uses the hidden states at Prefix-32 and at the answer end. Metrics use rows whose answers can be graded (91.2% at the answer end; Section B.2). Table 1 reports means and standard deviations over the three training seeds; the three reasoning-trace cells whose 32-label budget is undefined are excluded from every row of their seed.

Table 1: **Main results: risk estimators after post-training.**

| | Signal | Estimator | Labels needed | after SFT AUROC↑ | after SFT ECE↓ | after GRPO AUROC↑ | after GRPO ECE↓ | after OPD AUROC↑ | after OPD ECE↓ |
|---|---|---|---|---|---|---|---|---|---|
| Short answers 5 VLMs, 8 tasks | Probe before the update | | | .858 | .079 | .858 | .079 | .858 | .079 |
| | Output | Answer probability | 0 | $\mathbf{.850}_{\pm.006}$ | $\mathbf{.087}_{\pm.002}$ | $\mathbf{.845}_{\pm.009}$ | $\mathbf{.099}_{\pm.001}$ | $\mathbf{.851}_{\pm.008}$ | $\mathbf{.086}_{\pm.001}$ |
| | | Self-consistency | 0 | $.813_{\pm.006}$ | $.095_{\pm.002}$ | $.800_{\pm.008}$ | $.107_{\pm.001}$ | $.810_{\pm.004}$ | $.091_{\pm.002}$ |
| | Hidden state | REFIT | 2,138 | $.824_{\pm.005}$ | $.091_{\pm.007}$ | $.825_{\pm.006}$ | $.090_{\pm.003}$ | $.822_{\pm.005}$ | $.098_{\pm.012}$ |
| | | RECALIBRATE | ≈188 | $.739_{\pm.002}$ | $.089_{\pm.005}$ | $.739_{\pm.004}$ | $.088_{\pm.002}$ | $.719_{\pm.009}$ | $.096_{\pm.004}$ |
| | | FROZEN | 0 | $.737_{\pm.004}$ | $.456_{\pm.005}$ | $.740_{\pm.002}$ | $.437_{\pm.005}$ | $.721_{\pm.008}$ | $.506_{\pm.009}$ |
| | | **Re-center (ours)** | 0 | $.737_{\pm.004}$ | $.196_{\pm.001}$ | $.740_{\pm.002}$ | $.194_{\pm.005}$ | $.721_{\pm.008}$ | $.201_{\pm.005}$ |
| Reasoning traces Qwen3-VL, 2 tasks | Probe before the update | | | .829 | .031 | .829 | .031 | .829 | .031 |
| | Output | Verbal confidence | 0 | $.568_{\pm.021}$ | $.105_{\pm.017}$ | $.574_{\pm.018}$ | $.107_{\pm.019}$ | $.525_{\pm.027}$ | $.147_{\pm.013}$ |
| | | P(True) | 0 | $.728_{\pm.035}$ | $.361_{\pm.036}$ | $.722_{\pm.034}$ | $.339_{\pm.025}$ | $.679_{\pm.083}$ | $.405_{\pm.055}$ |
| | | Answer probability | 0 | $.698_{\pm.023}$ | $.089_{\pm.009}$ | $.720_{\pm.027}$ | $.095_{\pm.019}$ | $.682_{\pm.049}$ | $.112_{\pm.004}$ |
| | | Token history ($L$) | 0 | $.668_{\pm.033}$ | $.190_{\pm.017}$ | $.674_{\pm.037}$ | $.154_{\pm.021}$ | $.687_{\pm.017}$ | $.131_{\pm.019}$ |
| | | Self-consistency | 0 | $\mathbf{.827}_{\pm.021}$ | $.091_{\pm.030}$ | $\mathbf{.825}_{\pm.026}$ | $.091_{\pm.032}$ | $\mathbf{.851}_{\pm.018}$ | $.049_{\pm.007}$ |
| | Hidden state | REFIT | 1,024 | $.812_{\pm.008}$ | $.032_{\pm.004}$ | $.814_{\pm.004}$ | $.044_{\pm.008}$ | $.802_{\pm.018}$ | $.042_{\pm.003}$ |
| | | RECALIBRATE | 512 | $.720_{\pm.039}$ | $.029_{\pm.005}$ | $.748_{\pm.012}$ | $.038_{\pm.013}$ | $.763_{\pm.034}$ | $.039_{\pm.013}$ |
| | | FROZEN | 0 | $.731_{\pm.020}$ | $.655_{\pm.068}$ | $.748_{\pm.012}$ | $.486_{\pm.085}$ | $.763_{\pm.034}$ | $.594_{\pm.035}$ |
| | | **Re-center (ours)** | 0 | $.731_{\pm.020}$ | $\mathbf{.040}_{\pm.005}$ | $.748_{\pm.012}$ | $\mathbf{.041}_{\pm.006}$ | $.763_{\pm.034}$ | $\mathbf{.037}_{\pm.003}$ |

**Estimators and baselines.** We compare, on identical rows, verbalized confidence $V$, the number the model writes after its answer, scored as $1 - V$ (Tian et al., 2023; Xiong et al., 2024); P(True), the model's own probability that its answer is correct (Kadavath et al., 2022); single-pass token statistics, among them the answer probability, scored as one minus the probability of the answer's first token (Hendrycks & Gimpel, 2017); a token-history estimator $L$, gradient-boosted trees on 56 entropy, margin, and surprisal statistics of every consumed token; self-consistency $Y_K$, the disagreement of $K = 8$ sampled answers (Wang et al., 2023) ($K \in \{32, 64\}$ on a matched subset); and the hidden-state probe trained on $E$ or on $Y_8$. Short-answer error probes are logistic regressions whose L2 penalty is chosen by image-grouped cross-validation, and the $Y_8$ probe is a ridge regression; long-trajectory probes use ridge scores with a logistic map. The probe is maintained by FROZEN, RECALIBRATE, REFIT, or label-free re-centering (Proposition 4), which uses the mean score on unlabeled inputs of the updated model: half of the test images for short answers, and the map inputs of the 32-label budget (about 15 per cell) for reasoning traces. The calibration budget $n$ is swept over $\{32, 64, 128, 256, 512\}$ labeled groups, split between fitting the map and choosing a decision threshold. Intervals for the $Y_8$ probe are 95% paired bootstrap intervals over image clusters and training seeds.

### 4.2 MAIN RESULTS

We take the conclusions of Section 3 in turn; Table 1 reports every estimator.

**The frozen probe still ranks errors (Conclusion 1).** A probe fitted to the base model keeps ranking the updated models' errors (Figure 2a). On short answers the frozen error probe reaches AUROC 0.721–0.740 after SFT, GRPO, and OPD, against 0.822–0.825 for a probe refit on 2,138 labels, keeping 69–74% of the refit margin above chance; at the end of reasoning traces it reaches 0.731–0.763 against 0.802–0.814 (74–87%). This holds in every model (0.652–0.793; Table 15) and training seed (Figure 6). The loss follows how much an update reorders the score, not how far it moves the weights: over 477 pairs of states, the loss of the $Y_8$ probe rises with rank disagreement (within-model Spearman 0.65) and not with weight distance ($-0.08$, $p = 0.10$); SFT runs with different seeds lie $1.31\times$ the SFT update apart, yet a probe moved between them loses only 0.002 AUROC (Figure 5). Early layers transfer worse than the final ones (Figure 2c).

**Rankings can be learned without correctness labels (Conclusion 2).** A single-pass probe trained only to predict the disagreement of eight sampled answers ranks errors better than those samples:

AUROC 0.776 versus 0.730 on base models (ratio 1.062, 95% interval [1.043, 1.085]) and ratios of 1.020–1.033 after the updates (Table 8). It averages away the noise of a finite bank, which larger banks ($K = 32, 64$) reduce by 0.051 and 0.066 AUROC (Figure 7). Applied frozen, this label-free probe ranks the updated models' errors with AUROC 0.794–0.795 (Table 9).

**Hidden states provide the error signal after reasoning (Conclusion 3).** On ScienceQA, the same seven Qwen3-VL states answering immediately give first-token entropy AUROC 0.877 against 0.883 for the frozen probe; after a reasoning preamble entropy falls to 0.635 while the probe keeps 0.753 (Figure 8). In the full cohorts, the answer probability is the best zero-label monitor for short answers (AUROC 0.845–0.851, ECE 0.086–0.099) but falls below the frozen probe after reasoning (0.682–0.720); verbalized confidence is near chance (0.525–0.574), and P(True) is poorly calibrated (ECE 0.339–0.405). Self-consistency ranks best (0.825–0.851) at the cost of eight extra generations per input.

**The frozen probe's probabilities drift (Conclusion 4).** At the end of reasoning traces the frozen probe predicts a mean error probability of 0.688 where the observed rate is 0.115; its ECE rises from 0.031 to 0.486–0.655, and its Brier score (0.467) is 5.8× that of a constant predictor. On short answers ECE rises from 0.079 to 0.437–0.506. The whole score distribution moves (Figure 4): on the same inputs, scores shift by 3.71 base standard deviations on average (0.77–6.64, positive in every cell) with unchanged spread (ratio 0.99). The weights alone produce this shift: on identical base-generated text the score moves by 1.2–9.5 standard deviations, and 0.82 of the hidden-state change is a translation common to all inputs (Section F). The shift predicts the calibration error (Spearman 0.98) but not the frozen AUROC (−0.14).

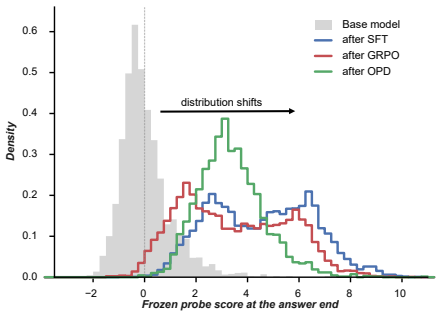



Figure 4: **The frozen score shifts as a whole.** Frozen probe score at the answer end, in base standard deviations, before and after each update (reasoning traces, three seeds).

**A few unlabeled inputs repair the probability scale (Conclusion 5).** Re-centering the frozen map on about 15 unlabeled inputs lowers the ECE of reasoning traces to 0.037–0.041 and the Brier score from 0.467 to 0.093, improving on the frozen probe in all 27 cells (Figure 2b, Table 11), whereas recalibration on 32 labels does worse (Brier 0.115), since small labeled samples often contain no error. With all ≈233 unlabeled inputs, re-centering and re-scaling match recalibration on 512 labels (Brier 0.090 versus 0.091). As Corollary 2 predicts, the excess Brier of re-centering on ≈15 inputs is at most 0.0045, within the bound in 25 of 27 cells (Table 12). On short answers, where post-training lowered the error rate from 0.440 to 0.309, re-centering more than halves the ECE (0.437–0.506 → 0.194–0.201), and the remaining over-prediction tracks the change in error rate (Pearson −0.92; Figure 10), the part that Proposition 3 puts out of reach of unlabeled inputs; an intercept fitted on 32 labels lowers ECE to 0.139 (Table 14). Only refitting improves the ranking (AUROC 0.810 from 128 labels).

### 4.3 Discussion

**Where the ranking fails.** The ranking survives in every model and seed, but not on every task. On SLAKE-zh the updates remove most errors (error rate 0.83 → 0.23), and the frozen probe falls from AUROC 0.932 in-state to 0.463–0.468, below chance, while a refit probe reaches 0.817–0.822 (Table 16): the base probe had learned which questions the base model gets wrong, and those errors are gone. This is the change of error labels that Corollary 1 leaves open; averaged over all cells it is small, as 18% of inputs change error status but move the frozen AUROC by only 0.003 (Table 7).

**Which estimator to use.** For immediate answers, the answer probability is already a calibrated error estimate after post-training. After a reasoning preamble, token statistics and verbalized confidence fail and self-consistency costs eight extra generations; the hidden-state probe is the single-pass alternative, and its scale can be maintained without labels while the error rate stays put. Because no unlabeled statistic reveals a change in the error rate (Proposition 3), a small labeled sample remains the way to verify a re-centered probe.In practice, this establishes a hierarchical maintenance workflow: retain

the base probe across routine updates, re-center its probability scale using unannotated queries, and reserve valuable annotation budgets strictly for auditing error-rate shifts or triggering full probe refitting.

## 5 Related Work

**Uncertainty from sampling and token probabilities.** Semantic entropy and self-consistency quantify uncertainty via agreement across multiple sampled completions (Kuhn et al., 2023; Farquhar et al., 2024; Wang et al., 2023; Manakul et al., 2023), but repeated generation incurs substantial computational overhead. Conversely, single-pass token probabilities provide lightweight risk signals, though their calibration strongly depends on the training objective (Kadavath et al., 2022; Desai & Durrett, 2020; Hendrycks & Gimpel, 2017). Nakkiran et al. (2026) theoretically analyze when sampling-based semantic confidence is calibrated in base models and anticipate that instruction tuning and chain-of-thought degrade it. Our findings provide a token-level counterpart for error ranking, demonstrating that token uncertainty becomes uninformative when reasoning preambles precede the answer, where internal probes remain effective (Conclusion 3).

**Verbalized confidence and calibration.** Rather than extracting probabilities, verbalized confidence prompts the model to self-report confidence in its output (Tian et al., 2023; Xiong et al., 2024). However, models are chronically overconfident, and reinforcement learning with binary rewards further exacerbates this miscalibration (Damani et al., 2025; Leng et al., 2025; Kirichenko et al., 2025). While recent advances like RLCR (Damani et al., 2025) and related techniques (Xu et al., 2024; Stangel et al., 2025) optimize verbal calibration using proper scoring rules, they necessitate fine-tuning the model itself. In contrast, we examine external monitors that remain non-invasive and leave model weights untouched.

**Internal hidden-state probing.** A parallel line of work trains linear probes on internal states to detect errors, hallucinations, and factual truthfulness (Azaria & Mitchell, 2023; Orgad et al., 2025; Burns et al., 2023; Marks & Tegmark, 2024; Kossen et al., 2024). Most relevant to our setting, semantic entropy probes (Kossen et al., 2024) predict sampling dispersion in a single forward pass. However, existing probing literature operates strictly within a static regime, fitting and testing probes on a single checkpoint. We ask a fundamentally new question: whether a probe remains useful when the underlying model evolves through post-training.

**Compatibility and calibration under model updates.** Continual updates can silently break downstream tools and user trust (Bansal et al., 2019), a challenge addressed in embedding spaces via backward-compatible training (Shen et al., 2020). While fine-tuning primarily shifts top-layer representations and reuses core mechanisms (Merchant et al., 2020; Prakash et al., 2024; Jain et al., 2024), post-training notoriously impairs model calibration (OpenAI, 2023; Xiao et al., 2025; Kapoor et al., 2024). This aligns with findings that calibration degrades under distribution shift even when ranking survives (Ovadia et al., 2019; Guo et al., 2017). Although post-hoc transformations (Platt, 1999; Zadrozny & Elkan, 2002) and label-free adjustments (Saerens et al., 2002; Lipton et al., 2018; Li et al., 2016) are widely used to realign probabilities, their applicability to probe transfer across post-training remains unexplored. We show that internal error rankings remain compatible across updates, while probability scale drift can be repaired using only unlabeled scores.

## 6 Conclusion

We asked what part of a hidden-state error estimator survives SFT, GRPO, and OPD. The order in which it ranks inputs by risk survives; the scale on which it reports probabilities does not. The ranking–calibration decomposition of proper scores explains the split and predicted three further regularities that we confirmed: because the drift is largely a shift, re-centering on a handful of unlabeled inputs removes it, and labels are needed only for the part of the change that moves the error rate itself; the ranking can be learned from self-consistency without labels; and token-level uncertainty suffices only when the answer is committed in the first tokens. For practitioners this gives a maintenance rule: keep the probe, re-center it on unlabeled inputs, verify it on a small labeled sample, and refit only when the ranking must improve (Section G.2).

### Reproducibility statement

Probe definitions and maintenance policies are specified in Sections 2 and 3, with proofs in Section A. Models, datasets, training hyperparameters, data splits, prompts, and evaluation protocols are described in Sections B.1, B.2 and 4.1. Detailed results by model and task and analyses of calibration budgets are provided in Sections G, G.1 and H.

### Ethics statement

We use public question-answering datasets and collect no human-subject data. Several tasks are medical; our error estimates are measurements of model behavior, not clinical validation, and a well-ranked monitor can still assign low risk to a wrong answer.

### AI use statement

Generative AI tools were used to assist with drafting and revising the manuscript, including improving its organization, clarity, and language. The authors take full responsibility for the paper's content, including its scientific claims, results, and references.

## References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *Empirical Methods in Natural Language Processing Findings*, 2023. URL https://arxiv.org/abs/2304.13734.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2429–2437, 2019.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*, 2025.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Empirical Methods in Natural Language Processing*, pp. 295–302, 2020.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.08500.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Gurari_VizWiz_Grand_Challenge_CVPR_2018_paper.html.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. URL https://arxiv.org/abs/2003.10286.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. URL https://doi.org/10.48550/arXiv.2207.05221.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew G Wilson. Large language models must be taught to know what they don't know. In *Advances in Neural Information Processing Systems*, 2024.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. Abstentionbench: Reasoning LLMs fail on unanswerable questions. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=kYbojsAOBj.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5: 180251, 2018. doi: 10.1038/sdata.2018.251. URL https://www.nature.com/articles/sdata2018251.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0tg0jzsdL.

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, 2018.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *IEEE International Symposium on Biomedical Imaging*, 2021. URL https://arxiv.org/abs/2102.09542.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 33–44, 2020.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Preetum Nakkiran, Arwen Bradley, Adam Goliński, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. Trained on tokens, calibrated on concepts: The emergence of semantic calibration in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/39b228b1a83232b14a1c81b0ed5a7a05-Abstract-Conference.html.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024.

Qwen Team. Qwen3.5-4B and Qwen3.5-9B: Post-trained model cards. Hugging Face model repositories, 2026. URL https://huggingface.co/Qwen/Qwen3.5-4B. Companion card: https://huggingface.co/Qwen/Qwen3.5-9B. Accessed 2026-09-16.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv:2206.01718*, 2022. URL https://arxiv.org/abs/2206.01718.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Paul Stangel, David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Kamilia Zaripova, Matthias Keicher, and Nassir Navab. Rewarding doubt: A reinforcement learning approach to confidence calibration of large language models. *CoRR*, abs/2503.02623, March 2025. URL https://doi.org/10.48550/arXiv.2503.02623.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=g3faCfrwm7.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J Su, and Li Shen. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *International Conference on Machine Learning*, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. Sayself: Teaching llms to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5985–5998, 2024.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

# APPENDIX

Table 2: **Where each conclusion of Section 3 is tested.**

| Conclusion | Tables | Figures |
|---|---|---|
| 1 The frozen probe still ranks errors | 15, 16, 17, 18, 6, 7 | 5, 6 |
| 2 Rankings can be learned without correctness labels | 8, 9 | 7 |
| 3 Hidden states provide the error signal after reasoning | – | 8 |
| 4 The frozen probe's probabilities drift | 10 | 9, 10, 11 |
| 5 A few unlabeled inputs repair the probability scale | 11, 12, 13, 14 | 12 |

## A PROOFS

**Proposition 1.** AUROC equals $\Pr(s(X_1) > s(X_0)) + \frac{1}{2}\Pr(s(X_1) = s(X_0))$ for independent draws $X_1$ from the erroneous and $X_0$ from the correct inputs. A strictly increasing $\phi$ preserves every strict inequality and every tie, so the AUROC of $\phi \circ s$ equals that of $s$. AURC and the risk at a fixed accepted coverage depend on the score only through the order in which inputs are accepted, which $\phi$ preserves. For the Brier score, condition on $s$: $\mathbb{E}[(\phi(s) - E)^2 \mid s] = (\phi(s) - \pi_s(s))^2 + \pi_s(s)(1 - \pi_s(s))$, because $\mathbb{E}[E \mid s] = \pi_s(s)$. Taking expectations, $\mathrm{Brier}(\phi(s)) = \mathbb{E}[(\phi(s) - \pi_s(s))^2] + \mathbb{E}[\pi_s(s)(1 - \pi_s(s))]$; the second term does not depend on $\phi$, so the minimum is attained by $\phi = \pi_s$ and the excess is the first term. Applied to $q = \phi(s)$ this is the decomposition (1).

**Proposition 2.** For a fixed input $x$, let $f(t) = s_{\theta_0 + tUu}(x)$ for $t \in [0, 1]$. Taylor's theorem with the Lipschitz condition on the gradient gives $|f(1) - f(0) - f'(0)| \le \frac{\Lambda}{2}\|Uu\|_2^2 = \frac{\Lambda}{2}\|u\|_2^2$, and $f'(0) = \nabla_\theta s_{\theta_0}(x)^\top Uu$. By Cauchy–Schwarz, $|f'(0)| \le \|U^\top \nabla_\theta s_{\theta_0}(x)\|_2 \|u\|_2$. Taking the $L_2(P)$ norm over $x$ and applying Minkowski's inequality yields (2). The bound concerns a fixed input distribution; responses generated by the updated model change the prefix itself, which is one reason we probe hidden states at several positions.

**Corollary 1.** Let $D = s_{\theta_0}(X_1) - s_{\theta_0}(X_0)$ and let $G$ be the event that both $|s_{\theta_1}(X_e) - s_{\theta_0}(X_e)| \le \delta$. On $G \cap \{D > 2\delta\}$ the updated score still orders the pair correctly. Hence $\mathrm{AUROC}(s_{\theta_1}) \ge \Pr(D > 2\delta) - \Pr(G^c) \ge \Pr(D > 0) - \Pr(0 < D \le 2\delta) - \eta_1 - \eta_0$, and $\Pr(D > 0) \ge \mathrm{AUROC}(s_{\theta_0}) - \frac{1}{2}\Pr(D = 0)$; for continuous scores the tie term vanishes. By Chebyshev's inequality, $\eta_e \le \|s_{\theta_1} - s_{\theta_0}\|_{L_2(P_e)}^2 / \delta^2$, which links the corollary to Proposition 2.

**Proposition 3.** Given $\mu$ and $\pi_k$, draw $s \sim \mu$ and then $E \mid s \sim \mathrm{Bernoulli}(\pi_k(s))$. Both joint laws have score marginal $\mu$ and calibration functions $\pi_1$ and $\pi_2$. A map $\hat{\phi}$ computed from unlabeled scores is a function of $\mu$ alone, so it is the same map under both laws. If $\pi_1 \ne \pi_2$ on a set of positive $\mu$-measure, $\hat{\phi}$ differs from at least one of them on that set, and by Proposition 1 it incurs positive excess Brier score under that law.

**Proposition 4.** Under the shift, $\mathbb{E}[s_{\theta_1}(X)] - \mathbb{E}[s_{\theta_0}(X)] = \Delta$ for any input distribution, so $\hat{\Delta}$ computed from unlabeled scores of the two models on the same inputs is consistent. For the re-centered score $\tilde{s} = s_{\theta_1} - \Delta = s_{\theta_0}$, $\Pr_1(E = 1 \mid \tilde{s} = t) = \pi_1(t + \Delta) = \pi_0(t)$, so $\pi_0(\tilde{s})$ is calibrated for the updated model; continuity of $\pi_0$ gives calibration in the limit $\hat{\Delta} \to \Delta$. Re-centering is an increasing map, so AUROC is unchanged (Proposition 1). In our implementation the frozen map standardizes the score with the base model's mean and scale; re-centering replaces the mean by the mean of the base probe's score on unlabeled calibration-role inputs of the updated model, and the re-scaled variant also replaces the scale.

**Corollary 2.** The logistic function has slope at most $1/4$, so $c \mapsto \sigma(a(s - c)/\sigma_0 + b)$ is $a/(4\sigma_0)$-Lipschitz for every $s$. As in Proposition 4, take the frozen map to be the base calibration function in its own coordinates. Under the shift, $\pi_1(s) = \pi_0(s - \Delta)$ is then the frozen map with centering $\hat{\mu}_0 + \Delta$, while re-centering uses $\hat{\mu}_1$. The re-centered map $q$ is increasing in $s$, so its calibration function is $\pi_1$ and its calibration term is $\mathbb{E}_s[(q(s) - \pi_1(s))^2] \le \frac{a^2}{16\sigma_0^2}(\hat{\mu}_1 - \hat{\mu}_0 - \Delta)^2$. With $m_j$ the population mean

score of model $j$, so that $m_1 - m_0 = \Delta$, we have $\hat{\mu}_1 - \hat{\mu}_0 - \Delta = (\hat{\mu}_1 - m_1) - (\hat{\mu}_0 - m_0)$, a difference of independent sample-mean errors with variances $\sigma_1^2/n$ and $\sigma_0^2/n_0$; taking expectations gives the bound, which contains no error rate. Finally, $m$ independent labeled examples at error rate $p$ contain no error with probability $(1-p)^m$. The likelihood of a calibration map is then increased by lowering every predicted error probability, so no maximum-likelihood map exists; our implementation falls back to a constant map. Table 12 evaluates the bound on the long-trajectory cells.

**Supporting identities.** For a probe trained on the proxy $Y$ rather than on $E$, writing $m_Z(H) = \mathbb{E}[Z \mid H]$,

$$\mathbb{E}\big[(E - m_Y(H))^2\big] = \mathbb{E}\big[\operatorname{Var}(E \mid H)\big] + \mathbb{E}\big[(m_E(H) - m_Y(H))^2\big], \tag{3}$$

the conditional bias–variance identity for the predictor $m_Y(H)$ of the binary variable $E$; label-free supervision loses only the second term, the *proxy gap*, which is small when consistently wrong answers are rare. For the denoising statement, write $Y_K = Y_\infty + \varepsilon_K$ with $\mathbb{E}[\varepsilon_K \mid x] = 0$. Then $m_{Y_K}(H) = \mathbb{E}[Y_\infty \mid H]$, so $\mathbb{E}[(Y_\infty - m_{Y_K}(H))^2] = \mathbb{E}[\operatorname{Var}(Y_\infty \mid H)]$, whereas $\mathbb{E}[(Y_\infty - Y_K)^2] = \mathbb{E}[\operatorname{Var}(\varepsilon_K \mid x)]$, which for the plug-in disagreement scales as $1/K$. The first quantity is smaller whenever $H$ explains most of the variation of $Y_\infty$ across inputs, which is the regime measured in Figure 7. The oracle version of the incremental value of $H$ over token statistics, $\mathrm{Brier}(m_E(L)) - \mathrm{Brier}(m_E(L,H)) = \mathbb{E}[(m_E(L,H) - m_E(L))^2] \geq 0$, is nonnegative, but finite fitted estimators need not realize it (Section E).

# B Setup

## B.1 Post-training operators

All updated states are rank-16 LoRA adapters ($\alpha = 32$) on the frozen base model, trained on image-disjoint splits; POPE and VizWiz are evaluation-only. GRPO and OPD continue the SFT adapter: their increments beyond it are 5.6–8.8% of the SFT update (mean 7.0%, 72 states), and within a model the update norm varies by at most 1.1% across the 87 post-trained states. Update size is not associated with the refit-minus-frozen AUROC gap (within-model Spearman $-0.19$, permutation $p = 0.09$).

Table 3: **Post-training recipes** (Qwen3-VL-2B shown; other models use the same recipe; three seeds each).

| Operator | Recipe |
|---|---|
| SFT | Target-token cross-entropy on the gold answer; one epoch over 5,629 examples; learning rate $10^{-4}$; gradient accumulation 8 (704 steps). |
| GRPO | From SFT; 512 prompts, groups of 4 responses (at most 32 new tokens), 2 iterations, learning rate $5 \cdot 10^{-6}$, clip $\epsilon = 0.2$, exact $\mathrm{KL}(\pi_\theta \,\|\, \pi_{\text{SFT}})$ with $\beta = 0.04$, binary correctness reward; 256 steps. Short answers add archived-rollout and base-anchored variants. |
| OPD | From SFT; teacher: SFT state of the larger model of the family (Qwen3-VL-8B for 2B, Qwen3.5-9B for 4B); 512 prompts, 4 student samples, 2 iterations, learning rate $5 \cdot 10^{-6}$; full-vocabulary $\mathrm{KL}(q_{\text{teacher}} \,\|\, p_\theta)$ on the student's own tokens. Short answers add an archived-rollout variant. |

### B.2 MEASUREMENT PROTOCOLS

Table 4: **The two protocols.**

| | **Short answers** | **Reasoning traces** |
|---|---|---|
| Models | Qwen3-VL-2B/8B, InternVL3.5-8B, Qwen3.5-4B/9B | Qwen3-VL-2B/8B |
| Tasks | SLAKE-en/zh, VQA-RAD, PathVQA, ScienceQA, A-OKVQA, POPE, VizWiz | ScienceQA, GSM8K |
| Updates | SFT; GRPO (4 variants); OPD (2 variants; 2B and 4B only) | SFT, GRPO; OPD (2B only) |
| Decoding | assistant turn prefilled with `<answer>`, at most 16 tokens | greedy, at most 2,048 tokens; reasoning, then `<answer>` and `<conf>` |
| Hidden state | final layer, last prompt token | final layer, Prefix-32 and answer end (before `<conf>`) |
| Probe | logistic regression; L2 penalty by image-grouped 5-fold CV (Brier) | ridge on $E$ ($\lambda \in \{0.01, 0.1, 1, 10\}$, grouped 3-fold CV), then a logistic map |
| Labels | REFIT 2,138, RECALIBRATE $\approx$ 188 per cell | REFIT 1,024, RECALIBRATE 512 per cell |
| Re-center | half A of the test images, labels unused | map inputs of the 32-label budget, labels unused |
| Scoring | half B of the test images | evaluation role; gradable answers only |
| Self-cons. | $K$=8 samples at temperature 1 | $K$=8 samples |
| Metrics | AUROC; ECE with ten equal-width bins; Brier. Mean$_{\pm \text{s.d.}}$ over three training seeds. | |

**Prompts.**

**Short answers: open-ended prompt**

`<image>`
Answer the medical visual question concisely.
If the question is yes/no, answer exactly yes or no.
Question: *{question}*
Put your final answer inside
`<answer></answer>` tags.

**Short answers: multiple-choice prompt**

`<image>`
Answer the multiple-choice question about the image.
Choose exactly one option and reply with its letter only.
Question: *{question}*
*A. {option} . . . one line per option*
Put your final answer inside
`<answer></answer>` tags.

**Reasoning traces: prompt**

Question:
*{question}*
Options: (multiple choice only)
*A. {option} . . . one line per option*

Output instructions:
Explain how you arrive at your answer in ordinary text. A brief explanation is enough. Do not put the explanation inside tags.
After the explanation, write exactly one `<answer>...</answer>` block containing only your final answer.
Then write exactly one `<conf>...</conf>` block containing only a finite number from 0 to 1 for the probability that your final answer is correct. Do not write a percentage.
Close both blocks. Do not nest or repeat the answer or conf tags, and do not write text after the conf block.
Put exactly the selected option letter in the answer block. (multiple choice only)

**Reasoning traces: P(True) baseline prompt**

Question:
*{question}*
Options: (multiple choice only)
*A. {option} . . . one line per option*
Proposed answer: *{final answer of the trace}*
Is the proposed answer correct? Reply with exactly one word: True or False.
Score, from the next-token distribution: $1 - P(\text{True})$, with $P(\text{True}) = p(\texttt{True})/(p(\texttt{True}) + p(\texttt{False}))$.

Rows without a gradable answer or valid features are excluded from the metrics (91.2% of evaluation rows remain at the answer end, seed 0; lowest for GRPO on Qwen3-VL-8B ScienceQA, 63%).

Table 5: **Availability of long-trajectory landmarks** (seed 0). Errors are labeled only at the final answer.

| Role | Rows | Prefix-0 | Prefix-32 | Prefix-64 | Prefix-128 | Prefix-256 | Prefix-512 |
|---|---|---|---|---|---|---|---|
| Fit | 7,168 | 100.0% | 97.8% | 78.3% | 34.2% | 6.7% | 3.6% |
| Calibration | 7,168 | 100.0% | 97.7% | 78.8% | 33.7% | 6.4% | 3.5% |
| Evaluation | 7,168 | 100.0% | 98.0% | 79.6% | 35.3% | 7.3% | 4.2% |

## C THE FROZEN PROBE STILL RANKS ERRORS

**Conclusion 1.** The frozen probe's ranking loss follows how much the update reorders its score, not how far the weights move (Figure 5, Tables 6 and 7); its AUROC stays well above chance for every training seed and model (Figure 6). The exception is SLAKE-zh, where the update removes most errors (Table 16). **Depth.** Error information rises with depth in every state, and the frozen probe transfers best at the final layers (1,200 cell–layer rows).

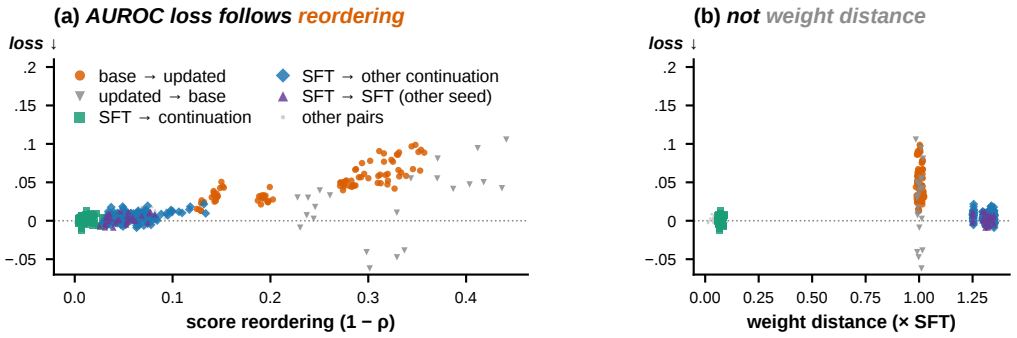



Figure 5: **The ranking loss follows score reordering, not weight distance** (short answers; one point per pair of fitting and application states, averaged over tasks).

Table 6: **Score drift and ranking loss by pair type** (short answers). Distance: relative to the SFT update; shift and residual: in s.d. of the fitted state's score; loss: refit minus frozen AUROC.

| Fitted → applied | Pairs | Weight dist. | Shift | Spread ratio | Residual | Rank corr. | Frozen | Refit | Ranking loss |
|---|---|---|---|---|---|---|---|---|---|
| Base → updated | 87 | 1.00 | +1.77 | 0.63 | 0.63 | 0.75 | 0.795 | 0.844 | +0.049 |
| Updated → base | 22 | 1.00 | +0.23 | 0.99 | 0.69 | 0.68 | 0.753 | 0.778 | +0.025 |
| SFT → its GRPO/OPD continuation | 72 | 0.07 | −0.17 | 1.04 | 0.11 | 0.99 | 0.843 | 0.844 | +0.001 |
| SFT → another seed's continuation | 144 | 1.31 | −0.10 | 1.03 | 0.25 | 0.94 | 0.840 | 0.844 | +0.004 |
| SFT → SFT of another seed | 30 | 1.31 | +0.07 | 0.99 | 0.22 | 0.95 | 0.844 | 0.846 | +0.002 |
| Other updated pairs | 122 | 0.95 | +0.07 | 0.98 | 0.21 | 0.95 | 0.839 | 0.843 | +0.005 |

Table 7: **Drift versus turnover** (short answers, base probe). Turnover: $A(s_0; E_1) - A(s_0; E_0)$; drift: $A(s_1; E_1) - A(s_0; E_1)$.

| Update | Cells | In-state $A(s_0; E_0)$ | Base scores, new labels $A(s_0; E_1)$ | Frozen $A(s_1; E_1)$ | Refit | Turnover effect | Inputs whose error flips |
|---|---|---|---|---|---|---|---|
| SFT | 120 | 0.776 | 0.780 | 0.794 | 0.846 | +0.004 | 18% |
| GRPO | 480 | 0.776 | 0.780 | 0.795 | 0.844 | +0.003 | 18% |
| OPD | 96 | 0.795 | 0.788 | 0.795 | 0.844 | −0.007 | 18% |
| All | 696 | 0.776 | 0.780 | 0.795 | 0.845 | +0.003 | 18% |

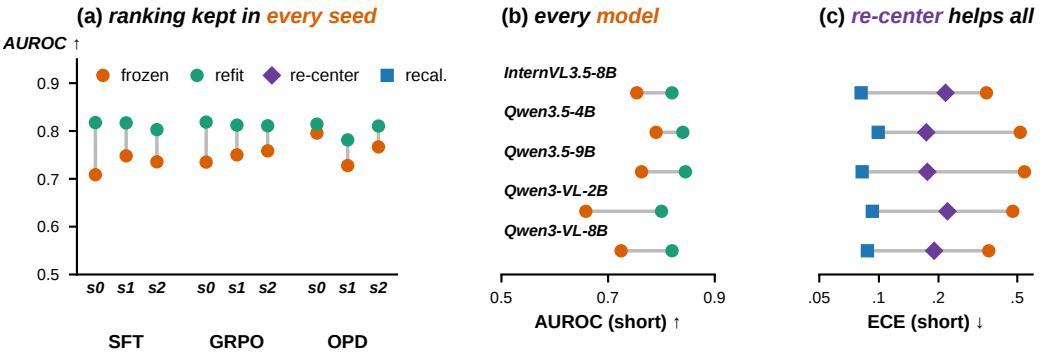



Figure 6: **Robustness.** (a) Frozen and refit AUROC for every training seed (reasoning traces); (b, c) every model (short answers).

## D RANKINGS CAN BE LEARNED WITHOUT CORRECTNESS LABELS

**Conclusion 2.** A probe trained on the eight-sample disagreement, without correctness labels, ranks errors slightly better than the samples it imitates (Figure 7, Tables 8 and 9).

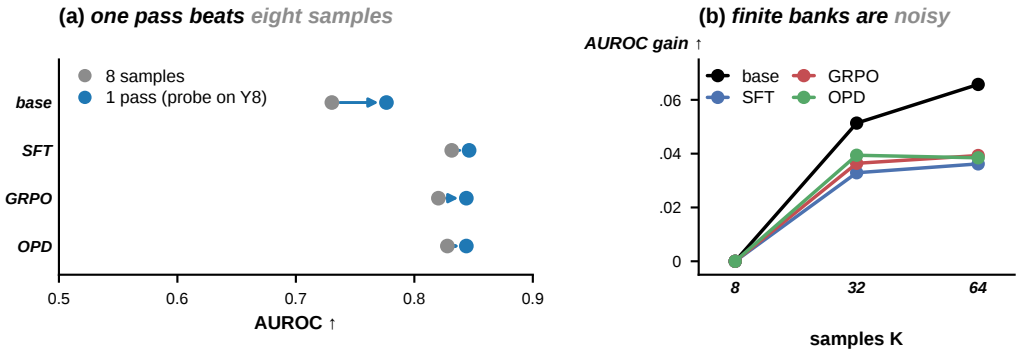



Figure 7: **One forward pass versus eight samples** (short answers). (a) Error AUROC of $Y_8$ and of a probe trained on $Y_8$ without labels; (b) gain of larger banks over $K=8$.

Table 8: **A probe trained on self-consistency versus the self-consistency it imitates** (short answers; 95% paired bootstrap intervals). Ratio: probe AUROC / bank AUROC.

| State | Cells | $Y_8$ AUROC | Probe AUROC | Ratio | Pearson$(\hat{Y}, Y_8)$ |
|---|---|---|---|---|---|
| Base | 40 | 0.730 [0.713,0.747] | 0.776 [0.760,0.793] | 1.062 [1.043,1.085] | 0.677 |
| SFT | 120 | 0.832 [0.815,0.846] | 0.846 [0.832,0.860] | 1.020 [1.007,1.037] | 0.783 |
| GRPO | 480 | 0.820 [0.804,0.836] | 0.844 [0.828,0.858] | 1.033 [1.020,1.051] | 0.777 |
| OPD | 96 | 0.828 [0.811,0.844] | 0.844 [0.828,0.859] | 1.022 [1.006,1.042] | 0.789 |

Table 9: **The probe trained on $Y_8$, applied frozen to updated models** (short answers; 95% image-cluster bootstrap intervals). Held-out: six held-out task cohorts.

| | Error AUROC ↑ | | | MSE to $Y_8$ ↓ | | | Held-out |
|---|---|---|---|---|---|---|---|
| **Update** | FROZEN | RECALIBRATE | REFIT | FROZEN | RECALIBRATE | REFIT | FROZEN AUROC |
| None (in-state) | 0.776 [0.760,0.793] | | | 0.026 | | | – |
| SFT | 0.794 [0.777,0.810] | 0.794 | 0.846 | 0.110 | 0.056 | 0.023 | 0.775 |
| GRPO | 0.795 [0.778,0.810] | 0.794 | 0.844 | 0.109 | 0.052 | 0.023 | 0.774 |
| OPD | 0.795 [0.773,0.815] | 0.794 | 0.844 | 0.101 | 0.051 | 0.024 | 0.797 |

**Target choice.** On a matched cohort, the $E$-trained probe ranks better in-state ($Y-E$ AUROC $-0.081$, $[-0.098,-0.064]$), while the $Y$-trained probe transfers better after the update (+0.051, [+0.035,+0.067]; reversed for InternVL3.5-8B).

## E HIDDEN STATES PROVIDE THE ERROR SIGNAL AFTER REASONING

**Conclusion 3.** First-token uncertainty is a strong error signal when the answer begins immediately and a weak one after a reasoning preamble (Figure 8).

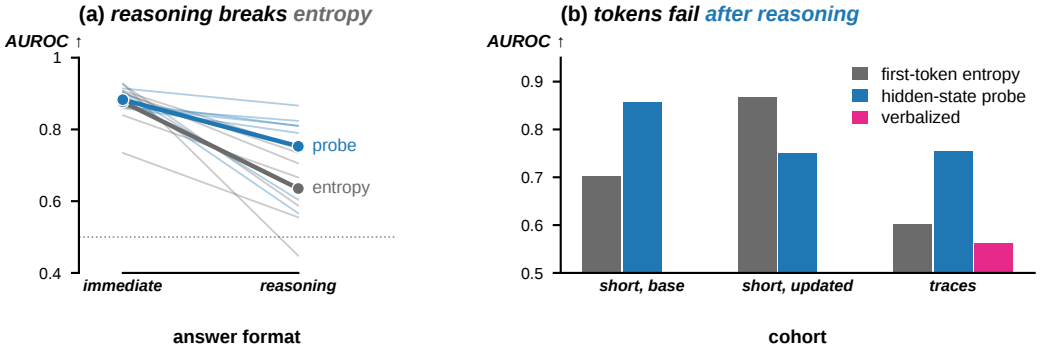



Figure 8: **Token uncertainty depends on the answer format.** (a) ScienceQA, seven Qwen3-VL states answering immediately or after reasoning (thin lines: states); (b) all cohorts.

**Answer length.** Within short-answer cells (67 cells), first-token entropy ranks one-token and multi-token answers equally well (AUROC 0.805 and 0.806): what separates the regimes is a reasoning preamble, not answer length. **Fusion.** Adding the hidden score to $L$ changes Brier by $\Delta B \in [-0.0013, +0.0036]$ (reasoning traces, seed 0); a frozen short-answer fusion changes AUROC by $-0.024$ $[-0.035,-0.013]$. These are negative results for the tested fusion rules.

## F THE FROZEN PROBE'S PROBABILITIES DRIFT

**Conclusion 4.** After every update the frozen probe over-predicts error because the whole score distribution moves (Figures 9 and 10, Table 10); Figure 11 places every estimator on the ranking–calibration plane.

## G A FEW UNLABELED INPUTS REPAIR THE PROBABILITY SCALE

**Conclusion 5.** On reasoning traces, re-centering on a few unlabeled inputs calibrates as well as labeled recalibration (Figure 12, Tables 11 and 12); on short answers, where the error rate changed, a few labels for an intercept are still needed (Table 14). Only refitting improves the ranking and selective risk (Table 13).

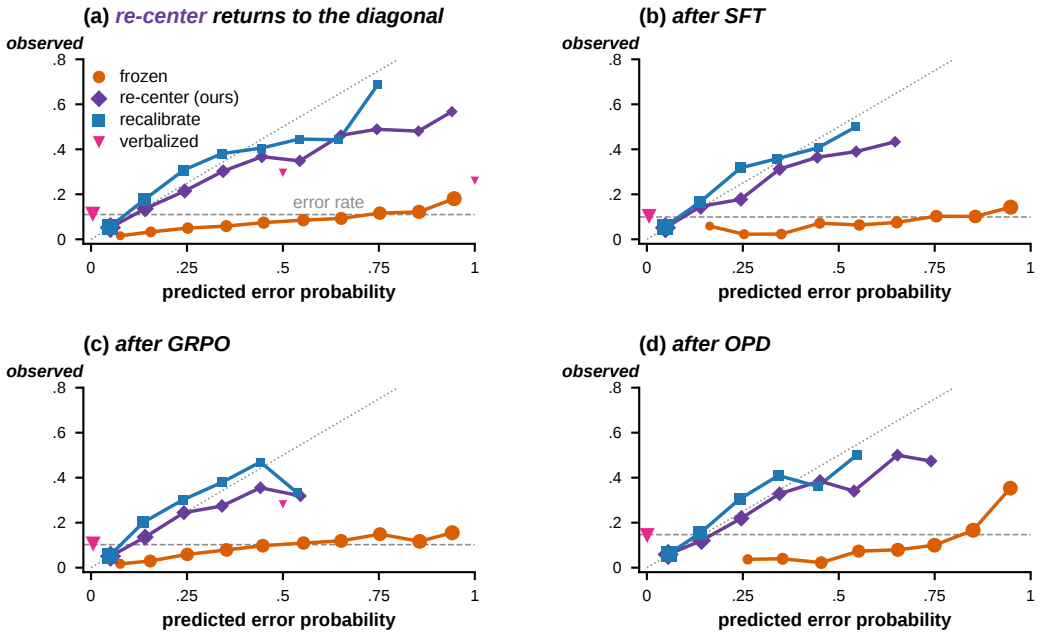



Figure 9: **Reliability on reasoning traces** (answer end, three seeds). Marker area: bin size; dashed: observed error rate; dotted: diagonal.

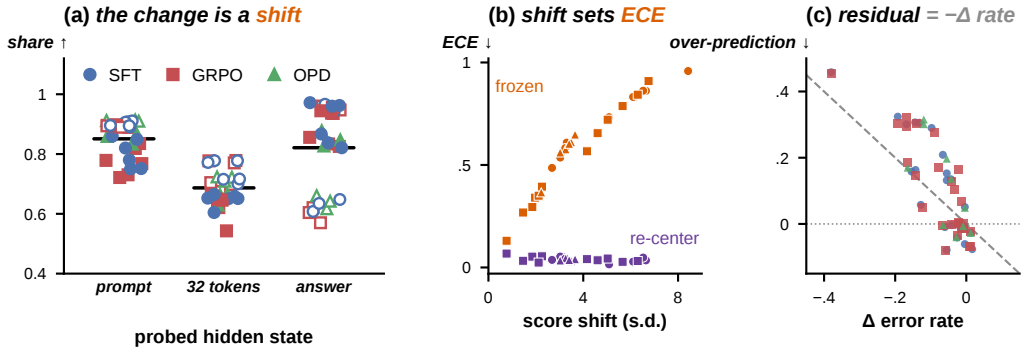



Figure 10: **The update shifts the score.** (a) Share of the hidden-state change that is a common translation (three seeds; filled: ScienceQA, open: GSM8K; bar: mean). (b) Frozen and re-centered ECE against the paired shift. (c) Short answers: over-prediction after re-centering against the change in error rate (dashed: $y = -x$).

### G.1 Repair hierarchy on short answers

Parameters are added one at a time to the re-centered logit; the slope is fitted with a ridge penalty ($10^{-3}$) toward one.

### G.2 A maintenance recipe

1. **Keep** the base probe, normalization, and map.
2. **Re-center** the map on unlabeled inputs of the updated model (Corollary 2; Table 12).
3. **Verify** on a small labeled sample; if the error rate moved, fit an intercept, and with 32 or more labels a slope (Table 14). Never fit a map on a sample without errors.
4. **Refit** when the ranking itself must improve (Tables 11 and 13).

Table 10: **Paired score shift per cell** (reasoning traces, answer end unless noted; base s.d.; three seeds). Own text: hidden states on each model's own trajectories; base text: on the base model's trajectories (weights only).

| | | | Own text | | Base text | | Prefix-32 | Slope |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Task** | **Update** | Shift | Scale ratio | Shift | Drift s.d. | Shift | $w_1/w_0$ |
| 2B | GSM8K | SFT | $+2.95_{\pm 0.24}$ | $1.21_{\pm 0.05}$ | $+3.26_{\pm 0.27}$ | $1.21_{\pm 0.19}$ | $+0.90_{\pm 0.10}$ | $+0.82_{\pm 0.19}$ |
| | | GRPO | $+1.76_{\pm 0.26}$ | $1.06_{\pm 0.03}$ | $+2.10_{\pm 0.24}$ | $1.15_{\pm 0.26}$ | $+0.72_{\pm 0.14}$ | $+0.76_{\pm 0.19}$ |
| | | OPD | $+3.26_{\pm 0.23}$ | $1.24_{\pm 0.05}$ | $+3.58_{\pm 0.24}$ | $1.23_{\pm 0.20}$ | $+0.95_{\pm 0.11}$ | $+0.76_{\pm 0.10}$ |
| 2B | ScienceQA | SFT | $+2.81_{\pm 0.81}$ | $1.05_{\pm 0.06}$ | $+3.13_{\pm 0.84}$ | $0.64_{\pm 0.15}$ | $+0.04_{\pm 0.07}$ | $+1.17_{\pm 0.18}$ |
| | | GRPO | $+1.71_{\pm 0.82}$ | $0.93_{\pm 0.04}$ | $+2.06_{\pm 0.79}$ | $0.55_{\pm 0.10}$ | $+0.03_{\pm 0.09}$ | $+1.12_{\pm 0.16}$ |
| | | OPD | $+3.10_{\pm 0.76}$ | $1.02_{\pm 0.07}$ | $+3.40_{\pm 0.80}$ | $0.62_{\pm 0.16}$ | $+0.07_{\pm 0.12}$ | $+1.14_{\pm 0.26}$ |
| 8B | GSM8K | SFT | $+5.97_{\pm 0.80}$ | $0.61_{\pm 0.02}$ | $+5.91_{\pm 0.86}$ | $1.16_{\pm 0.04}$ | $+1.36_{\pm 0.06}$ | $+0.96_{\pm 0.17}$ |
| | | GRPO | $+5.52_{\pm 0.85}$ | $0.65_{\pm 0.03}$ | $+5.42_{\pm 0.91}$ | $1.18_{\pm 0.06}$ | $+1.30_{\pm 0.07}$ | $+1.08_{\pm 0.22}$ |
| 8B | ScienceQA | SFT | $+7.02_{\pm 1.24}$ | $1.13_{\pm 0.14}$ | $+8.66_{\pm 0.90}$ | $0.76_{\pm 0.01}$ | $+0.31_{\pm 0.15}$ | $-0.03_{\pm 0.49}$ |
| | | GRPO | $+5.31_{\pm 1.32}$ | $1.39_{\pm 0.36}$ | $+7.58_{\pm 1.28}$ | $0.90_{\pm 0.08}$ | $+0.29_{\pm 0.15}$ | $+0.49_{\pm 0.43}$ |

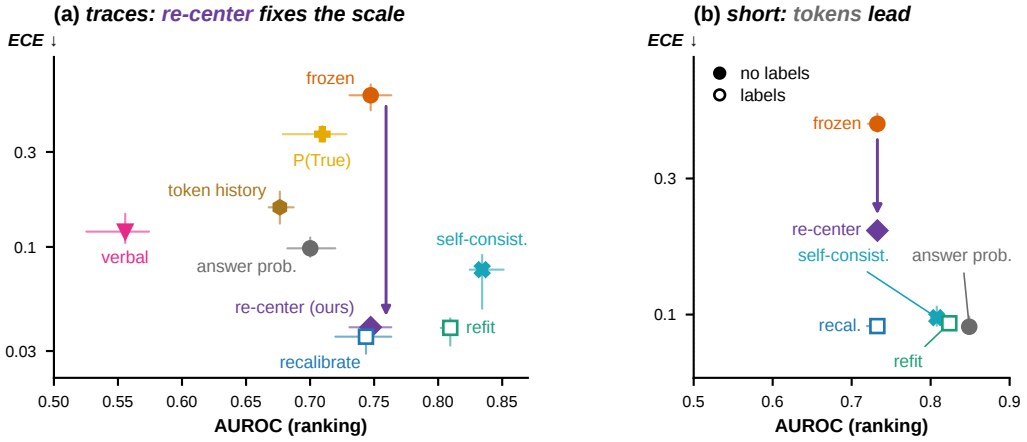



Figure 11: **Ranking versus calibration for every estimator of Table 1.** Lines: range over SFT, GRPO, and OPD; open markers need labels; arrow: re-centering.

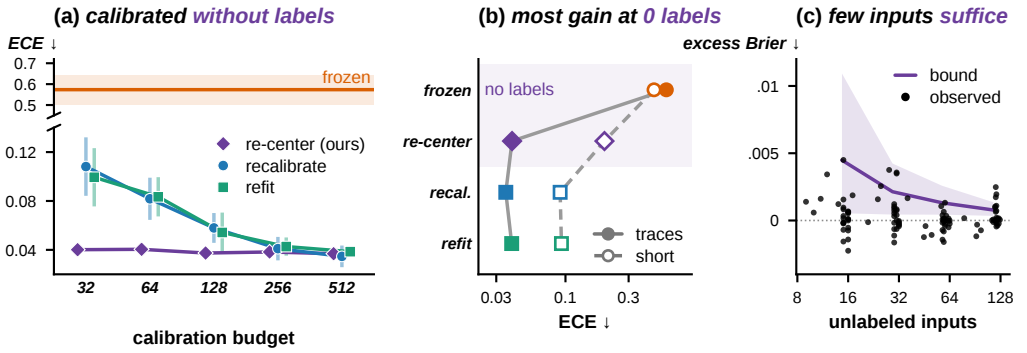



Figure 12: **Repair is cheap.** (a) ECE by labeled budget (reasoning traces; band: frozen ± s.d.); (b) ECE of each policy on both protocols; (c) excess Brier of re-centering on $n$ unlabeled inputs and the bound of Corollary 2 (band: range over cells).

5. **Choose the signal by answer format**: token uncertainty for immediate answers, the hidden state after reasoning (Figure 8).

Table 11: **Calibration budget on reasoning traces.** 27 post-training cells at the answer end; mean$_{\pm \text{s.d.}}$ over three training seeds. Bold: best policy at each budget.

| | | Labels | Labeled calibration budget $n$ | | | | |
|---|---|---|---|---|---|---|---|
| | **Policy** | **needed** | 32 | 64 | 128 | 256 | 512 |
| ECE↓ | FROZEN | 0 | $.574_{\pm.071}$ | $.574_{\pm.071}$ | $.574_{\pm.071}$ | $.574_{\pm.071}$ | $.574_{\pm.071}$ |
| | RECALIBRATE | $n$ | $.108_{\pm.024}$ | $.082_{\pm.017}$ | $.058_{\pm.012}$ | $.041_{\pm.010}$ | $.035_{\pm.009}$ |
| | REFIT | $512{+}n$ | $.099_{\pm.024}$ | $.083_{\pm.016}$ | $.054_{\pm.016}$ | $.043_{\pm.007}$ | $.039_{\pm.002}$ |
| | Re-center + re-scale | 0 | $.044_{\pm.010}$ | $.042_{\pm.005}$ | $\mathbf{.033}_{\pm.003}$ | $\mathbf{.031}_{\pm.001}$ | $\mathbf{.032}_{\pm.001}$ |
| | **Re-center (ours)** | 0 | $\mathbf{.040}_{\pm.002}$ | $\mathbf{.040}_{\pm.004}$ | $.037_{\pm.004}$ | $.038_{\pm.003}$ | $.037_{\pm.001}$ |
| Brier↓ | FROZEN | 0 | $.467_{\pm.072}$ | $.467_{\pm.072}$ | $.467_{\pm.072}$ | $.467_{\pm.072}$ | $.467_{\pm.072}$ |
| | RECALIBRATE | $n$ | $.115_{\pm.014}$ | $.102_{\pm.014}$ | $.095_{\pm.014}$ | $.091_{\pm.012}$ | $.091_{\pm.012}$ |
| | REFIT | $512{+}n$ | $.110_{\pm.016}$ | $.101_{\pm.013}$ | $.091_{\pm.015}$ | $\mathbf{.087}_{\pm.012}$ | $\mathbf{.087}_{\pm.012}$ |
| | Re-center + re-scale | 0 | $.094_{\pm.008}$ | $\mathbf{.093}_{\pm.010}$ | $\mathbf{.090}_{\pm.012}$ | $.090_{\pm.012}$ | $.090_{\pm.011}$ |
| | **Re-center (ours)** | 0 | $\mathbf{.093}_{\pm.012}$ | $.094_{\pm.012}$ | $.093_{\pm.012}$ | $.093_{\pm.012}$ | $.093_{\pm.012}$ |
| AUROC↑ | FROZEN | 0 | $.743_{\pm.006}$ | $.743_{\pm.006}$ | $.743_{\pm.006}$ | $.743_{\pm.006}$ | $.743_{\pm.006}$ |
| | RECALIBRATE | $n$ | $.606_{\pm.060}$ | $.656_{\pm.032}$ | $.732_{\pm.026}$ | $.739_{\pm.014}$ | $.739_{\pm.014}$ |
| | REFIT | $512{+}n$ | $.637_{\pm.046}$ | $.694_{\pm.023}$ | $\mathbf{.810}_{\pm.006}$ | $\mathbf{.810}_{\pm.006}$ | $\mathbf{.810}_{\pm.006}$ |
| | **Re-center (ours)** | 0 | $\mathbf{.743}_{\pm.006}$ | $\mathbf{.743}_{\pm.006}$ | $.743_{\pm.006}$ | $.743_{\pm.006}$ | $.743_{\pm.006}$ |
| | Unlabeled inputs used by re-centering | | 15 | 29 | 58 | 116 | 233 |

Table 12: **Sample cost of re-centering** (Corollary 2; 27 cells, three seeds). Excess: Brier of re-centering at budget $n$ minus at 512.

| **Labeled calibration budget** $n$ | 32 | 64 | 128 | 256 |
|---|---|---|---|---|
| Unlabeled inputs used by re-centering | $14.7_{\pm 0.6}$ | $29.4_{\pm 1.3}$ | $57.7_{\pm 1.4}$ | $116.4_{\pm 3.3}$ |
| Bound on excess Brier, max over cells | 0.0109 | 0.0042 | 0.0026 | 0.0014 |
| Observed excess Brier, max over cells | 0.0045 | 0.0038 | 0.0017 | 0.0025 |
| Cells within the bound (of 27) | 25 | 23 | 26 | 22 |
| Re-center beats FROZEN in Brier (of 27) | 27 | 27 | 27 | 27 |
| Labeled maps with no error: observed | 15 | 10 | 0 | 0 |
| expected, sum of $(1-p)^m$ | 5.7 | 1.9 | 0.4 | 0.0 |

Table 13: **Selective prediction at the answer end** (reasoning traces, seed 0). Risk@$c$: error rate among the $c$ fraction of answers judged safest.

| Update | Policy | AUROC | AURC | Risk@80% | Risk@90% |
|---|---|---|---|---|---|
| SFT | FROZEN | 0.709 | 0.041 | 0.066 | 0.078 |
| SFT | RECALIBRATE | 0.676 | 0.047 | 0.068 | 0.079 |
| SFT | REFIT | 0.818 | 0.027 | 0.054 | 0.071 |
| GRPO | FROZEN | 0.735 | 0.039 | 0.065 | 0.076 |
| GRPO | RECALIBRATE | 0.735 | 0.039 | 0.065 | 0.076 |
| GRPO | REFIT | 0.819 | 0.034 | 0.051 | 0.068 |
| OPD | FROZEN | 0.796 | 0.051 | 0.081 | 0.111 |
| OPD | RECALIBRATE | 0.796 | 0.051 | 0.081 | 0.111 |
| OPD | REFIT | 0.814 | 0.050 | 0.073 | 0.098 |

## H RESULTS BY MODEL AND TASK

Tables 15 to 18 break Table 1 down with the same estimators and bold rule; rotated labels give the error rate before and after the update.

Table 14: **Repair hierarchy on short answers** (71 post-training cells). Labels come from half A, capped at its size; scored on half B.

| | FROZEN | Re-center | + intercept | | | | + intercept and slope | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Labels (at most) | 0 | 0 | 16 | 32 | 64 | 128 | 16 | 32 | 64 | 128 |
| ECE↓ | .441 | .212 | .147 | .139 | .142 | .133 | .136 | .103 | .096 | **.080** |
| Brier↓ | .381 | .229 | .188 | .180 | .182 | .181 | .164 | .152 | .149 | **.146** |

Table 15: **Short answers by model.** Each model pools eight tasks and the rollout variants; mean$_{\pm \text{s.d.}}$ over three training seeds. Bold: best estimator that needs no labels.

| | | | Labels | after SFT | | after GRPO | | after OPD | |
|---|---|---|---|---|---|---|---|---|---|
| | Signal | Estimator | needed | AUROC↑ | ECE↓ | AUROC↑ | ECE↓ | AUROC↑ | ECE↓ |
| InternVL3.5-8B error .47→.29 | Probe before the update | | | .859 | .075 | .859 | .075 | | |
| | Output | Answer probability | 0 | **.861**$_{\pm.009}$ | .090$_{\pm.004}$ | **.861**$_{\pm.015}$ | .095$_{\pm.003}$ | | |
| | | Self-consistency | 0 | .828$_{\pm.012}$ | **.082**$_{\pm.009}$ | .827$_{\pm.010}$ | **.091**$_{\pm.004}$ | | |
| | Hidden state | REFIT | 2,138 | .818$_{\pm.012}$ | .089$_{\pm.009}$ | .821$_{\pm.013}$ | .093$_{\pm.006}$ | - | |
| | | RECALIBRATE | ≈188 | .757$_{\pm.009}$ | .081$_{\pm.004}$ | .751$_{\pm.004}$ | .081$_{\pm.005}$ | | |
| | | FROZEN | 0 | .757$_{\pm.009}$ | .356$_{\pm.010}$ | .751$_{\pm.004}$ | .343$_{\pm.008}$ | | |
| | | **Re-center (ours)** | 0 | .757$_{\pm.009}$ | .219$_{\pm.004}$ | .751$_{\pm.004}$ | .215$_{\pm.005}$ | | |
| Qwen3.5-4B error .44→.30 | Probe before the update | | | .886 | .084 | .886 | .084 | .886 | .084 |
| | Output | Answer probability | 0 | **.867**$_{\pm.005}$ | **.096**$_{\pm.004}$ | **.856**$_{\pm.015}$ | **.102**$_{\pm.005}$ | **.864**$_{\pm.013}$ | **.092**$_{\pm.004}$ |
| | | Self-consistency | 0 | .809$_{\pm.005}$ | **.096**$_{\pm.008}$ | .796$_{\pm.008}$ | .107$_{\pm.005}$ | .809$_{\pm.010}$ | .094$_{\pm.002}$ |
| | Hidden state | REFIT | 2,138 | .841$_{\pm.006}$ | .102$_{\pm.016}$ | .839$_{\pm.008}$ | .100$_{\pm.004}$ | .840$_{\pm.004}$ | .104$_{\pm.014}$ |
| | | RECALIBRATE | ≈188 | .793$_{\pm.012}$ | .098$_{\pm.005}$ | .787$_{\pm.006}$ | .099$_{\pm.004}$ | .789$_{\pm.007}$ | .101$_{\pm.006}$ |
| | | FROZEN | 0 | .793$_{\pm.012}$ | .526$_{\pm.014}$ | .787$_{\pm.006}$ | .501$_{\pm.010}$ | .789$_{\pm.007}$ | .525$_{\pm.012}$ |
| | | **Re-center (ours)** | 0 | .793$_{\pm.012}$ | .174$_{\pm.010}$ | .787$_{\pm.006}$ | .171$_{\pm.015}$ | .789$_{\pm.007}$ | .176$_{\pm.006}$ |
| Qwen3.5-9B error .40→.28 | Probe before the update | | | .879 | .086 | .879 | .086 | | |
| | Output | Answer probability | 0 | **.850**$_{\pm.010}$ | **.078**$_{\pm.010}$ | **.845**$_{\pm.014}$ | **.092**$_{\pm.009}$ | | |
| | | Self-consistency | 0 | .813$_{\pm.011}$ | .096$_{\pm.004}$ | .800$_{\pm.016}$ | .106$_{\pm.002}$ | | |
| | Hidden state | REFIT | 2,138 | .845$_{\pm.014}$ | .088$_{\pm.002}$ | .845$_{\pm.015}$ | .085$_{\pm.002}$ | - | |
| | | RECALIBRATE | ≈188 | .761$_{\pm.009}$ | .084$_{\pm.012}$ | .764$_{\pm.009}$ | .081$_{\pm.006}$ | | |
| | | FROZEN | 0 | .761$_{\pm.009}$ | .558$_{\pm.012}$ | .764$_{\pm.009}$ | .529$_{\pm.010}$ | | |
| | | **Re-center (ours)** | 0 | .761$_{\pm.009}$ | .175$_{\pm.006}$ | .764$_{\pm.009}$ | .176$_{\pm.003}$ | | |
| Qwen3-VL-2B error .45→.34 | Probe before the update | | | .821 | .078 | .821 | .078 | .821 | .078 |
| | Output | Answer probability | 0 | **.841**$_{\pm.005}$ | **.085**$_{\pm.005}$ | **.827**$_{\pm.005}$ | **.101**$_{\pm.003}$ | **.838**$_{\pm.007}$ | **.080**$_{\pm.006}$ |
| | | Self-consistency | 0 | .808$_{\pm.004}$ | .097$_{\pm.002}$ | .788$_{\pm.001}$ | .117$_{\pm.001}$ | .811$_{\pm.006}$ | .089$_{\pm.006}$ |
| | Hidden state | REFIT | 2,138 | .798$_{\pm.006}$ | .096$_{\pm.017}$ | .799$_{\pm.004}$ | .087$_{\pm.005}$ | .803$_{\pm.008}$ | .092$_{\pm.012}$ |
| | | RECALIBRATE | ≈188 | .650$_{\pm.013}$ | .092$_{\pm.007}$ | .666$_{\pm.023}$ | .094$_{\pm.008}$ | .649$_{\pm.013}$ | .091$_{\pm.004}$ |
| | | FROZEN | 0 | .653$_{\pm.009}$ | .478$_{\pm.009}$ | .669$_{\pm.017}$ | .458$_{\pm.013}$ | .652$_{\pm.011}$ | .488$_{\pm.008}$ |
| | | **Re-center (ours)** | 0 | .653$_{\pm.009}$ | .222$_{\pm.001}$ | .669$_{\pm.017}$ | .218$_{\pm.009}$ | .652$_{\pm.011}$ | .227$_{\pm.006}$ |
| Qwen3-VL-8B error .44→.31 | Probe before the update | | | .846 | .073 | .846 | .073 | | |
| | Output | Answer probability | 0 | **.833**$_{\pm.013}$ | **.086**$_{\pm.004}$ | **.836**$_{\pm.005}$ | **.102**$_{\pm.002}$ | | |
| | | Self-consistency | 0 | .809$_{\pm.017}$ | .102$_{\pm.003}$ | .790$_{\pm.012}$ | .113$_{\pm.003}$ | | |
| | Hidden state | REFIT | 2,138 | .820$_{\pm.011}$ | .082$_{\pm.010}$ | .820$_{\pm.009}$ | .085$_{\pm.002}$ | - | |
| | | RECALIBRATE | ≈188 | .733$_{\pm.008}$ | .088$_{\pm.010}$ | .729$_{\pm.008}$ | .087$_{\pm.007}$ | | |
| | | FROZEN | 0 | .722$_{\pm.012}$ | .361$_{\pm.006}$ | .726$_{\pm.009}$ | .357$_{\pm.009}$ | | |
| | | **Re-center (ours)** | 0 | .722$_{\pm.012}$ | .191$_{\pm.003}$ | .726$_{\pm.009}$ | .189$_{\pm.009}$ | | |

Table 16: **Short answers by task: medical VQA.** Model-equal means over five VLMs (OPD: the two VLMs with an OPD student); mean$_{\pm \text{s.d.}}$ over three training seeds.

| | | | Labels | after SFT | | after GRPO | | after OPD | |
|---|---|---|---|---|---|---|---|---|---|
| | Signal | Estimator | needed | AUROC↑ | ECE↓ | AUROC↑ | ECE↓ | AUROC↑ | ECE↓ |
| SLAKE-en error .48→.31 | Probe before the update | | | .892 | .133 | .892 | .133 | .892 | .133 |
| | Output | Answer probability | 0 | **.824**$_{\pm.019}$ | **.139**$_{\pm.019}$ | **.812**$_{\pm.038}$ | .157$_{\pm.012}$ | **.850**$_{\pm.037}$ | **.135**$_{\pm.010}$ |
| | | Self-consistency | 0 | .811$_{\pm.013}$ | .141$_{\pm.006}$ | .795$_{\pm.012}$ | **.155**$_{\pm.007}$ | .842$_{\pm.039}$ | .150$_{\pm.017}$ |
| | Hidden state | REFIT | 2,138 | .849$_{\pm.019}$ | .133$_{\pm.017}$ | .845$_{\pm.021}$ | .136$_{\pm.005}$ | .884$_{\pm.038}$ | .126$_{\pm.023}$ |
| | | RECALIBRATE | ≈49 | .675$_{\pm.020}$ | .161$_{\pm.019}$ | .668$_{\pm.013}$ | .159$_{\pm.016}$ | .706$_{\pm.044}$ | .161$_{\pm.025}$ |
| | | FROZEN | 0 | .675$_{\pm.020}$ | .488$_{\pm.002}$ | .668$_{\pm.013}$ | .481$_{\pm.009}$ | .706$_{\pm.044}$ | .488$_{\pm.034}$ |
| | | **Re-center (ours)** | 0 | .675$_{\pm.020}$ | .239$_{\pm.012}$ | .668$_{\pm.013}$ | .249$_{\pm.021}$ | .706$_{\pm.044}$ | .217$_{\pm.024}$ |
| SLAKE-zh error .83→.23 | Probe before the update | | | .932 | .101 | .932 | .101 | .932 | .101 |
| | Output | Answer probability | 0 | **.873**$_{\pm.023}$ | **.106**$_{\pm.007}$ | **.865**$_{\pm.040}$ | .119$_{\pm.011}$ | **.877**$_{\pm.044}$ | **.106**$_{\pm.006}$ |
| | | Self-consistency | 0 | .868$_{\pm.022}$ | .118$_{\pm.007}$ | **.865**$_{\pm.017}$ | **.111**$_{\pm.007}$ | .868$_{\pm.012}$ | .126$_{\pm.021}$ |
| | Hidden state | REFIT | 2,138 | .817$_{\pm.037}$ | .123$_{\pm.017}$ | .817$_{\pm.020}$ | .113$_{\pm.002}$ | .822$_{\pm.014}$ | .111$_{\pm.006}$ |
| | | RECALIBRATE | ≈46 | .486$_{\pm.055}$ | .167$_{\pm.025}$ | .472$_{\pm.022}$ | .167$_{\pm.009}$ | .463$_{\pm.049}$ | .220$_{\pm.021}$ |
| | | FROZEN | 0 | .468$_{\pm.045}$ | .544$_{\pm.021}$ | .468$_{\pm.018}$ | .538$_{\pm.016}$ | .463$_{\pm.049}$ | .520$_{\pm.020}$ |
| | | **Re-center (ours)** | 0 | .468$_{\pm.045}$ | .627$_{\pm.018}$ | .468$_{\pm.018}$ | .621$_{\pm.013}$ | .463$_{\pm.049}$ | .602$_{\pm.019}$ |
| VQA-RAD error .50→.44 | Probe before the update | | | .856 | .083 | .856 | .083 | .856 | .083 |
| | Output | Answer probability | 0 | .820$_{\pm.008}$ | .138$_{\pm.017}$ | .805$_{\pm.006}$ | .138$_{\pm.008}$ | .814$_{\pm.001}$ | .154$_{\pm.012}$ |
| | | Self-consistency | 0 | **.839**$_{\pm.014}$ | **.107**$_{\pm.010}$ | **.836**$_{\pm.004}$ | **.115**$_{\pm.009}$ | **.827**$_{\pm.009}$ | **.086**$_{\pm.005}$ |
| | Hidden state | REFIT | 2,138 | .830$_{\pm.009}$ | .117$_{\pm.016}$ | .835$_{\pm.012}$ | .115$_{\pm.006}$ | .831$_{\pm.012}$ | .121$_{\pm.025}$ |
| | | RECALIBRATE | ≈112 | .742$_{\pm.031}$ | .107$_{\pm.003}$ | .748$_{\pm.029}$ | .106$_{\pm.003}$ | .691$_{\pm.062}$ | .103$_{\pm.011}$ |
| | | FROZEN | 0 | .747$_{\pm.023}$ | .423$_{\pm.004}$ | .755$_{\pm.019}$ | .417$_{\pm.008}$ | .703$_{\pm.042}$ | .470$_{\pm.009}$ |
| | | **Re-center (ours)** | 0 | .747$_{\pm.023}$ | .161$_{\pm.004}$ | .755$_{\pm.019}$ | .156$_{\pm.003}$ | .703$_{\pm.042}$ | .207$_{\pm.026}$ |
| PathVQA error .60→.46 | Probe before the update | | | .909 | .054 | .909 | .054 | .909 | .054 |
| | Output | Answer probability | 0 | **.893**$_{\pm.003}$ | .064$_{\pm.004}$ | **.895**$_{\pm.004}$ | **.066**$_{\pm.002}$ | **.902**$_{\pm.008}$ | .069$_{\pm.006}$ |
| | | Self-consistency | 0 | .859$_{\pm.005}$ | **.061**$_{\pm.011}$ | .867$_{\pm.005}$ | .081$_{\pm.005}$ | .861$_{\pm.005}$ | **.055**$_{\pm.006}$ |
| | Hidden state | REFIT | 2,138 | .886$_{\pm.006}$ | .062$_{\pm.004}$ | .889$_{\pm.002}$ | .062$_{\pm.006}$ | .883$_{\pm.008}$ | .071$_{\pm.013}$ |
| | | RECALIBRATE | ≈275 | .824$_{\pm.005}$ | .091$_{\pm.007}$ | .828$_{\pm.002}$ | .097$_{\pm.004}$ | .810$_{\pm.011}$ | .101$_{\pm.012}$ |
| | | FROZEN | 0 | .824$_{\pm.005}$ | .428$_{\pm.006}$ | .828$_{\pm.002}$ | .422$_{\pm.005}$ | .810$_{\pm.011}$ | .479$_{\pm.015}$ |
| | | **Re-center (ours)** | 0 | .824$_{\pm.005}$ | .292$_{\pm.005}$ | .828$_{\pm.002}$ | .284$_{\pm.002}$ | .810$_{\pm.011}$ | .304$_{\pm.015}$ |

Table 17: **Short answers by task: general VQA.** Model-equal means over five VLMs (OPD: the two VLMs with an OPD student); mean$_{\pm \text{s.d.}}$ over three training seeds.

| | | | Labels | after SFT | | after GRPO | | after OPD | |
|---|---|---|---|---|---|---|---|---|---|
| | Signal | Estimator | needed | AUROC↑ | ECE↓ | AUROC↑ | ECE↓ | AUROC↑ | ECE↓ |
| ScienceQA error .07→.07 | Probe before the update | | | .923 | .035 | .923 | .035 | .923 | .035 |
| | Output | Answer probability | 0 | **.937**$_{\pm.008}$ | .026$_{\pm.004}$ | **.934**$_{\pm.003}$ | .027$_{\pm.002}$ | **.926**$_{\pm.009}$ | .031$_{\pm.003}$ |
| | Output | Self-consistency | 0 | .864$_{\pm.011}$ | **.024**$_{\pm.002}$ | .828$_{\pm.011}$ | .032$_{\pm.002}$ | .894$_{\pm.020}$ | **.028**$_{\pm.002}$ |
| | Hidden state | REFIT | 2,138 | .920$_{\pm.006}$ | .026$_{\pm.004}$ | .919$_{\pm.004}$ | .026$_{\pm.002}$ | .905$_{\pm.022}$ | .038$_{\pm.005}$ |
| | Hidden state | RECALIBRATE | ≈256 | .907$_{\pm.007}$ | .030$_{\pm.002}$ | .909$_{\pm.003}$ | .026$_{\pm.003}$ | .887$_{\pm.023}$ | .031$_{\pm.008}$ |
| | Hidden state | FROZEN | 0 | .907$_{\pm.007}$ | .460$_{\pm.010}$ | .909$_{\pm.003}$ | .412$_{\pm.011}$ | .887$_{\pm.023}$ | .569$_{\pm.024}$ |
| | Hidden state | **Re-center (ours)** | 0 | .907$_{\pm.007}$ | .029$_{\pm.005}$ | .909$_{\pm.003}$ | **.026**$_{\pm.002}$ | .887$_{\pm.023}$ | .047$_{\pm.009}$ |
| A-OKVQA error .15→.15 | Probe before the update | | | .843 | .040 | .843 | .040 | .843 | .040 |
| | Output | Answer probability | 0 | **.846**$_{\pm.004}$ | **.053**$_{\pm.003}$ | **.848**$_{\pm.001}$ | **.067**$_{\pm.002}$ | **.858**$_{\pm.011}$ | **.051**$_{\pm.005}$ |
| | Output | Self-consistency | 0 | .796$_{\pm.006}$ | .066$_{\pm.005}$ | .776$_{\pm.014}$ | .079$_{\pm.004}$ | .801$_{\pm.024}$ | .063$_{\pm.011}$ |
| | Hidden state | REFIT | 2,138 | .843$_{\pm.002}$ | .036$_{\pm.004}$ | .846$_{\pm.000}$ | .039$_{\pm.003}$ | .848$_{\pm.005}$ | .042$_{\pm.012}$ |
| | Hidden state | RECALIBRATE | ≈271 | .827$_{\pm.004}$ | .048$_{\pm.006}$ | .832$_{\pm.002}$ | .049$_{\pm.006}$ | .830$_{\pm.005}$ | .053$_{\pm.007}$ |
| | Hidden state | FROZEN | 0 | .827$_{\pm.004}$ | .515$_{\pm.009}$ | .832$_{\pm.002}$ | .473$_{\pm.008}$ | .830$_{\pm.005}$ | .563$_{\pm.009}$ |
| | Hidden state | **Re-center (ours)** | 0 | .827$_{\pm.004}$ | .076$_{\pm.003}$ | .832$_{\pm.002}$ | .072$_{\pm.003}$ | .830$_{\pm.005}$ | .092$_{\pm.007}$ |
| POPE error .10→.09 | Probe before the update | | | .854 | .046 | .854 | .046 | .854 | .046 |
| | Output | Answer probability | 0 | **.855**$_{\pm.018}$ | **.048**$_{\pm.003}$ | **.853**$_{\pm.013}$ | .060$_{\pm.001}$ | **.832**$_{\pm.028}$ | .059$_{\pm.003}$ |
| | Output | Self-consistency | 0 | .692$_{\pm.023}$ | .059$_{\pm.002}$ | .658$_{\pm.019}$ | .066$_{\pm.005}$ | .624$_{\pm.026}$ | .068$_{\pm.005}$ |
| | Hidden state | REFIT | 2,138 | .855$_{\pm.006}$ | .039$_{\pm.003}$ | .853$_{\pm.005}$ | .039$_{\pm.004}$ | .822$_{\pm.019}$ | .042$_{\pm.005}$ |
| | Hidden state | RECALIBRATE | ≈258 | .817$_{\pm.007}$ | .032$_{\pm.003}$ | .821$_{\pm.013}$ | .029$_{\pm.001}$ | .744$_{\pm.019}$ | .030$_{\pm.010}$ |
| | Hidden state | FROZEN | 0 | .817$_{\pm.007}$ | .635$_{\pm.018}$ | .821$_{\pm.013}$ | .602$_{\pm.021}$ | .744$_{\pm.019}$ | .754$_{\pm.021}$ |
| | Hidden state | **Re-center (ours)** | 0 | .817$_{\pm.007}$ | .055$_{\pm.005}$ | .821$_{\pm.013}$ | **.053**$_{\pm.002}$ | .744$_{\pm.019}$ | **.041**$_{\pm.005}$ |
| VizWiz error .79→.73 | Probe before the update | | | .655 | .143 | .655 | .143 | .655 | .143 |
| | Output | Answer probability | 0 | .756$_{\pm.005}$ | .124$_{\pm.019}$ | .748$_{\pm.006}$ | .154$_{\pm.011}$ | .749$_{\pm.005}$ | **.083**$_{\pm.010}$ |
| | Output | Self-consistency | 0 | **.778**$_{\pm.004}$ | .182$_{\pm.008}$ | **.775**$_{\pm.006}$ | .214$_{\pm.008}$ | **.761**$_{\pm.003}$ | .153$_{\pm.011}$ |
| | Hidden state | REFIT | 2,138 | .596$_{\pm.010}$ | .194$_{\pm.033}$ | .596$_{\pm.011}$ | .192$_{\pm.027}$ | .577$_{\pm.007}$ | .235$_{\pm.061}$ |
| | Hidden state | RECALIBRATE | ≈239 | .633$_{\pm.002}$ | .074$_{\pm.012}$ | .637$_{\pm.004}$ | .073$_{\pm.004}$ | .622$_{\pm.005}$ | .073$_{\pm.004}$ |
| | Hidden state | FROZEN | 0 | .633$_{\pm.002}$ | .153$_{\pm.005}$ | .637$_{\pm.004}$ | .156$_{\pm.008}$ | .622$_{\pm.005}$ | .208$_{\pm.007}$ |
| | Hidden state | **Re-center (ours)** | 0 | .633$_{\pm.002}$ | **.089**$_{\pm.011}$ | .637$_{\pm.004}$ | **.091**$_{\pm.010}$ | .622$_{\pm.005}$ | .099$_{\pm.004}$ |

Table 18: **Reasoning traces by model and task.** Answer end; mean$_{\pm \text{s.d.}}$ over training seeds (two seeds for 2B ScienceQA OPD, 8B ScienceQA GRPO, 8B ScienceQA SFT, where one seed's 32-label budget is undefined).

| | | | Labels | after SFT | | after GRPO | | after OPD | |
|---|---|---|---|---|---|---|---|---|---|
| | Signal | Estimator | needed | AUROC↑ | ECE↓ | AUROC↑ | ECE↓ | AUROC↑ | ECE↓ |
| Qwen3-VL-2B ScienceQA error .12→.13 | Probe before the update | | | .810 | .030 | .810 | .030 | .810 | .030 |
| | Output | Verbal confidence | 0 | $.602_{\pm 024}$ | $.130_{\pm 016}$ | $.573_{\pm 043}$ | $.138_{\pm 009}$ | $.595_{\pm 000}$ | $.138_{\pm 011}$ |
| | | P(True) | 0 | $\mathbf{.858}_{\pm 010}$ | $.241_{\pm 037}$ | $\mathbf{.854}_{\pm 013}$ | $.186_{\pm 027}$ | $\mathbf{.867}_{\pm 023}$ | $.245_{\pm 013}$ |
| | | Answer probability | 0 | $.762_{\pm 013}$ | $.064_{\pm 004}$ | $.791_{\pm 036}$ | $.085_{\pm 020}$ | $.761_{\pm 007}$ | $.073_{\pm 000}$ |
| | | Token history ($L$) | 0 | $.704_{\pm 013}$ | $.163_{\pm 019}$ | $.726_{\pm 021}$ | $.125_{\pm 025}$ | $.707_{\pm 035}$ | $.153_{\pm 014}$ |
| | | Self-consistency | 0 | $.769_{\pm 065}$ | $.076_{\pm 030}$ | $.781_{\pm 048}$ | $.065_{\pm 024}$ | $.792_{\pm 019}$ | $.066_{\pm 016}$ |
| | Hidden state | REFIT | 1,024 | $.825_{\pm 004}$ | $.041_{\pm 015}$ | $.815_{\pm 011}$ | $.058_{\pm 013}$ | $.802_{\pm 020}$ | $.037_{\pm 014}$ |
| | | RECALIBRATE | 512 | $.750_{\pm 035}$ | $.032_{\pm 019}$ | $.759_{\pm 056}$ | $.043_{\pm 014}$ | $.760_{\pm 070}$ | $.040_{\pm 031}$ |
| | | FROZEN | 0 | $.750_{\pm 035}$ | $.499_{\pm 168}$ | $.759_{\pm 056}$ | $.292_{\pm 141}$ | $.760_{\pm 070}$ | $.626_{\pm 032}$ |
| | | **Re-center (ours)** | 0 | $.750_{\pm 035}$ | $\mathbf{.046}_{\pm 008}$ | $.759_{\pm 056}$ | $\mathbf{.049}_{\pm 022}$ | $.760_{\pm 070}$ | $\mathbf{.038}_{\pm 003}$ |
| Qwen3-VL-2B GSM8K error .16→.15 | Probe before the update | | | .778 | .055 | .778 | .055 | .778 | .055 |
| | Output | Verbal confidence | 0 | $.490_{\pm 002}$ | $.154_{\pm 012}$ | $.503_{\pm 002}$ | $.159_{\pm 012}$ | $.489_{\pm 004}$ | $.157_{\pm 021}$ |
| | | P(True) | 0 | $.592_{\pm 014}$ | $.487_{\pm 037}$ | $.587_{\pm 025}$ | $.462_{\pm 044}$ | $.585_{\pm 009}$ | $.491_{\pm 024}$ |
| | | Answer probability | 0 | $.629_{\pm 034}$ | $.135_{\pm 015}$ | $.641_{\pm 017}$ | $.143_{\pm 019}$ | $.648_{\pm 024}$ | $.137_{\pm 022}$ |
| | | Token history ($L$) | 0 | $.675_{\pm 020}$ | $.097_{\pm 025}$ | $.670_{\pm 042}$ | $.071_{\pm 014}$ | $.674_{\pm 014}$ | $.110_{\pm 035}$ |
| | | Self-consistency | 0 | $\mathbf{.878}_{\pm 021}$ | $\mathbf{.036}_{\pm 002}$ | $\mathbf{.878}_{\pm 025}$ | $\mathbf{.043}_{\pm 001}$ | $\mathbf{.885}_{\pm 016}$ | $\mathbf{.037}_{\pm 010}$ |
| | Hidden state | REFIT | 1,024 | $.794_{\pm 014}$ | $.029_{\pm 014}$ | $.807_{\pm 017}$ | $.040_{\pm 011}$ | $.799_{\pm 021}$ | $.045_{\pm 015}$ |
| | | RECALIBRATE | 512 | $.752_{\pm 017}$ | $.037_{\pm 008}$ | $.763_{\pm 011}$ | $.045_{\pm 016}$ | $.764_{\pm 018}$ | $.038_{\pm 005}$ |
| | | FROZEN | 0 | $.752_{\pm 017}$ | $.530_{\pm 042}$ | $.763_{\pm 011}$ | $.301_{\pm 037}$ | $.764_{\pm 018}$ | $.583_{\pm 026}$ |
| | | **Re-center (ours)** | 0 | $.752_{\pm 017}$ | $.047_{\pm 009}$ | $.763_{\pm 011}$ | $.045_{\pm 011}$ | $.764_{\pm 018}$ | $\mathbf{.037}_{\pm 007}$ |
| Qwen3-VL-8B ScienceQA error .07→.05 | Probe before the update | | | .866 | .016 | .866 | .016 | | |
| | Output | Verbal confidence | 0 | $.543_{\pm 019}$ | $.044_{\pm 008}$ | $.584_{\pm 009}$ | $.039_{\pm 009}$ | | |
| | | P(True) | 0 | $\mathbf{.910}_{\pm 006}$ | $.075_{\pm 018}$ | $\mathbf{.897}_{\pm 005}$ | $.066_{\pm 001}$ | | |
| | | Answer probability | 0 | $.636_{\pm 119}$ | $.073_{\pm 000}$ | $.670_{\pm 072}$ | $.058_{\pm 013}$ | | |
| | | Token history ($L$) | 0 | $.494_{\pm 030}$ | $.041_{\pm 014}$ | $.512_{\pm 109}$ | $\mathbf{.027}_{\pm 015}$ | | |
| | | Self-consistency | 0 | $.746_{\pm 129}$ | $.226_{\pm 163}$ | $.779_{\pm 169}$ | $.224_{\pm 193}$ | - | |
| | Hidden state | REFIT | 1,024 | $.795_{\pm 021}$ | $.025_{\pm 024}$ | $.825_{\pm 036}$ | $.039_{\pm 002}$ | | |
| | | RECALIBRATE | 512 | $.532_{\pm 139}$ | $.012_{\pm 010}$ | $.648_{\pm 062}$ | $.026_{\pm 008}$ | | |
| | | FROZEN | 0 | $.598_{\pm 045}$ | $.847_{\pm 021}$ | $.648_{\pm 062}$ | $.644_{\pm 109}$ | | |
| | | **Re-center (ours)** | 0 | $.598_{\pm 045}$ | $\mathbf{.038}_{\pm 014}$ | $.648_{\pm 062}$ | $.042_{\pm 002}$ | | |
| Qwen3-VL-8B GSM8K error .07→.08 | Probe before the update | | | .862 | .023 | .862 | .023 | | |
| | Output | Verbal confidence | 0 | $.619_{\pm 033}$ | $.064_{\pm 009}$ | $.635_{\pm 031}$ | $.062_{\pm 011}$ | | |
| | | P(True) | 0 | $.628_{\pm 013}$ | $.534_{\pm 046}$ | $.619_{\pm 002}$ | $.542_{\pm 054}$ | | |
| | | Answer probability | 0 | $.742_{\pm 027}$ | $.076_{\pm 007}$ | $.753_{\pm 032}$ | $.074_{\pm 008}$ | | |
| | | Token history ($L$) | 0 | $.728_{\pm 055}$ | $.402_{\pm 005}$ | $.724_{\pm 028}$ | $.344_{\pm 018}$ | | |
| | | Self-consistency | 0 | $\mathbf{.881}_{\pm 042}$ | $.073_{\pm 067}$ | $\mathbf{.842}_{\pm 042}$ | $.080_{\pm 070}$ | - | |
| | Hidden state | REFIT | 1,024 | $.829_{\pm 019}$ | $.031_{\pm 004}$ | $.814_{\pm 047}$ | $.033_{\pm 010}$ | | |
| | | RECALIBRATE | 512 | $.773_{\pm 063}$ | $.029_{\pm 005}$ | $.787_{\pm 054}$ | $.031_{\pm 007}$ | | |
| | | FROZEN | 0 | $.773_{\pm 063}$ | $.809_{\pm 067}$ | $.787_{\pm 054}$ | $.762_{\pm 096}$ | | |
| | | **Re-center (ours)** | 0 | $.773_{\pm 063}$ | $\mathbf{.027}_{\pm 010}$ | $.787_{\pm 054}$ | $\mathbf{.031}_{\pm 004}$ | | |