CALIBRATION, NOT ANSWER SELECTION: DISTILLING INTERNAL CONFIDENCE IN REASONING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with binary correctness rewards trains correctness, not calibrated confidence. The confidence that reasoning models verbalize is systematically overconfident, and the problem is not merely one of scale: verbalized confidence tracks how willing a model is to commit to an answer, not how likely the answer is to be right. Post-hoc rescaling therefore fits one distribution but rarely transfers. We look inside the model instead. On factual question answering, a linear probe on the hidden state between the chain of thought and the answer is substantially better calibrated: its expected calibration error is 5–38$\times$ lower than that of the verbalized score across four benchmarks and two model families. However, when used to pick among N sampled answers, that same probe nearly ties majority voting yet falls far short of the oracle. Internal states answer "how certain am I" well and "which answer is right" poorly, so the signal should be reported as a confidence rather than used to select answers. As a result, we introduce *probe-guided self-distillation* (Probe-SD): score a model's own sampled traces with the probe, overwrite the confidence each trace states, and finetune the base checkpoint of the same family, so nothing but the model itself remains at test time. On Qwen3-14B, Probe-SD cuts ECE from 0.178 to 0.024 in-domain and from 0.542 to 0.113 out-of-domain, where it also beats post-hoc recalibration and self-consistency distillation. The resulting confidence is well-calibrated and useful for weighted voting—behaviors previously attributed to online RL, here obtained with supervised finetuning alone. Code will be released upon publication.

1 INTRODUCTION

Reasoning models—trained with reinforcement learning to think in natural language before answering (OpenAI, 2024; DeepSeek-AI, 2025)—account for much of the recent progress in language modeling. A common recipe, RLVR, rewards binary correctness: a lucky guess pays as well as a confident correct answer, abstention costs as much as an error, and calibration receives no direct reward (Damani et al., 2025). Where the stakes are high, being right is not enough: a model must also say when it is unsure (Xiong et al., 2024). We study a model's *verbalized confidence* on factual QA. It requires no auxiliary judge to interpret. Uncertainty estimation is less reliable on knowledge-heavy factual tasks than on reasoning-intensive tasks such as mathematics (Tao et al., 2025).

Miscalibration here runs deeper than a mis-set scale. Kumaran (2026) find that a stated confidence predicts whether a model will submit or abstain far better than whether it will be right, including in reasoning models: the number reads as a decision about commitment rather than as evidence. Ji et al. (2025) reach the same conclusion from the wording of an answer rather than from the number: expressed uncertainty lies almost entirely along one linear direction, only loosely tied to semantic uncertainty. **The channel that expresses uncertainty has its own cause, and that cause is not the evidence for the answer.** Consequently, temperature scaling (Guo et al., 2017) and isotonic regression (Zadrozny & Elkan, 2002) rescale a signal that measures something else: they reduce ECE most on the fit distribution, transfer poorly under distribution shift, and add no discriminative power. Repairing verbalized confidence means changing its source. That raises two questions: does a reasoning model hold an internal signal that points at correctness on a trustworthy scale, and, if so, can the model be taught to say it?

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.

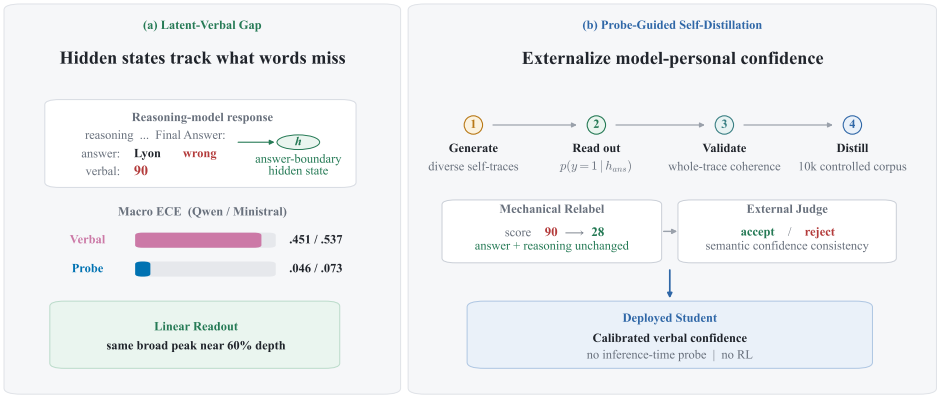

Figure 1: Probe-guided self-distillation. A hidden-state probe scores the reasoning model's own traces and overwrites the confidence each one states. The base checkpoint of the same family is then finetuned on the traces that pass a coherence filter.

A calibration signal exists in hidden states. Hidden states linearly encode output correctness (Alain & Bengio, 2018; Marks & Tegmark, 2024; Orgad et al., 2025; Gekhman et al., 2025), but prior evidence comes largely from non-reasoning models and focuses on *discrimination* rather than *calibration*. Across 80 models, the two are nearly rank-uncorrelated (Tao et al., 2025). We address this gap on factual QA with reasoning models. A linear probe at the boundary between the chain of thought and the answer is substantially better calibrated: its probability scale remains stable as questions become harder, it is most stable when read from the middle layers of both model families, and its ECE is markedly lower than that of the verbalized score.

But the probe is a poor answer selector. Good calibration does not imply reliable answer recovery from the model's internal states: when used to choose among N samples, the same probe nearly ties majority voting yet stays far below the oracle, echoing Servedio et al. (2025) on the limits of factuality encoding. **It is therefore more useful for estimating trace reliability than for identifying the correct answer among candidates.** Reranking candidates or filtering hallucinations during decoding requires information the probe does not reliably encode.

Probe-guided self-distillation (Probe-SD). A probe cannot be carried into deployment: it is unusable behind a text-only API, which is why deployment-oriented evaluations often exclude internal-state approaches (Tao et al., 2025). As an external readout, it does not shape what the model says. We therefore move it offline: the reasoning model samples its own traces, the probe scores them, the stated confidence is overwritten with the probe score, and the base checkpoint of the same family is finetuned on the result (Figure 1), so that only the model remains at test time. A suitable label must remain within the student's knowledge and representational capacity, score traces rather than questions, and be calibrated to the correctness of a single *trace*. An external judge fails the first, self-consistency the second, and a hidden-state probe meets all three.

Our contributions are the following.

- **Measurement.** On two model families and four benchmarks, a probe at the answer boundary outperforms the verbalized score on all 24 ECE, AUROC and Brier comparisons, with ECE 5–38$\times$ lower (§3.1–§3.2).
- **Boundary.** The signal calibrates but does not select: best-of-N selection nearly ties majority voting (§3.3), and dropping probe relabeling from training substantially worsens calibration while leaving accuracy intact (§5.3).
- **Method.** Supervised finetuning alone transfers the calibration signal into verbalized confidence: Qwen3-14B's ECE drops from 0.178 to 0.024 in-domain and from 0.542 to 0.113 out-of-domain (§5.2). §6 traces how the student inherits its teacher's calibration.

2 Related Work

Verbalized confidence and sampling-consistency proxies. A model can state a numerical confidence outright or hedge in its wording (Yang et al., 2024; Xiong et al., 2024; Yona et al., 2024; Belem et al., 2024). Stated numbers beat token-level conditional probabilities (Kadavath et al., 2022; Tian et al., 2023) but are systematically overconfident (Xiong et al., 2024; Mei et al., 2025). A reasoning mode only softens the error (Yoon et al., 2025; Tao et al., 2025). Kumaran (2026) identify the failure: the number predicts submit-or-abstain decisions better than correctness, so it is behavioral rather than evidential. This is why post-hoc recalibration (Guo et al., 2017; Zadrozny & Elkan, 2002) neither transfers across distributions nor adds discriminative power. Sampling consistency instead scores a question rather than a trace: self-consistency (Wang et al., 2023) and semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) cost an order of magnitude more inference, and semantic-entropy probes (Kossen et al., 2024) cut that cost. Offline distillation can amortize self-consistency into a single-generation predictor (Zollo et al., 2026). Trace length (Devic et al., 2025) is free but unscaled. None of these signals is both trace-level *and* on a trustworthy probability scale.

Training a model to express calibrated confidence. These methods differ mainly in what supplies the target. Mielke et al. (2022) recalibrate the wording of confidence, while Lin et al. (2022) use empirical accuracy over sampled questions, which motivates our SC-SD baseline. ConfTuner (Li et al., 2025) optimizes the probability of the final confidence token with a tokenized Brier score against answer correctness. LACIE (Stengel-Eskin et al., 2024) targets what a listener accepts. In parallel work, Cacioli (2026) use a hidden-state probe as the confidence target via LoRA on instruction-tuned models, focusing on discrimination rather than calibration. We instead target reasoning models and distill the probe's calibrated score into the confidence they state. HALT (Franzmeyer et al., 2026) uses fragment-level correctness to learn when to abstain. Ji et al. (2025) steer activations without correctness targets. A reinforcement-learning branch optimizes the stated confidence online instead: RLCR (Damani et al., 2025) adds a Brier term to the binary correctness reward, and SaySelf (Xu et al., 2024), Liu et al. (2026b), Guo et al. (2026), Stangel et al. (2025) and Ma et al. (2026) proceed similarly, at the cost of online sampling and reward design. Unlike these online methods, we calibrate an existing reasoning model with supervised finetuning.

Internal-state probes. Hidden states carry a linearly readable correctness signal, used for truthfulness discrimination (Azaria & Mitchell, 2023; Orgad et al., 2025; Burns et al., 2023), activation intervention (Li et al., 2024), and pre-generation awareness of knowledge boundaries (Ni et al., 2025). Its strength is contested: Gekhman et al. (2025) argue that internal factual knowledge exceeds what surfaces in the output, while Servedio et al. (2025) show that internal states pick the correct response out of a model's own samples only weakly. These probes use non-reasoning models and validate discrimination rather than calibration. The two are nearly rank-uncorrelated across models (Tao et al., 2025). The probes that do report ECE either target mathematics for early stopping (Zhang et al., 2025; Li et al., 2026), read a judge model's assessment of someone else's output (Radharapu et al., 2025), or reside at the last layer of a non-reasoning model (Liu et al., 2024). All keep the probe in the deployment path. We instead apply calibration probes to factual QA with reasoning models, distill their signal offline, and remove the probe from deployment.

Self-distillation. Within knowledge distillation (Hinton et al., 2015), self-distillation includes STaR (Zelikman et al., 2022) and ReST^{EM} (Singh et al., 2024), which bootstrap from self-generated, externally verified trajectories, and SSD (Zhang et al., 2026), which improves performance using only self-generated data. We keep the paradigm and change the target: probe scores replace the verbalized confidence score, and the student starts from the base model.

3 A Calibration Signal in Reasoning-Model Hidden States

We test whether reasoning models hold a trustworthy internal correctness signal. Across four benchmarks and two model families, a post-CoT probe is substantially better calibrated, but as an answer selector it nearly ties majority voting and falls far short of the oracle.

Setup. We evaluate on four factual QA benchmarks—TriviaQA (Joshi et al., 2017), EntityQuestions (Sciavolino et al., 2021), NQ-open (Kwiatkowski et al., 2019) and SimpleQA-Verified (Wei et al., 2024; Haas et al., 2025)—using 1,000 questions and 10 traces each, on Qwen3-14B (Qwen Team, 2025) and Ministral-3-8B-Reasoning-2512 (Liu et al., 2026a) at T=0.6, top-k=20, top-p=0.95.

Table 1: Verbalized score (V) versus hidden-state probe (P). All calibration metrics are computed on the subset that the source can score. Brier decomposition and missing-score sensitivity analysis are in Appendix E.

		Qwen3-14B					*Ministral-3-8B*				
Dataset		Acc	ECE ↓	AUROC ↑	Brier ↓	Cov. ↑	Acc	ECE ↓	AUROC ↑	Brier ↓	Cov. ↑
TriviaQA (ID)	V	.729	.178±.020	.825±.017	.184±.017	.980	.664	.247±.024	.669±.015	.263±.021	.957
	P		**.015**±.009	**.948**±.009	**.081**±.008	**1.000**		**.014**±.008	**.917**±.012	**.105**±.009	**.977**
EntityQ	V	.267	.529±.022	.812±.017	.455±.018	.903	.207	.627±.022	.734±.016	.564±.019	.950
	P		**.014**±.009	**.913**±.014	**.097**±.010	**1.000**		**.024**±.012	**.909**±.015	**.091**±.009	**.999**
NQ-open	V	.440	.436±.029	.705±.021	.416±.025	.723	.370	.513±.026	.643±.018	.490±.022	.798
	P		**.078**±.017	**.826**±.020	**.177**±.012	**1.000**		**.108**±.018	**.816**±.020	**.188**±.012	**.994**
SimpleQA	V	.066	.661±.016	.627±.042	.532±.013	.849	.061	.761±.013	.531±.040	.671±.012	.887
	P		**.075**±.012	**.704**±.048	**.076**±.008	**1.000**		**.146**±.013	**.640**±.048	**.116**±.009	**.979**

We fit an ℓ_2-regularized logistic probe at the *last pre-answer token* to trace correctness on 10,000 TriviaQA training questions. TriviaQA is in-domain and the other benchmarks are out-of-domain. Layer, readout, and regularization are chosen on validation data alone (Appendix C). Qwen3-32B-Instruct grades answers. Because all labels come from it, we manually audited 700 samples and validated its reliability (Appendix D). We report ECE (Naeini et al., 2015), AUROC, Brier (Brier, 1950) and Coverage (Cov.; the fraction of traces for which the confidence source produces a score); Appendix A gives detailed definitions and computation conventions.

3.1 Internal States Give Better-Calibrated Confidence

Table 1 compares the verbalized score with the hidden-state probe on the same sampled traces.

The probe improves every confidence metric. On EntityQ, Qwen's ECE falls from 0.529 to 0.014 and its AUROC rises from 0.812 to 0.913. Although the probe is trained only on TriviaQA, its out-of-domain EntityQ ECE (0.014 for Qwen, 0.024 for Ministral) matches the in-domain value, while the verbalized score deteriorates from 0.178/0.247 to 0.529/0.627. The probe's ECE rises only on NQ-open and SimpleQA (0.078–0.146), the lowest-base-rate benchmarks, where AUROC also drops (to 0.704/0.640 on SimpleQA): the probe remains better calibrated under distribution shift, while ranking weakens as base rates fall.

The two sources fail differently. ECE and Brier for the verbalized score are both higher and move together—Qwen on SimpleQA has ECE 0.661 and Brier 0.532 against the probe's 0.075 and 0.076, and the other seven model–dataset cells agree in direction. That high ECE reflects scale overconfidence: the verbalized confidence exceeds observed accuracy. The probe lowers ECE and improves Brier and AUROC, so its low error does not sacrifice discrimination.

3.2 The Signal Is Robust to Task, Layer, and Readout

Both models read out a better-than-chance correctness signal from the early layers, while the best calibration and ranking coincide on a broad plateau at roughly 60–65% relative depth: about L21–L27 for Qwen and L20–L24 for Ministral, with out-of-domain performance degrading toward the output layers (Appendix C). The signal is therefore *readable across layers and most stable in the middle*. The probe is equally insensitive to the readout: probes trained on the residual stream and on the MLP output differ by far less than the Probe–Verbal gap (Table 6).

3.3 A Well-Calibrated Probe Is a Poor Answer Selector

The probe now satisfies the requirements for a confidence source, but that does not imply that internal states hold more recoverable correct answers. Selecting, from a question's N samples, the trace with the highest probe score (Max-Probe) tracks lexically normalized majority voting and stays far below oracle pass@N, which counts a question solved whenever any correct answer appears among the samples. Selection accuracy is flat across layers, like calibration (Figure 2a), and in the Qwen3-14B TriviaQA curve the gap to the oracle widens with N: at N=32, Max-Probe reaches 0.778, majority voting 0.771, and the oracle 0.879 (Figure 2b). A 32-sample extension over both models

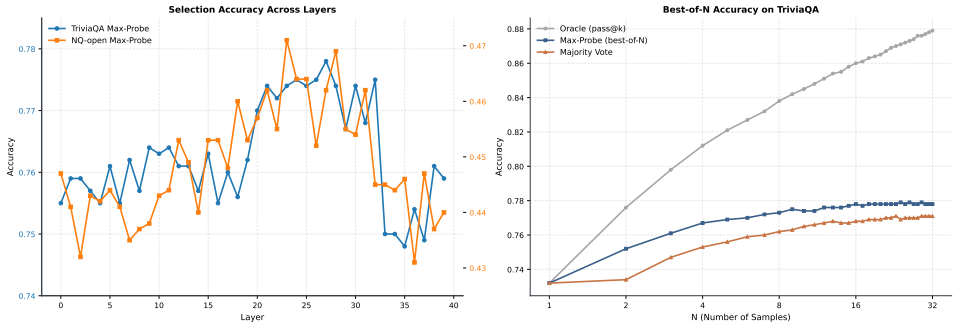

Figure 2: **The probe scores traces well and selects among them poorly.** (a) Max-Probe selection accuracy for Qwen3-14B on TriviaQA and NQ-open as a function of the layer. (b) Best-of-N curves on the TriviaQA 32-sample bench: Max-Probe tracks lexically normalized majority voting while the oracle pass@N pulls ahead.

and all four datasets (Appendix C) confirms the same picture: Max-Probe nearly ties majority voting, while both stay far below the oracle. This agrees with Servedio et al. (2025) and points the probe to trace-level calibration rather than answer selection.

4 Probe-Guided Self-Distillation

The probe cannot serve a text-only interface, so we move it from inference to offline supervision.

4.1 Problem Setting

Let the reasoning model π generate, for question q, a response $r = (\tau, a, s)$, where τ is the chain of thought, a the final answer, and $s \in \{1, \ldots, 100\}$ the verbalized confidence. The frozen probe reads the pre-answer hidden state h of that trace and yields $p = g_\phi(h) \in [0, 1]$. Relabeling keeps a and constructs $r' = (\tau', a, s')$, with s' aligned to p and τ' preserving the reasoning while staying linguistically consistent with s'. The relabeled traces supervise the same-family base model. We call the signal providing s' the *teacher* and the supervised base model the *student*.

4.2 Candidate Sampling and Probe Scoring

We sample N=20 candidate responses per question from π on TriviaQA training questions. Zhang et al. (2026) report that the benefit of self-distillation grows with sampling temperature, so the training samples use T=1.5, top-k=20, top-p=0.95—which also widens within-question trace variation—while evaluation uniformly uses T=0.6. A temperature ablation (Appendix F) shows only a modest gain within the bootstrap half-widths. We discard samples for which the answer boundary cannot be located or no probe readout is available. For a retained trace i we read h_i at the pre-answer position and compute

$$p_i = \sigma(w^\top h_i + b), \qquad s'_i = \mathrm{clip}(\mathrm{round}(100 p_i), 1, 100)\,. \tag{1}$$

Because temperature shifts the hidden-state distribution, we reselect the probe on a high-temperature validation set instead of reusing the T=0.6 configuration. Qwen picks the same configuration at both temperatures. Ministral switches from L21 MLP to L22 residual, where reusing the T=0.6 probe would raise macro ECE from 0.054 to 0.100.

4.3 Confidence Relabeling and Expression Coherence

Relabeling changes numbers only. We replace s_i after the final marker with s'_i and rewrite the in-chain confidence numbers, leaving the reasoning steps and the final answer unchanged. This can still produce self-contradictory samples, e.g. a chain of thought saying "definitely true" next to a confidence of 35.

We therefore run a coherence filter. With $\bar{p}_q = N^{-1} \sum_i p_i$ the question's mean probe score, we examine candidates in increasing order of $|p_i - \bar{p}_q|$. A candidate is rejected if its chain of

thought contains multiple or inconsistent confidence numbers. The remaining candidates are rated by Gemma-4-31B-it (Team, 2026), which judges the relabeled triple of reasoning tone, explicit confidence statements, and final score on a 1–5 coherence scale. The first candidate rated 5 enters the training set. The filter edits nothing and carries no answer knowledge, so it enforces text–score consistency without injecting correctness information.

To isolate the role of the teacher signal itself, we construct two controls, EA-SD and SC-SD. EA-SD relabels a trace with the empirical accuracy of its question across the multiple samples. SC-SD replaces the probe score with the answer-cluster frequency $\mathrm{round}(100 f_c)$. Both share the sampling, relabeling, filtering, and training of Probe-SD, differing only in the teacher source.

4.4 Training Objective and Initialization

Standard SFT weights every assistant token equally, so long reasoning traces dilute supervision on the few score tokens. On Qwen3-14B, we therefore apply a weight $\omega > 1$ to the final score span $\mathcal{S}_{\mathrm{score}}$, leaving the targets of reasoning and answer tokens unchanged:

$$\mathcal{L}_{\mathrm{CW}} = -\sum_{t \in \mathrm{assistant}} w_t \log P_\theta(y_t \mid y_{<t}, q), \qquad w_t = \begin{cases} \omega, & t \in \mathcal{S}_{\mathrm{score}}, \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

Ministral repeats the answer and score inside and outside its thinking block, so standard SFT suffices. Although the traces come from the Instruct/Reasoning checkpoint, we initialize the student from the Base checkpoint of the same family and compare Base and Instruct/Reasoning initialization on identical distillation data.

5 Experiments

5.1 Setup

Training and evaluation. Students start from Qwen3-14B-Base and Ministral-3-8B-Base. All SFT methods use the same 10,000 TriviaQA traces and three seeds, with each trained checkpoint evaluated with 10 samples per question. The out-of-domain result is the macro average over EntityQ, NQ-open, and SimpleQA. SFT estimates average the three seeds with 95% question-level bootstrap half-widths (B=5000). Verbal, Answer Probability, and post-hoc recalibration use one generation of the original model. Per-dataset results are in Appendix G.

Data downsampling. The coherence filter keeps one trajectory per question, biasing the retained set toward mid-to-high-confidence trajectories. We downsample to 10,000 traces with a policy that re-aligns the training accuracy to the candidate pool. Re-alignment yields a modest gain only on Ministral, where the shift is largest, and elsewhere it barely affects performance (Appendix H).

Compared methods. Verbal is the original model's verbalized confidence; Answer Probability is the mean generation probability of the answer tokens; Verbal-Iso and Verbal-TS+bias fit isotonic and Platt scaling on TriviaQA and transfer them unchanged out of domain; Surface-Feat fits an L2 logistic readout on surface features—trace length, verbal score, hesitation count, and token probability. EA-SD uses the question's empirical accuracy as its label, SC-SD the answer-cluster frequency; Probe-SD-LRM is Probe-SD with Instruct/Reasoning initialization instead of Base, identical in all else; Probe-SD is the full method.

5.2 Main Results

Probe-SD improves calibration substantially and carries the improvement out of domain. For Qwen, in-domain ECE falls from the Verbal baseline's 0.178 to 0.024 and out-of-domain ECE from 0.542 to 0.113, with AUROC rising from 0.825/0.715 to 0.928/0.783. Ministral moves in the same direction: in-domain ECE from 0.247 to 0.044, out-of-domain from 0.634 to 0.129, and AUROC from 0.669/0.636 to 0.895/0.792 (Table 2). The student transfers the teacher's scale and part of its ranking to held-out test sets under distribution shift.

Post-hoc recalibration behaves as expected. Verbal-Iso beats Probe-SD on the domain it was fit to (Qwen 0.013 versus 0.024, Ministral 0.020 versus 0.044) and degrades to 0.274 and 0.316 off that

Table 2: Main results. TriviaQA is in-domain. Out-of-domain is the macro average of the other three benchmarks. Readout and rescaling methods reuse the Verbal answers and therefore have no separate accuracy.

		In-domain (TriviaQA)				Out-of-domain (macro of 3)			
Model	Method	Acc ↑	ECE ↓	AUROC ↑	Brier ↓	Acc ↑	ECE ↓	AUROC ↑	Brier ↓
Qwen3-14B	Verbal	.729±.024	.178±.020	.825±.017	.184±.017	.258±.013	.542±.013	.715±.017	.468±.011
	Verbal-Iso	—	**.013±.009**	.825±.017	.128±.010	—	.274±.012	.714±.017	.236±.007
	Verbal-TS+bias	—	.047±.011	.825±.017	.131±.010	—	.294±.011	.715±.017	.245±.009
	Answer Probability	—	.251±.024	.636±.029	.253±.023	—	.695±.013	.586±.020	.686±.013
	Surface-Feat	—	.042±.010	.858±.016	.128±.010	—	.332±.012	.697±.017	.285±.007
	EA-SD	**.750±.024**	.093±.011	.878±.014	.119±.011	.278±.013	.227±.009	.747±.014	.220±.008
	SC-SD	.748±.023	.063±.013	.876±.014	.116±.010	**.281±.013**	.236±.010	.753±.013	.222±.007
	Probe-SD-LRM	.729±.024	.031±.011	.926±.011	.096±.009	.258±.012	.135±.009	.774±.017	.154±.006
	Probe-SD	.749±.024	.024±.009	**.928±.011**	**.092±.009**	.279±.013	**.113±.008**	**.783±.015**	**.147±.006**
Ministral-3-8B	Verbal	.664±.025	.247±.024	.669±.015	.263±.021	.213±.012	.634±.012	.636±.016	.575±.010
	Verbal-Iso	—	**.020±.014**	.669±.015	.188±.009	—	.316±.012	.636±.016	.267±.005
	Verbal-TS+bias	—	.067±.016	.669±.015	.193±.008	—	.341±.006	.636±.009	.280±.003
	Answer Probability	—	.302±.025	.569±.027	.306±.025	—	.753±.013	.520±.018	.746±.012
	Surface-Feat	—	.042±.010	.770±.016	.174±.009	—	.329±.011	.673±.017	.276±.006
	EA-SD	.696±.024	.123±.015	.826±.014	.160±.012	.232±.012	.251±.009	.734±.015	.234±.007
	SC-SD	**.698±.025**	.052±.007	.809±.015	.153±.010	**.232±.012**	.169±.008	.722±.014	.192±.005
	Probe-SD-LRM	.644±.026	.064±.011	**.904±.012**	.128±.008	.206±.012	.142±.009	.778±.016	**.147±.006**
	Probe-SD	.694±.025	.044±.009	.895±.014	**.120±.008**	.232±.012	**.129±.008**	**.792±.015**	.148±.005

Table 3: Component ablation on Qwen3-14B. "relabel" is probe relabeling of the verbalized confidence and CW the confidence-weighted loss.

	Components		In-domain				Out-of-domain			
Variant	relabel	CW	Acc	ECE ↓	AUROC ↑	Brier ↓	Acc	ECE ↓	AUROC ↑	Brier ↓
Probe-SD (full)	yes	yes	**.749**	**.024**	**.928**	**.092**	.279	**.113**	**.783**	**.147**
w/o CW	yes	no	.748	.033	.919	.097	.280	.129	.768	.157
w/o relabel	no	no	.747	.159	.814	.173	**.281**	.522	.723	.446

domain, with AUROC unchanged throughout. Probe-SD improves out-of-domain ECE and AUROC together, because it writes the internal teacher's signal into the output rather than rescaling it.

Trace-level labels beat question-level frequency labels. SC-SD is close to Probe-SD in accuracy but markedly worse out of domain on ECE (Qwen 0.236 versus 0.113, Ministral 0.169 versus 0.129). EA-SD, which labels every trace with the question's empirical accuracy, is worse still on calibration (Ministral out-of-domain ECE 0.251 versus 0.129)—the probe's advantage comes from its trace-level internal probability, not merely from using an external correctness signal. Answer Probability is weaker still, with out-of-domain AUROC of only 0.586 and 0.520. The difference is not a global score offset but comes from self-consistency being unable to distinguish the internal states behind traces of the same question.

Base initialization is consistently better than Instruct/Reasoning initialization. Against Probe-SD-LRM, which differs only in initialization, Probe-SD reaches out-of-domain Acc/ECE of 0.279/0.113 versus 0.258/0.135 on Qwen and 0.232/0.129 versus 0.206/0.142 on Ministral. The in-domain accuracy gap is on the order of the bootstrap half-width, whereas the out-of-domain accuracy and calibration advantages are stable. A factual-recall audit attributes most of this advantage to retrieval accessibility, with model-dependent contributions from fact use and quality.

5.3 Component Ablation

Relabeling is decisive (Table 3): removing it raises in-domain ECE from 0.024 to 0.159 and out-of-domain ECE from 0.113 to 0.522, while accuracy changes by under one percentage point. It

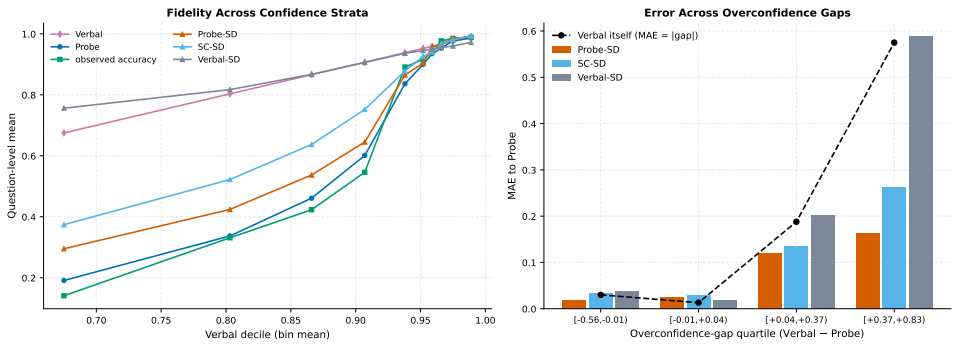

Figure 3: **Each student ends up closest to the teacher it was trained on.** (a) Probe, source verbal score, observed accuracy, and the three students' verbalized confidence on TriviaQA, aggregated by source-verbal decile. (b) Mean absolute error of each student relative to Probe on TriviaQA, by quartile of the overconfidence gap between source verbal and Probe.

changes the confidence scale, not the answers. CW provides a secondary gain—removing it raises out-of-domain ECE to 0.129 and lowers out-of-domain AUROC to 0.768. The small accuracy gain also appears without relabeling, so it comes from Base-initialized SFT.

6 Analysis

Probe-SD primarily improves calibration. Base initialization adds a modest accuracy gain. We ask whether the student inherits the teacher's calibration shape, why trace-level Probe outperforms question-level self-consistency, where the Base-initialization gain comes from, and how internalized confidence behaves under test-time aggregation.

6.1 The Student Inherits the Teacher's Calibration Shape

On Qwen3-14B, Probe-SD, SC-SD, and Verbal-SD share the same source sampling, filtering, and Base initialization, and differ only in the confidence target—the frozen probe probability for Probe-SD, the self-consistency cluster frequency for SC-SD, and the source model's own verbalized score for Verbal-SD. Each student ends up closest to its teacher (Figure 3a). On TriviaQA, question-level correlation with Probe is 0.913 for Probe-SD, 0.872 for SC-SD, and 0.767 for Verbal-SD, versus 0.794 for the source verbal score. Probe-SD and SC-SD therefore track Probe more closely than the original score. The same ordering holds across all four datasets (Appendix I). The gap is largest where the source is overconfident (Figure 3b): MAE to Probe in the top versus bottom quartile of the source–Probe gap is 0.164/0.018, 0.263/0.034, and 0.588/0.037 for Probe-SD, SC-SD, and Verbal-SD, respectively. Probe-SD does not uniformly lower scores. It corrects traces where the verbal channel and internal teacher disagree. The w/o-relabel ablation returns near the original verbal calibration, showing that the training label determines the student's confidence structure.

6.2 Self-Consistency Fails Where It Disagrees With the Probe

Probe and self-consistency disagree on a minority of questions, and that minority carries the gap between Probe-SD and SC-SD. We bin questions by the difference between mean Probe score and SC frequency. The SC frequency is shared within a cluster, whereas Probe varies by trace. We compare residuals (score minus observed accuracy) across bins.

The two tails carry the difference (Appendix I). Where Probe$\ll$SC (2.3% of questions, accuracy 0.635), self-consistency overshoots by $+0.219$ and Probe undershoots by -0.165. Where Probe$\gg$SC (13.4% of questions, accuracy 0.766), self-consistency undershoots badly at -0.381 while Probe sits near zero ($+0.078$). Probe-SD follows its teacher at both ends (-0.100 and $+0.023$). SC-SD fails to reproduce its label's direction where that label is most misleading. In that region, answer frequency is too noisy a label for the student, whereas Probe retains resolution.

6.3 Base Initialization Improves Factual Retrieval

The accuracy gain from Base initialization is not a probe effect. To locate it, we compare Probe-SD with the original Instruct/Reasoning model on matched TriviaQA questions using independent traces. For each trace, we use string matching to check whether the gold answer or an alias appears in the chain of thought, then measure accuracy conditional on that hit. This separates retrieval from post-hit use and the miss branch. A fact audit checks whether extracted facts are true and answer-bridging, while ten-sample hit counts classify questions as easy, medium, or hard (Appendix J).

For Qwen, Probe-SD gains 2.36 points over Instruct, with retrieval accounting for 2.21 (93.7%). Conditional accuracy after a hit is unchanged (0.8915 versus 0.8912), as are true bridging facts (4.95 versus 4.98). Ministral shows the same pattern: 2.60 of its 3.56-point gain (73.0%) comes from retrieval, with smaller usage and miss-branch contributions. The audit finds fact-quality differences (true bridging facts 4.81 versus 4.32, hallucination 26.7% versus 27.6%).

Difficulty counts locate the gain in the marginal region. Each model has exactly one easy-to-hard transition, while most hard questions remain hard (124/147 for Qwen, 141/192 for Ministral). The main movement is in the middle band: Qwen and Ministral respectively move 51 and 73 medium questions to easy versus 9 and 21 to hard. Base initialization thus raises access to latent factual knowledge at the margin without shifting the boundary of retrievable knowledge.

6.4 SFT Alone Produces Test-Time Confidence Behavior

We adapt two diagnostics from RLCR (Damani et al., 2025) to test whether related test-time confidence behaviors can arise from supervised finetuning alone. We do not train an online-RL baseline, and differences in model, data, and output format preclude a controlled comparison with RLCR. Our results therefore show that these behaviors can arise without RL in our setting. They do not establish equivalence to an RL-trained model. For Qwen3-14B on TriviaQA, internalized confidence is more useful for weighting than selecting, mirroring the probe's behavior in §3.3: at $N{=}16$, weighted voting reaches 79.05%, versus 78.59% for majority voting and 77.75% for max-confidence selection, while the oracle reaches 87.19%. The three non-oracle methods largely saturate past $N \geq 8$. Averaging independently sampled traces that agree on an answer lowers Brier from 0.0886 at $K{=}1$ to 0.0755 at $K{=}8$, near the group-mean bound of 0.0748, while ECE remains 0.033–0.036: it removes variance, not within-bin bias.

At $K{=}16$, 37.4% of questions have $\sigma < 0.02$ and 13.1% have $\sigma \geq 0.20$. For mutually exclusive answers, the original Instruct model's confidences sum to 2.287 on average and exceed 2 on 33.3% of questions, versus 1.383 and 13.3% for Probe-SD. Together with low ECE, these checks rule out mere saturated overconfidence. More details appear in Appendix K.

7 Conclusion and Limitations

Across two model families and four factual QA benchmarks, a linear probe at the pre-answer position is well calibrated but nearly ties majority voting as an answer selector, with both far below the oracle. **On factual QA the value of internal states is calibration, not answer knowledge that decoding failed to extract.** The method follows from that separation. Rather than serving the probe at decoding time, we apply it offline as a trace-level teacher: the probe scores the model's own traces, the stated confidence is overwritten with the probe score, and the base checkpoint of the same family is finetuned on the result, so only the model remains at deployment. Removing the relabeling step confirms the separation from the training side: calibration collapses and accuracy does not move. In our setting, supervised finetuning alone suffices to internalize the probe-derived confidence signal, without online RL or a test-time judge.

Two limits bound the claim: we evaluate two open-weight reasoning families (up to 14B parameters) on four factual QA benchmarks with short, checkable answers, leaving larger or closed models, reasoning-intensive tasks, and long-form or subjective tasks untested. Building the training data requires white-box access to hidden states and per-model probe configuration, so the pipeline cannot be built from a text-only interface. Appendix L covers these and the remaining limitations. Looking ahead, whether the calibration–selection separation holds at larger scales, and whether calibrated confidence can drive model cascading, are natural next steps.

Ethics Statement

This work studies the calibration of confidence reported by language models on factual question answering, using publicly available benchmarks and open-weight models. No human subjects, private data, or personally identifying information are involved. Better-calibrated confidence is intended to support honest communication of uncertainty and safer downstream thresholding. We note one dual-use consideration: a model that states well-calibrated confidence may be trusted more readily, so residual miscalibration on out-of-distribution or low base-rate inputs could be more consequential than the same error from a visibly overconfident model.

Reproducibility Statement

All datasets and base models used are publicly available. The experimental protocol and supporting analyses are documented in Appendices A–L.

AI Use Statement

We used generative AI tools as research artifacts inside the experimental pipeline: answer extraction and correctness grading (Qwen3-32B-Instruct), coherence filtering of relabeled traces (Gemma-4-31B-it), and chain-of-thought fact extraction and labeling for the factual priming audit (glm-5-2). All of these are documented in the appendices, and the grader was validated against a human audit. We additionally used generative AI assistance for writing polish and for code implementation, all of which was reviewed by the authors. We did not use generative AI to generate research ideas, experimental results, or claims. We take responsibility for the final content of this work.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying, 2023. URL https://arxiv.org/abs/2304.13734.

Catarina G Belem, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. Perceptions of linguistic uncertainty by language models and humans. *arXiv preprint arXiv:2407.15814*, 2024. URL https://arxiv.org/abs/2407.15814.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. URL https://api.semanticscholar.org/CorpusID:122906757.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2212.03827.

Jon-Paul Cacioli. Making llms say what they know: Probe-targeted fine-tuning for verbal confidence calibration, 2026. URL https://zenodo.org/records/20436841. Zenodo preprint (submitted to JAIR), not on arXiv.

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, 2025. URL https://arxiv.org/abs/2507.16806.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Siddartha Devic, Charlotte Peale, Arwen Bradley, Sinead Williamson, Preetum Nakkiran, and Aravind Gollakota. Trace length is a simple uncertainty signal in reasoning models, 2025. URL https://arxiv.org/abs/2510.10409.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. URL https://doi.org/10.1038/s41586-024-07421-0.

Tim Franzmeyer, Archie Sravankumar, Lijuan Liu, Yuning Mao, Rui Hou, Sinong Wang, Jakob N. Foerster, Luke Zettlemoyer, and Madian Khabsa. High accuracy, less talk (HALT): Reliable LLMs through capability-aligned finetuning, 2026. URL https://arxiv.org/abs/2506.04051.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in LLMs, 2025. URL https://arxiv.org/abs/2503.15299.

Zorik Gekhman, Roee Aharoni, Eran Ofek, Mor Geva, Roi Reichart, and Jonathan Herzig. Thinking to recall: How reasoning unlocks parametric knowledge in LLMs, 2026. URL https://arxiv.org/abs/2603.09906.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Junyu Guo, Shangding Gu, Ming Jin, Costas Spanos, and Javad Lavaei. LLMs should express uncertainty explicitly, 2026. URL https://arxiv.org/abs/2604.05306.

L. Haas, G. Yona, G. D'Antonio, S. Goldshtein, and D. Das. SimpleQA-Verified: A reliable factuality benchmark to measure parametric knowledge, 2025. URL https://arxiv.org/abs/2509.07968.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations, 2025. URL https://arxiv.org/abs/2503.14477.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017. URL https://arxiv.org/abs/1705.03551.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 (ICLR)*, 2023. URL https://arxiv.org/abs/2302.09664.

Dharshan Kumaran. Reported confidence in llms tracks commitment more than correctness, 2026. URL https://arxiv.org/abs/2606.29490.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7:452–466, 2019. URL https://aclanthology.org/Q19-1026/.

Ke Li, Chongzhe Zhang, Zifan Zeng, Feng Liu, Qunli Zhang, and Zheng Hu. Calibrating overconfidence without sacrificing confidence: Probe-conditioned head intervention for LLMs, 2026. URL https://arxiv.org/abs/2606.09876.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL https://arxiv.org/abs/2306.03341.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. *arXiv preprint arXiv:2508.18847*, 2025. URL https://arxiv.org/abs/2508.18847.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022. URL https://arxiv.org/abs/2205.14334.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026a. URL https://arxiv.org/abs/2601.08584.

Gabrielle Kaili-May Liu, Avi Caciularu, Gal Yona, Idan Szpektor, and Arman Cohan. Reinforcement learning with metacognitive feedback elicits faithful uncertainty expression in LLMs, 2026b. URL https://arxiv.org/abs/2606.32032. https://github.com/yale-nlp/RLMF.

Xin Liu, Farima Fatahi Bayat, and Lu Wang. Enhancing language model factuality via activation-based confidence calibration and guided decoding, 2024. URL https://arxiv.org/abs/2406.13230.

Zhengzhao Ma, Xueru Wen, Boxi Cao, Yaojie Lu, Hongyu Lin, Jinglin Yang, Min He, Xianpei Han, and Le Sun. Decoupling reasoning and confidence: Resurrecting calibration in reinforcement learning from verifiable rewards. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2603.09117.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know?, 2025. URL https://arxiv.org/abs/2506.18183.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. URL https://arxiv.org/abs/2012.14983.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. Towards fully exploiting LLM internal states to enhance knowledge boundary perception, 2025. URL https://arxiv.org/abs/2502.11677.

OpenAI. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. URL https://arxiv.org/abs/2412.16720.

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, 2025. URL https://arxiv.org/abs/2410.02707.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Bhaktipriya Radharapu, Eshika Saxena, Kenneth Li, Chenxi Whitehouse, Adina Williams, and Nicola Cancedda. Calibrating llm judges: Linear probes for fast and reliable uncertainty estimation, 2025. URL https://arxiv.org/abs/2512.22245.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL https://arxiv.org/abs/2109.08535.

Giovanni Servedio, Alessandro De Bellis, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. Are the hidden states hiding something? testing the limits of factuality-encoding capabilities in LLMs, 2025. URL https://arxiv.org/abs/2505.16520.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, et al. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2312.06585.

Paul Stangel, David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Kamilia Zaripova, Matthias Keicher, and Nassir Navab. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models, 2025. URL https://arxiv.org/abs/2503.02623.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. LACIE: Listener-aware finetuning for calibration in large language models, 2024. URL https://arxiv.org/abs/2405.21028.

Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Philip Torr, and Chang Xu. Revisiting uncertainty estimation and calibration of large language models, 2025. URL https://arxiv.org/abs/2505.23854.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://arxiv.org/abs/2305.14975.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Karina Nguyen, Hyung Won Chung, et al. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024. URL https://arxiv.org/abs/2411.04368.

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, 2024. URL https://arxiv.org/abs/2306.13063.

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, 2024. URL https://arxiv.org/abs/2405.20974.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for LLMs, 2024. URL https://arxiv.org/abs/2412.14737.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*, 2024. URL https://arxiv.org/abs/2405.16908.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence, 2025. URL https://arxiv.org/abs/2505.14489.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, KDD '02, pp. 694–699, 2002. URL https://dl.acm.org/doi/10.1145/775047.775151.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.14465.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification, 2025. URL https://arxiv.org/abs/2504.05419.

Ruixiang Zhang, Richard He Bai, Huangjie Zheng, Navdeep Jaitly, Ronan Collobert, and Yizhe Zhang. Embarrassingly simple self-distillation improves code generation, 2026. URL https://arxiv.org/abs/2604.01193. Apple. https://github.com/apple/ml-ssd.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A Scalable lightWeight Infrastructure for Fine-Tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Thomas Zollo, Jimmy Wang, and Richard Zemel. Unsupervised confidence calibration for reasoning llms from a single generation, 2026. URL https://arxiv.org/abs/2604.19444.

A Experimental Setup

This appendix expands the setup summarized in §3.

Data and models. We use four factual QA benchmarks with short, automatically checkable answers: TriviaQA (Joshi et al., 2017), EntityQuestions (EntityQ) (Sciavolino et al., 2021), NQ-open (Kwiatkowski et al., 2019), and SimpleQA-Verified (Wei et al., 2024; Haas et al., 2025), with 1,000 test questions each and 10 traces per question, on Qwen3-14B (Qwen Team, 2025) (40 layers) and Ministral-3-8B-Reasoning-2512 (Liu et al., 2026a) (34 layers), sampled at T=0.6, top-k=20, top-p=0.95. Probes are trained on 10,000 TriviaQA training questions, so TriviaQA is in-domain and the other three are out-of-domain throughout. All test sets are held out from configuration selection. Qwen3-32B-Instruct extracts and grades answers. Its reliability is audited against human labels in Appendix D.

Readout and probe. The prompt asks for a complete chain of thought followed by `##FINAL ANSWER##\n{answer}, {score}`, with the score an integer from 1 to 100 normalized to $[0, 1]$ (Appendix B). Among seven candidate extraction positions, the *last token before the answer*—after the chain of thought has ended and before the answer begins—was stable across settings and always close to optimal, so all experiments use it. Per layer we fit an ℓ_2-regularized logistic regression $p_{\text{probe}} = \sigma(w^\top h + b)$ on the residual stream and on the MLP output separately, with cross-entropy against trace correctness. Validation data alone select layer 24 residual with C=0.005 for Qwen3-14B and layer 21 MLP with C=2.0 for Ministral-3-8B, at 60.0% and 61.8% relative depth. The candidate positions, the full grid, and the selection criterion are in Appendix C.

Metrics. Let $y_i \in \{0, 1\}$ indicate whether trace i is correct and let $p_i \in [0, 1]$ be the confidence assigned to that trace. Unless stated otherwise, confidence metrics are computed over the *scorable subset*: traces for which the source produces a valid confidence score. Expected calibration error (ECE) (Naeini et al., 2015) partitions scores into 10 equal-width bins and computes

$$\text{ECE} = \sum_{b=1}^{10} \frac{|I_b|}{n} \left| \frac{1}{|I_b|} \sum_{i \in I_b} p_i - \frac{1}{|I_b|} \sum_{i \in I_b} y_i \right|,$$

where I_b is the set of scored traces in bin b. ECE is zero for perfect empirical calibration and is lower when stated confidence more closely matches observed accuracy; because it depends on binning, we use the same bins for every method. The Brier score is the mean squared probabilistic error, $n^{-1} \sum_i (p_i - y_i)^2$. It is a strictly proper score that reflects both calibration and discrimination, with lower values better. AUROC is the probability that a randomly chosen correct trace receives a higher score than a randomly chosen incorrect trace (with ties handled in the standard rank-based manner). It measures ranking rather than probability-scale calibration: 0.5 is chance and 1 is perfect discrimination. Coverage is the number of traces with a valid confidence score divided by the total number of generated traces; higher is better. Thus a source may have low ECE but weak AUROC, or strong AUROC but a miscalibrated scale, so we report both together with Brier and Coverage. Appendix E.1 tests whether differences in the scorable subset affect the comparison.

B PROMPTS AND SAMPLING SETTINGS

Sampling. Training data are generated at T=1.5, top-k=20, top-p=0.95 with 20 samples per question, identically for both models. Evaluation uses T=0.6, top-k=20, top-p=0.95 with 10 samples per question. The verbalized score is parsed as `,\d{1,3}` after the last `##FINAL ANSWER##` marker and normalized to $[0, 1]$. For Ministral-3-8B we always take the last marker with `rfind`, because the model rehearses the marker inside its thinking block. All prompts ask the model to reason step by step about both the answer and its confidence and then emit `##FINAL ANSWER##\n{answer}, {score}` with the score an integer from 1 to 100.

Qwen3-14B prompt. Used unchanged for both training-data generation and evaluation.

```
[system] You are a precise assistant. Think step by step about how confident you
         are (score 1-100), then give your answer.
[user]   Format: ##FINAL ANSWER##\n{answer}, {score}

         Question: {question}
```

Ministral-3-8B-Reasoning-2512 prompt. We keep the model's own default system prompt (the structured text + `thinking(closed)` + text template that triggers `mistral-common` reasoning mode) and append the format instruction to the user turn.

```
[system] # HOW YOU SHOULD THINK AND ANSWER

         First draft your thinking process (inner monologue) until you
         arrive at a response. Format your response using Markdown, and use
         LaTeX for any mathematical equations. Write both your thoughts and
         the response in the same language as the input.

         Your thinking process must follow the template below:
         [THINK]Your thoughts and/or draft, like working through an
         exercise on a scratch paper. Be as casual and as long as you want
         until you are confident to generate the response to the
         user.[/THINK]Here, provide a self-contained response.
[user]   Question: {question}

         Answer the question. Then append your confidence (1-100) on the
         same line.

         Format:
         ##FINAL ANSWER##\n{answer}, {score}
```

B.1 CONSISTENCY SEARCH AND RELABELING

Candidates are visited in the order of §4.3 and each one is passed through a three-step filter chain, stopping at the first candidate that survives. (1) **Score extraction check** (judge with thinking off): scan the chain of thought for every explicit confidence number. The candidate passes if no score is found, or if exactly one is found and it agrees with the score after `##FINAL ANSWER##`. It is rejected if several scores appear or if any contradicts the final one. (2) **Mechanical relabeling** (pure string operations, no prompt): the verbalized score in `##FINAL ANSWER##\n{answer}, {old_score}` is replaced by $s_{\text{new}} = \max(1, \min(100, \text{round}(p_{\text{probe}} \times 100)))$, and every 1–100 number in the chain of thought that is adjacent to a confidence word ("confident", "sure", "certain", "likely", "guess", ...) is replaced by the same value. Dates, counts and statistics are skipped. Ministral typically restates answer and score inside the `[THINK]` block before the final marker, so relabeling rewrites both occurrences and the rehearsal is kept rather than trimmed. (3) **Coherence check** (judge with thinking on): the relabeled response is rated 1–5 for end-to-end consistency between chain of thought, confidence wording and final score. A rating of 5 selects the candidate and stops the search. Otherwise the next candidate is tried. Relabeling must precede the coherence check, because what step 3 evaluates is the self-consistency of the *relabeled* response. Both judge steps use Gemma-4-31B-it (Team, 2026).

Step 1: score extraction (thinking off).

```
[system] You are a confidence-score extractor. Your task is to scan a model's
  thinking chain and identify every explicit confidence score (1-100) the model
  states about its own answer.
```

```
  A confidence score is a number 1-100 that answers "how confident am I?" - it
  represents the model's self-assessed likelihood of being correct. Look for
  patterns like:
  - "score X", "score of X", "score: X", "my score is X"
  - "I'm X% confident/sure/certain", "I am X% sure"
  - "confidence: X", "confidence is X", "confidence level X"
  - "about X% confident", "around X%", "maybe X", "probably X%"
  - "give it a X", "rate it X", "put it at X"

  Do NOT count: years (19xx, 20xx), dates, quantities of things, statistics, or
  any number that is NOT describing the model's own confidence level. If the
  thinking says "score is 1-100" (describing the scale), that is NOT a
  confidence score.

[user]
  ## Thinking chain
  {thinking}

  ## Task
  Identify EVERY distinct confidence score the model explicitly states in this
  thinking chain. Return the count and the complete list.

  Return ONLY a JSON object (no markdown, no extra text):
  {"count": <integer>, "scores": [<integer, ...>]}
```

Step 3: coherence check (thinking on).

```
[system] You are a coherence evaluator. Your task is to judge whether a model's
  entire response - from the beginning of its thinking process to its final
  answer - forms a harmonious, consistent whole where the confidence it
  expresses and the scores it gives all align with each other.

  The model's response contains a thinking chain inside <think>...</think>,
  followed by a final answer in the format "##FINAL ANSWER##\n{answer},
  {score}". The score is a number from 1 to 100 on a confidence scale:
  - 1 means "I am completely guessing, I have no idea."
  - 100 means "I am absolutely certain, beyond any doubt."
  A score of 5 or 10 is extremely low - it means the model is admitting it
  barely knows.

  Your evaluation scope is the ENTIRE response, not just the final score. Judge
  the thinking process itself - how the model arrives at its answer matters, not
  just where it ends up. A model that reasons uncertainly but then gives a high
  score is still contradictory, even if its final conclusion is correct.

  You must check:

  1. **End-to-end coherence**: From the first sentence of thinking to the last,
  the model's expressed certainty should be internally consistent. If it starts
  confident, wavers, then becomes confident again - that is not coherent. If it
  sounds sure in some sentences and clueless in others - that is not coherent.
  The whole process must flow naturally toward the final score.

  2. **Score-tone alignment everywhere**: Every time the model mentions a
  confidence level - whether a numeric score inside the thinking ("I'm about X%
  sure"), a verbal expression ("pretty confident", "not sure", "definitely"), or
  the final score - check that the number and the surrounding tone match. This
  is a hard requirement:
     - Confident words ("definitely", "pretty confident", "I'm sure",
       "certain", "no doubt") + any score below 40 -> ALWAYS at most score 2.
     - Uncertain words ("not sure", "guessing", "no idea", "don't know", "not
       confident", "not certain") + any score above 80 -> ALWAYS at most
       score 2.

  3. **Scale awareness**: You must recognize where numbers fall on the 1-100
  scale. A score of 3, 5, 7, 11, 15 - these are all very low. If the thinking
  calls such a score "high" (e.g., "the score would be high, like 5"), that is a
  direct contradiction. A low score (below 40) should only appear alongside
  hesitant, uncertain reasoning. A high score (above 80) should only appear
  alongside confident, decisive reasoning. Mid-range scores (40-79) should
  appear alongside moderately confident reasoning with some hedging.

  4. **No contradictions allowed for score 5**: Score 5 is reserved for perfect
  alignment. If ANY part of the thinking contradicts ANY part of the confidence
  expression - internal or final - it cannot be a 5.

  Rate the coherence on a 1-5 scale:
  1 = Severe conflict: direct contradiction between the thinking's tone and any
      score mentioned, or violation of the hard requirement in point 2.
```

```
  2 = Moderate conflict: significant mismatch, mixed signals that undermine
      alignment, or violation of the hard requirement in point 2.
  3 = Neutral: ambiguous, inconsistently coherent, thinking and scores feel
      loosely related.
  4 = Good alignment: the thinking generally supports the scores throughout,
      with only minor imperfections. No hard requirement violations.
  5 = Perfect alignment: every part of the thinking is tonally consistent with
      every score mentioned, from start to finish. The entire response is a
      harmonious, unified whole.

  Return ONLY a JSON object (no markdown, no extra text):
  {"score": <1-5 integer>, "reason": "<one concise sentence>"}

[user]
  ## Question
  {question}

  ## Model Response (full)
  {raw_response}

  ## Task
  Evaluate the coherence of this model response from start to finish. The
  confidence scale is 1-100, where 5 means near-zero confidence. Check every
  confidence expression - internal scores, verbal tones, and the final score -
  they must all align.

  Return ONLY a JSON object:
  {"score": <1-5 integer>, "reason": "<one concise sentence>"}
```

B.2 Audit of Mechanical Relabeling

A rating of 5 is a strict quality gate: a candidate enters the training set only if the relabeled chain of thought, the confidence wording and the final score are consistent end to end. Mechanical relabeling never touches the answer or non-confidence text. It only synchronizes explicit self-confidence numbers. The audit pool contains 14,951 SFT training traces, of which 493 candidates have a verbal–probe gap of 18–22 points. Table 4 lists five of them (Qwen3-14B, T=0.6, chain-of-thought length filtered to 300–1500 tokens). Grading is by Qwen3-32B-Instruct and the replacement score is $\mathrm{round}(p_{\mathrm{probe}} \times 100)$.

Table 4: Targeted audit of mechanical relabeling. Candidates were selected by gap, so the table verifies the replacement mechanism and the direction of the label, not the average size of the correction.

qid	idx	Gold	Model answer	Graded	Verbal → new	p_{probe}
qb_8463	10	Times Square	Times Square	correct	95 → 73	.727
qw_10780	17	USA	United States of America	correct	95 → 74	.742
jp_2478	17	Elizabeth	Elizabeth	correct	99 → 81	.806
sfq_19580	11	Chris Huhne	Chris Huhne	correct	95 → 75	.749
sfq_4960	0	Glasgow	York	**incorrect**	85 → 66	.665

The audit deliberately includes both correct and incorrect samples, showing that relabeling does not merely lower high scores on questions the model got right: row 5 lowers 85 to 66 while the answer remains the incorrect *York*.

Example 1 (correct answer, 95 → 73). *Which New York landmark is known as 'The Crossroads of the World'?* (gold: Times Square). The chain of thought contains "...maybe it's referring to a different location? ... Like Grand Central Terminal? No, I don't think so. ... I think my confidence here is about 95%." The final answer `Times Square, 95` is relabeled to `Times Square, 73` and the in-chain 95% is replaced by 73%. The answer and all other reasoning text are unchanged.

Example 2 (incorrect answer, 85 → 66). *In 1990, which was the first UK city to be a European City of Culture?* (gold: Glasgow, the model answers York). The chain of thought contains "I think the first UK city to be selected was York in 1990. But I'm not 100% sure. ... My confidence is around 85% because I've heard this before, but I'm not entirely certain if there was another city before it." The final answer `York, 85` becomes `York, 66` and the in-chain 85% becomes 66%.

The answer is still the incorrect *York*: the probe acts here as a calibration signal, not an answer selector.

C PROBE GRID SEARCH AND CONFIGURATION SELECTION

Features and training. The probe input is the hidden-state readout $\mathbf{x} \in \mathbb{R}^d$ at the pre-answer position—the token preceding the first answer token, i.e. immediately after the newline following the last `##FINAL ANSWER##`—with d=5120 for Qwen3-14B and d=4096 for Ministral-3-8B. That position was chosen from seven candidates: the first token after the chain of thought ends, the last token before the answer, the first answer token, the last answer token, the mean over all answer tokens, the token preceding the score, and the score token itself. It was the only one stable across models, datasets and temperatures while always scoring close to the best candidate, so §A fixes it for every experiment. We fit ℓ_2-regularized logistic regression

$$\min_{\mathbf{w},b} \frac{1}{N}\sum_{i=1}^{N} \ell\big(y_i, \sigma(\mathbf{w}^\top \mathbf{x}_i + b)\big) + \frac{1}{C}\|\mathbf{w}\|_2^2,$$

where $y_i \in \{0,1\}$ is the correctness label of sample i as decided by the Qwen3-32B-Instruct grader. Every configuration is trained with lbfgs and `max_iter=2000` under three random seeds $\{1,2,3\}$, and the reported probe score is the ensemble mean $p_{\text{probe}} = \frac{1}{3}\sum_{s=1}^{3} \sigma(\mathbf{w}_s^\top \mathbf{x} + b_s)$.

Search space and selection. We take the Cartesian product of all layers, both readouts and eleven regularization strengths (Table 5). Probes are trained on model-generated samples from the TriviaQA training split (Qwen3-14B: 9,987 samples, positive rate 0.735, Ministral-3-8B: 9,743 samples, positive rate 0.665), disjoint from the validation and test splits. Selection is performed entirely on TriviaQA, EntityQ, and NQ-open validation data with the composite rank described in Appendix F. The selected configurations are reported in Table 16 and only then evaluated on the four held-out test sets. SimpleQA has no validation split and does not participate in selection. Configurations with $C \geq 5$ raise lbfgs convergence warnings and are excluded from consideration.

Table 5: Probe grid search space.

Dimension	Qwen3-14B	Ministral-3-8B
Layers	40	34
Regularization C	$\{0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1.0, 2.0, 5.0\}$	
Readout	{residual, MLP}	
Total configurations	1,760	1,496

Layer and readout. The all-layer sweep shows correctness- and calibration-related signal already readable in fairly early layers, strengthening with depth and forming a stable plateau over the middle of the network. Layers close to the output overfit more easily out of domain. Both readouts carry the calibration signal—Qwen's winner is the residual stream, Ministral's is the MLP output—and Table 6 shows the two are essentially equivalent. On the macro average Qwen scores the same selection accuracy either way (.405) with AUROC .848 versus .849, and residual is preferred only on the slightly lower ECE (.046 versus .048). Ministral's macro selection accuracy is .353 versus .350 with AUROC .821 versus .819. In both models the residual–MLP difference is far smaller than the Probe–Verbal gap, so the conclusions do not depend on a single definition of the internal tensor.

Best-of-N selection scaling at N=32. To rule out a single-dataset artifact we extend the best-of-N analysis of §3.3 to both models and all four datasets, sampling 32 independent traces per question at T=0.6 and comparing selection by maximum probe score, lexically normalized majority voting, and the oracle pass@N.

Selection accuracy forms a broad, flat plateau across layers (Figure 2a): 0.748–0.778 over all 40 Qwen layers on TriviaQA and 0.431–0.471 on NQ-open, peaking at L27 and L23 with under 5 points of variation, so deeper representations do not keep releasing new answer knowledge. Panel (b) is more discriminative: on TriviaQA at N=16, Max-Probe reaches 0.778 and majority voting 0.768 while the oracle is already at 0.860. At N=32 the three are 0.778, 0.771, and 0.879.

Max-Probe and majority voting trade places with small margins throughout: Qwen is slightly ahead on TriviaQA and EntityQ (.778 versus .771, .315 versus .306) and slightly behind on NQ-open and

Table 6: Residual versus MLP readout, per dataset and macro-averaged over the four test sets. (i) is the validation-selected winner for that model, (ii) the best configuration of the other readout type (Qwen: (i) residual L24 C=0.005, (ii) MLP L24 C=0.01, Ministral: (i) MLP L21 C=2.0, (ii) residual L21 C=1.0). These test-set numbers describe configurations already fixed on validation data. They are not used for selection. Bold marks the better of the two within a dataset.

Dataset	Readout	sel. Acc ↑	ECE ↓	AUROC ↑	Brier ↓
Qwen3-14B					
TriviaQA	(i) residual	.772	**.015**	.948	.081
	(ii) MLP	**.779**	.019	.948	.081
EntityQ	(i) residual	**.310**	**.014**	**.913**	**.097**
	(ii) MLP	.302	.026	.909	.100
NQ-open	(i) residual	**.463**	.078	.826	.177
	(ii) MLP	.462	**.065**	**.829**	**.174**
SimpleQA	(i) residual	**.076**	**.075**	.704	**.076**
	(ii) MLP	.075	.082	**.708**	.078
Macro	(i) residual	.405	**.046**	.848	.108
	(ii) MLP	.405	.048	**.849**	.108
Ministral-3-8B					
TriviaQA	(i) MLP	**.718**	.014	.917	.105
	(ii) residual	.714	**.013**	**.918**	**.104**
EntityQ	(i) MLP	.225	.024	**.909**	**.091**
	(ii) residual	.225	**.015**	.901	.095
NQ-open	(i) MLP	**.401**	**.108**	**.816**	**.188**
	(ii) residual	.392	.126	.813	.194
SimpleQA	(i) MLP	.068	.146	.640	.116
	(ii) residual	**.069**	**.127**	**.645**	**.109**
Macro	(i) MLP	**.353**	.073	**.821**	.125
	(ii) residual	.350	**.070**	.819	.125

Table 7: Best-of-N accuracy at N=32. Bold marks the better of Max-Probe and majority voting. Oracle is the theoretical ceiling.

Model	Dataset	Oracle (pass@N)	Max-Probe	Majority vote
Qwen3-14B	TriviaQA	.879	**.778**	.771
	EntityQ	.447	**.315**	.306
	NQ-open	.695	.458	**.492**
	SimpleQA	.246	.074	**.077**
Ministral-3-8B	TriviaQA	.844	.717	**.728**
	EntityQ	.363	.231	**.237**
	NQ-open	.631	**.406**	.405
	SimpleQA	.262	.072	**.078**

SimpleQA (.458 versus .492, .074 versus .077). Ministral is overtaken on three of four datasets and leads slightly on NQ-open. Both are far below the oracle on every dataset and the gap widens with N (for Qwen on TriviaQA it grows from .024 at N=2 to .101 at N=32).

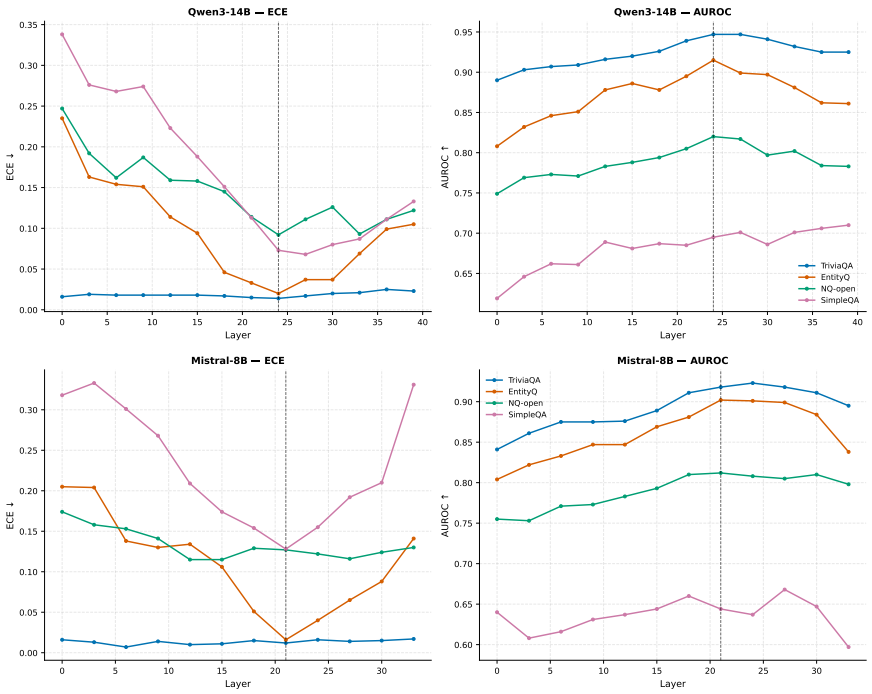

Figure 4: Per-layer ECE and AUROC (§3.2). Top: Qwen3-14B. Bottom: Ministral-3-8B. Each line is one dataset and the dashed vertical line marks the best layer.

D GRADING AND EXTRACTION RELIABILITY

Qwen3-32B extracts the model's final answer and decides its correctness. To assess this pipeline we sampled 200 already-graded question–answer pairs across 8 model columns and 4 datasets, pre-screened them with string matching to raise the yield of anomalies, then manually reviewed all 34 flagged cases plus 10 additional unflagged cases. The manual review inspects the question, the reference answer, the raw model output, the extraction result and the grading rationale. String matching is only a screening device and is never treated as ground truth.

Table 8: Extraction and grading reliability under a 200-question human audit.

Metric	String pre-screen	After human review
Extraction accuracy	92.0% (184/200)	**99.5% (199/200)**
Grading accuracy (answerable)	93.8% (181/193)	**99.5% (198/199)**
Grading precision	83.1%	**98.7% (77/78)**
Grading recall	100%	**100% (77/77)**
True grading errors	—	**1/200 (0.5%)**

Extraction and grading both end at 99.5%, but on different denominators: extraction is 199/200, whereas grading accuracy is computed on the 199 samples the review confirmed to be answerable (198/199). The grading denominator moves from 193 at the pre-screen stage to 199 because the review re-partitioned abstentions, valid synthesized answers, and genuinely gradable responses.

Table 9: Human classification of the audit sample. Left: extraction outcomes over all 200 cases. Right: attribution of the cases the string pre-screen flagged as grading anomalies.

Extraction outcome	n	Verdict	Flagged grading case	n	Verdict
Exact or prefix match	175	faithful	More specific answer contains gold	5	correct
Substring match	9	correct distillation	Alias not in gold list	3	correct
Surface mismatch, valid	5	valid synthesis	Functional description $\equiv$ entity	2	correct
Empty: explicit abstention	10	correct empty	Output contains gold substring	3	correct
Empty: answer was given	1	**true error**	Subset or containment relation	4	correct
			Diacritic difference	1	correct
Total	200	199 valid	**True grading error**	1	false accept

The string pre-screen underestimates extraction quality. All ten empty outputs come from responses that explicitly said "Unknown", "I don't know" or "not available", or that rejected the premise of the question, so they are correctly recognized abstentions rather than missed answers. Five surface-mismatch cases come from responses that ignored the single-answer format. The extractor distilled an entity or a list from long answers about the inventor of the CPU, the oldest house in the United States, a Revolutionary War battle, the start and end dates of the Soviet Union, and the next Olympic Games, and the review confirmed all five as valid syntheses. The single true extraction error is a response that stated "Singapore Island, 55" for which the extractor returned an empty string. Whether or not the answer was right, an explicitly stated answer should never be discarded.

Abstentions also raise a protocol difference between `INCORRECT` and `NOT_ATTEMPTED`. Our end-to-end convention keeps explicit abstentions in the accuracy denominator and counts them as errors, whereas the strict SimpleQA taxonomy would label them `NOT_ATTEMPTED`. This changes the name of the error type but not the end-to-end accuracy we report.

Of the 19 suspicious grading cases, 18 are false alarms produced by overly strict string rules. The single true grading error is a question about a 2010 South Africa World Cup stadium where the model substituted Moses Mabhida Stadium for Cape Town Stadium and the grader mistakenly treated Moses Mabhida Stadium and Nelson Mandela Bay Stadium as aliases. These are different stadiums, so the response should have been marked incorrect. All 10 additional unflagged cases were graded correctly.

Average label noise is therefore far smaller than the gaps between the methods we compare, but a low marginal error rate does not establish independence of errors. The extractor, the grader and the probe may still fail together on long-tail entities, ambiguous questions or malformed outputs. We

consequently do not treat automatic grading as unbiased ground truth and do not exclude correlation between grading error and probe error. The main conclusions are constrained jointly by the cross-dataset results, this human audit, and the recall metric of Appendix J, which does not depend on an LLM judge.

Human robustness check for calibration labels. To test whether automatic grading can change the calibration conclusions, we randomly sampled 500 rows from the 10,000-row TriviaQA pool behind the Phase A calibration tables and manually reviewed every row whose extracted answer was not an exact string match to the reference answer or its aliases. The full-pool and subset metrics are:

Table 10: Full-pool versus 500-row-subset metrics. The subset is used to audit labels, not to re-estimate the final ECE.

Source	ECE		AUROC		Brier	
	Full	Subset	Full	Subset	Full	Subset
Probe	.0147	.0363	.9477	.9621	.0809	.0718
Verbal	.1782	.1905	.8243	.8511	.1839	.1877
Token	.2514	.2602	.6360	.6181	.2530	.2586

Across ten resamples of 500 rows, probe ECE ranges from .028 to .053, while AUROC and Brier are more stable. ECE is therefore sensitive to the audit sample size, so this subset is useful for testing label reliability but not for reproducing the full-pool ECE.

Among the 500 rows, 258 exactly matched the reference answer and 28 exactly matched an alias. The remaining 214 were manually reviewed. The grading decision was confirmed for 210 rows, with one clear grading error, two boundary cases, and one protocol issue in which "I don't know" was routed as `INCORRECT`. The review exposed one clear false rejection caused by a defective reference answer and one boundary false acceptance of an answer with incorrect extra information, exposing residual risks in both the gold data and semantic grading. These 214 rows were selected by string mismatch and therefore do not provide an unbiased estimate of the overall error rate.

After manually correcting the two semantic decisions judged to require correction, subset ECE is .0363, .1905, and .2602 for the probe, verbal, and token sources, respectively, unchanged from the original values. Probe AUROC changes from .9621 to .9624 and Brier from .0718 to .0715. The audit correction therefore does not change the calibration gap between methods. It does not establish independence between grader and probe errors. Qwen3-32B remains an auditable but fallible proxy for correctness.

E ROBUSTNESS OF THE CONFIDENCE COMPARISON

This appendix supplements Table 1 with the per-term Brier decomposition $\mathrm{Brier} = \mathrm{Rel} - \mathrm{Res} + \mathrm{Unc}$ for both models and both confidence sources. The identity is *approximate*: it is exact only when the score variance inside each of the 10 equal-width bins is negligible. Probe scores are tightly concentrated within bins and reconstruct the Brier score to within ± 0.001. Verbalized scores span a wide range inside the same bin, so the within-bin variance is not captured by Rel or Res and the verbal rows show a systematic ± 0.003–0.006 residual. This is a property of the binning, not a transcription error, and was verified against the source data.

Table 11: Brier score and its decomposition into Reliability (↓), Resolution (↑) and Uncertainty. Uncertainty depends on the base accuracy of the scorable subset and is therefore comparable only within a (model, dataset) cell. Bold marks the better value per cell.

Dataset (Acc)	Source	Brier ↓	Rel ↓	Res ↑	Unc
Qwen3-14B					
TriviaQA (.729)	Verbal	.184	.052	.059	**.194**
	Probe	**.081**	**.001**	**.116**	.197
EntityQ (.267)	Verbal	.455	.297	.040	.207
	Probe	**.097**	**.000**	**.098**	**.196**
NQ-open (.440)	Verbal	.416	.207	.023	.249
	Probe	**.177**	**.009**	**.077**	**.246**
SimpleQA (.066)	Verbal	.532	.480	.001	.067
	Probe	**.076**	**.019**	**.004**	**.061**
Ministral-3-8B					
TriviaQA (.664)	Verbal	.263	.068	.018	**.216**
	Probe	**.105**	**.000**	**.112**	.218
EntityQ (.207)	Verbal	.564	.413	.016	.170
	Probe	**.091**	**.001**	**.075**	**.164**
NQ-open (.370)	Verbal	.490	.268	.013	.238
	Probe	**.188**	**.021**	**.066**	**.234**
SimpleQA (.061)	Verbal	.671	.612	.000	.059
	Probe	**.116**	**.058**	**.001**	.059

The pattern is identical in both families: the probe's Reliability is essentially zero (its probability scale is accurate) and its Resolution is the higher of the two (it separates correct from incorrect traces), whereas the verbalized score has high Reliability error (systematic overconfidence) and low Resolution. Figure 5 gives the per-bin version of this comparison: the probe curve lies close to the diagonal, while the verbalized and token-probability curves concentrate their error in the high-confidence region.

E.1 SENSITIVITY TO MISSING-SCORE SELECTION BIAS

ECE, AUROC and Brier can only be computed on samples for which a verbalized score is successfully parsed. Since format compliance differs across methods, the main tables report calibration conditional on the n_{scored} subset, and a method with low coverage could benefit from a different set of admitted samples. We check robustness under three conventions. **Convention 1** is the main-table convention with denominator n_{scored}. There, the extract rate measures how often an answer can be extracted and coverage how often a valid score is present, and their difference is the set of samples whose answer is gradable but whose score is missing. **Convention 2** keeps, per seed, only the questions for which every main-table column has at least 5 of 10 valid scores, and recomputes ECE, AUROC and Brier on that common question set. **Convention 3** assigns a score of 0 to every missing score and recomputes ECE and Brier over all samples. Because this creates a large tie at 0, AUROC is not reported.

Both the common-subset and full-sample results pool the three training seeds 42/43/44. Verbal has seed 42 only and is marked *. Error terms are question-level bootstrap half-widths with $B{=}5000$. Qwen's main-table columns are Probe-SD, Probe-SD-LRM, Probe-SD w/o CW, SC-SD and Verbal. Ministral omits the *w/o CW* variant. The readout probe has a different coverage mechanism and is reported separately. The standalone probe results in Table 1 use the Phase-A evaluation batch,

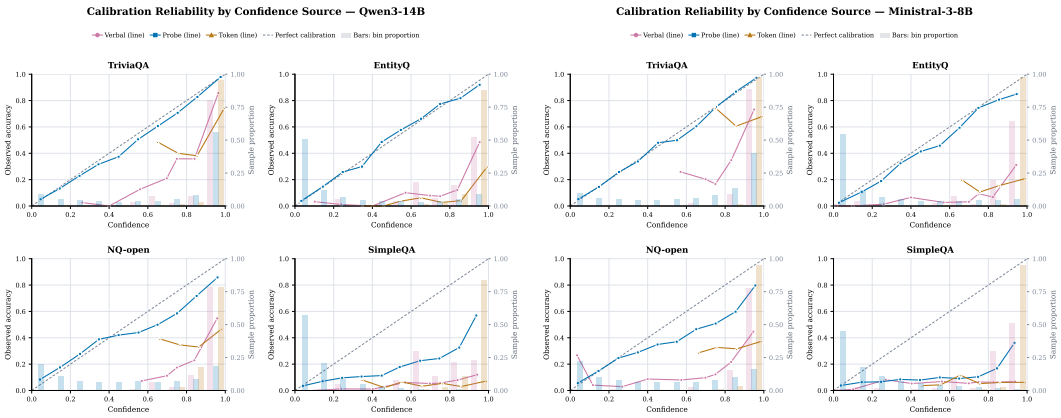

Figure 5: Reliability diagrams over four datasets and three sources (verbal / probe / token probability). Left: Qwen3-14B. Right: Ministral-3-8B. In each panel the solid line connects per-bin mean confidence to observed accuracy (left axis) and the dashed line is perfect calibration. Pale bars give the fraction of samples in each of the 10 equal-width confidence bins (right axis).

whereas the readout-probe check below reuses the Probe baseline from the SFT evaluation suite. They use the same L24 residual, C=0.005 recipe, but not the same evaluation batch or coverage, so their point estimates need not match.

Table 12: Convention 2 (common subset). Common question counts are: EntityQ 982/998, NQ-open 690/893, SimpleQA 846/929, TriviaQA 989/981 for Qwen/Ministral. Bold marks the best method per column within a model.

Model	Method	TriviaQA	EntityQ	NQ-open	SimpleQA	Macro
ECE ↓						
Qwen	Probe-SD	**.023±.006**	**.090±.008**	**.124±.012**	**.114±.007**	**.088±.004**
Qwen	Verbal*	.177±.021	.531±.022	.423±.032	.665±.015	.449±.012
Qwen	Probe-SD-LRM	.030±.007	.102±.008	.163±.013	.136±.007	.108±.004
Qwen	Probe-SD w/o CW	.032±.006	.105±.008	.146±.012	.126±.007	.102±.004
Qwen	SC-SD	.063±.008	.224±.009	.241±.013	.237±.009	.191±.005
Min.	Probe-SD	**.043±.006**	**.087±.009**	**.141±.011**	.162±.007	**.108±.004**
Min.	Verbal*	.246±.024	.627±.022	.515±.027	.767±.014	.539±.011
Min.	Probe-SD-LRM	.064±.007	.108±.009	.164±.011	**.158±.007**	.123±.004
Min.	SC-SD	.051±.005	.150±.009	.175±.011	.184±.007	.140±.004
AUROC ↑						
Qwen	Probe-SD	**.929±.007**	**.881±.009**	**.809±.013**	**.668±.024**	**.821±.007**
Qwen	Verbal*	.826±.017	.811±.017	.716±.022	.625±.043	.745±.013
Qwen	Probe-SD-LRM	.926±.007	.868±.010	.804±.013	.654±.027	.813±.008
Qwen	Probe-SD w/o CW	.919±.007	.872±.009	.800±.013	.644±.024	.809±.007
Qwen	SC-SD	.877±.009	.847±.009	.758±.012	.658±.020	.785±.007
Min.	Probe-SD	.891±.008	**.879±.009**	**.811±.012**	**.683±.025**	**.816±.008**
Min.	Verbal*	.669±.015	.734±.016	.640±.018	.523±.041	.641±.012
Min.	Probe-SD-LRM	**.904±.007**	.875±.010	**.811±.012**	.641±.027	.808±.008
Min.	SC-SD	.809±.009	.821±.011	.733±.011	.612±.021	.744±.007
Brier ↓						
Qwen	Probe-SD	**.092±.005**	**.131±.006**	**.197±.008**	**.106±.006**	**.132±.003**
Qwen	Verbal*	.183±.017	.457±.018	.404±.028	.536±.013	.395±.010
Qwen	Probe-SD-LRM	.096±.005	.137±.006	.213±.009	.109±.006	.139±.003
Qwen	Probe-SD w/o CW	.096±.005	.140±.006	.208±.009	.116±.006	.140±.003
Qwen	SC-SD	.116±.006	.212±.007	.263±.010	.185±.006	.194±.004
Min.	Probe-SD	**.122±.005**	**.126±.005**	**.200±.007**	.122±.005	**.142±.003**
Min.	Verbal*	.262±.021	.564±.019	.493±.024	.678±.012	.499±.010
Min.	Probe-SD-LRM	.127±.005	.127±.005	.204±.007	**.113±.005**	.143±.003
Min.	SC-SD	.153±.006	.174±.006	.237±.006	.166±.005	.183±.003

On Qwen, Probe-SD is best on all four datasets and on the macro average for all three metrics. Ministral has local exceptions—Probe-SD-LRM is better on TriviaQA AUROC and on SimpleQA ECE/Brier, and the two tie on NQ-open AUROC—but the hierarchy is unchanged: probe-based methods beat SC-SD, and both beat Verbal by a wide margin.

Table 13: Convention 3 (full-sample attribution, missing scores set to 0). AUROC is not reported because of the tie mass at 0.

Model	Method	TriviaQA	EntityQ	NQ-open	SimpleQA	Macro
ECE ↓						
Qwen	Probe-SD	**.024**±.009	**.089**±.013	**.162**±.018	**.104**±.010	**.095**±.006
Qwen	Verbal*	.182±.020	.493±.020	.410±.022	.545±.018	.408±.010
Qwen	Probe-SD-LRM	.031±.010	.101±.014	.205±.018	.122±.010	.115±.007
Qwen	Probe-SD w/o CW	.033±.010	.105±.013	.182±.017	.114±.010	.109±.006
Qwen	SC-SD	.064±.013	.222±.016	.266±.019	.208±.012	.190±.007
Min.	Probe-SD	**.045**±.008	**.087**±.014	**.146**±.018	.159±.011	**.109**±.007
Min.	Verbal*	.245±.023	.596±.021	.469±.020	.680±.016	.498±.010
Min.	Probe-SD-LRM	.066±.011	.106±.014	.165±.018	**.152**±.010	.122±.007
Min.	SC-SD	.054±.007	.147±.014	.178±.016	.175±.010	.138±.006
Brier ↓						
Qwen	Probe-SD	**.092**±.009	**.130**±.010	**.224**±.013	**.097**±.008	**.136**±.005
Qwen	Verbal*	.188±.017	.424±.017	.397±.020	.439±.015	.362±.009
Qwen	Probe-SD-LRM	.097±.009	.135±.009	.245±.015	.099±.008	.144±.005
Qwen	Probe-SD w/o CW	.097±.009	.139±.010	.237±.014	.105±.008	.144±.005
Qwen	SC-SD	.116±.010	.210±.011	.285±.015	.165±.008	.194±.006
Min.	Probe-SD	**.124**±.009	**.125**±.009	**.203**±.011	.121±.008	.143±.005
Min.	Verbal*	.260±.020	.537±.019	.451±.017	.601±.015	.462±.009
Min.	Probe-SD-LRM	.129±.008	.125±.009	.205±.011	**.110**±.008	**.142**±.005
Min.	SC-SD	.154±.010	.171±.010	.238±.010	.161±.008	.181±.005

Full-sample attribution preserves the overall ordering. This convention is not an upper bound: if the missing samples are mostly wrong answers, assigning them 0 makes their ECE contribution close to zero and can *lower* the total. Qwen's Verbal ECE on NQ-open, for instance, drops from .436 under the conditional convention to .410 under full-sample attribution. The common subset is therefore our primary test for selection bias, with full-sample attribution as a supplementary view.

Readout probe under all three conventions. We evaluate the Qwen readout-probe baseline (L24 residual, C=0.005) on the seed-42 run of the SFT evaluation suite, using the same common question set as the main-table columns and again assigning 0 to missing probe scores, with bootstrap B=5000. This is distinct from the standalone probe evaluation in Table 1: the recipe is the same, but the evaluation batch and coverage differ. Table 14 therefore tests within-run stability across denominator conventions rather than reproducing Table 1.

The probe's macro ECE is .050, .049 and .047 under the three conventions and its Brier .115, .114 and .108, so its results are stable under changes of denominator. Combining the main-table columns with the readout probe, we find that Probe-SD's calibration advantage cannot be explained by missing-score selection or by the probe's higher coverage. This conclusion concerns selection bias induced by missing scores only. It does not address grading error, answer-extraction error, or the causal origin of the missingness mechanism itself.

Table 14: Qwen readout-probe baseline from the SFT evaluation suite under the three denominator conventions. Brackets are 95% bootstrap intervals.

Dataset	Convention	ECE ↓	AUROC ↑	Brier ↓
TriviaQA	scorable subset	.014 [.010,.026]	.948 [.938,.957]	.080 [.072,.089]
	common subset (n_q=989)	.013 [.010,.025]	.948 [.938,.957]	.080 [.071,.089]
	full-sample attribution	.015 [.010,.027]	—	.080 [.072,.088]
EntityQ	scorable subset	.022 [.015,.038]	.898 [.882,.914]	.109 [.099,.120]
	common subset (n_q=982)	.023 [.015,.039]	.899 [.883,.915]	.109 [.099,.120]
	full-sample attribution	.020 [.014,.035]	—	.099 [.090,.109]
NQ-open	scorable subset	.074 [.058,.093]	.817 [.796,.837]	.181 [.168,.193]
	common subset (n_q=690)	.067 [.048,.090]	.834 [.809,.856]	.172 [.157,.187]
	full-sample attribution	.070 [.054,.089]	—	.175 [.163,.187]
SimpleQA	scorable subset	.089 [.077,.103]	.670 [.615,.720]	.089 [.080,.099]
	common subset (n_q=846)	.092 [.080,.108]	.658 [.600,.712]	.096 [.085,.107]
	full-sample attribution	.081 [.069,.094]	—	.079 [.070,.088]
Macro	scorable subset	.050	.833	.115
	common subset	.049	.835	.114
	full-sample attribution	.047	—	.108

F Training Hyperparameters and Probe Configurations

All SFT methods use full-parameter finetuning with `swift` (Zhao et al., 2024) and three training seeds (42/43/44), and differ only in the source of the training data. The two models share most hyperparameters. The differences are confined to the learning rate and confidence weighting (Table 15).

Table 15: Supervised finetuning hyperparameters.

Hyperparameter	Qwen3-14B	Ministral-3-8B	Note
Training set size	10,000	10,000	identical across methods
Learning rate	5×10^{-6}	3×10^{-6}	
Epochs	2	2	$\approx$625 steps
LR schedule	cosine to 0	cosine to 0	50 warmup steps
Weight decay	0.1	0.1	
Adam β_1/β_2	0.9 / 0.95	0.9 / 0.95	
Per-device batch	1	1	
Gradient accumulation	4	4	effective batch $1\times4\times8=32$
Max sequence length	8192	8192	
Precision	bf16	bf16	
Optimizer sharding	DeepSpeed ZeRO-1	DeepSpeed ZeRO-1	8 GPUs
Confidence weighting	$\omega=3$	not used	applied to score tokens only
Initialization	Qwen3-14B-Base	Ministral-3-8B-Base-2512	base checkpoint

Confidence weighting multiplies the cross-entropy loss on the verbalized score tokens by ω. Qwen3-14B uses $\omega=3$ to offset the gradient dilution caused by long chains of thought. Ministral-3-8B rehearses the score inside and outside its thinking block, which already duplicates the score tokens and acts as implicit weighting, so no extra weighting is applied.

Table 16: Probe configurations selected independently for each purpose and sampling temperature. Selection uses validation data only (Appendix C).

Purpose	T	Qwen3-14B	Ministral-3-8B
Evaluation probe scoring	0.6	L24 residual, $C=0.005$	L21 MLP, $C=2.0$
Probe-SD training labels	1.5	L24 residual, $C=0.005$	L22 residual, $C=0.25$

Each temperature is searched independently with the same all-layer validation procedure: on the TriviaQA validation split we compute selection accuracy, ECE and AUROC, together with the OOD-macro ECE and AUROC over EntityQuestions and NQ-open validation data, rank the five metrics separately and take the configuration with the lowest mean rank. No test set is used for selection. SimpleQA has only 1,000 questions and no validation split, so it does not participate in selection. Qwen selects L24 residual with $C=0.005$ at both temperatures. Ministral's two winners are the adjacent layers 21 and 22 with different readouts and C. The all-layer sweep shows that the calibration signal does not emerge abruptly at a particular relative depth but is readable across many layers and forms a broad plateau over the middle of the network. L24/40 and L21/34 are the specific winners picked on that plateau by the validation composite rank and should not be read as the only effective layers. Transferring the $T=0.6$ probe directly to $T=1.5$ samples instead of matching temperatures costs Qwen almost nothing (macro ECE .043 against .047 for the temperature-matched probe, i.e. direct transfer is even marginally better) but costs Ministral substantially (.100 against .054), which is why we reselect per temperature rather than reuse the evaluation probe.

High-temperature sampling gives a modest, dataset-dependent gain at best: relative to $T=0.6$, $T=1.5$ improves TriviaQA by 0.5 points and EntityQ by 0.2 points, leaves NQ-open unchanged, and lowers SimpleQA by 0.1 points. All differences fall within the bootstrap half-widths. We therefore keep $T=1.5$ for candidate diversity rather than claim a reliable accuracy gain, and apply top-k/top-p truncation to suppress generation breakdown, in particular Ministral's formatting and repetition failures at high temperature.

After training, Ministral-3-8B weights must be converted from HF format to Ministral-native format for vLLM deployment. The conversion has to apply the inverse of the wq/wk head interleaving

Table 17: Training-data generation temperature ablation for Qwen3-14B (§4.2). All resulting models are evaluated directly at T=0.6. Errors are 95% question-level bootstrap half-widths.

Training pool	TriviaQA	NQ-open	EntityQ	SimpleQA
T=0.6	.739±.024	.470±.027	.289±.025	.077±.013
T=1.0	.738±.024	.467±.026	.288±.025	.078±.013
T=1.5	.744±.024	.470±.026	.291±.025	.076±.012
T=1.5, no top-k/top-p	.745±.024	.471±.026	.286±.024	.078±.012

permutation. Without it the Q/K heads are scrambled and the output degenerates into repetitive noise. Qwen3-14B is deployed directly in HF format.

G Complete Per-Dataset Results

Tables 18 and 19 expand Table 2 into all four datasets. SFT methods (SC-SD, EA-SD, Probe-SD-LRM, Probe-SD) are averaged over the three training seeds 42/43/44. Verbal, the frozen Probe, Verbal-Iso, Verbal-TS+bias, Answer Prob. and Surface-Feat use the seed-42 evaluation run, with * marking the single-run baselines. All ± values are 95% question-level bootstrap half-widths with B=5000 and bootstrap seed 42. For SFT methods each training seed is evaluated first and the three seeds are then pooled. Out-of-domain macro half-widths in Table 2 are composed from the three datasets' bootstrap variances rather than resampled from the macro average itself. Where ECE is close to 0 the bootstrap interval is right-skewed, so the reported ± half-width is a symmetric approximation. ECE uses 10 equal-width bins. AUROC and Brier are computed on the scorable subset. A dash marks a readout or post-hoc method that reuses the Verbal answers and therefore has no accuracy of its own. The frozen Probe scores the original model's sampled answers rather than the student's outputs.

Table 18: Qwen3-14B, complete per-dataset results. TriviaQA is in-domain. The other three are held-out test datasets under distribution shift. Bold marks the best value per column within a dataset.

Dataset	Method	Acc ↑	ECE ↓	AUROC ↑	Brier ↓
TriviaQA	Verbal	.729±.024*	.178±.020	.825±.017	.184±.017
	Frozen Probe*	—	.014±.008	**.948±.010**	**.080±.009**
	Verbal-Iso	—	**.013±.009**	.825±.017	.128±.010
	Verbal-TS+bias	—	.047±.011	.825±.017	.131±.010
	Answer Prob.	—	.251±.024	.636±.029	.253±.023
	Surface-Feat	—	.042±.010	.858±.016	.128±.010
	EA-SD	**.750±.024**	.093±.011	.878±.014	.119±.011
	SC-SD	.748±.023	.063±.013	.876±.014	.116±.010
	Probe-SD-LRM	.729±.024	.031±.011	.926±.011	.096±.009
	Probe-SD	.749±.024	.024±.009	**.928±.011**	**.092±.009**
EntityQ	Verbal	.267±.024*	.529±.022	.812±.017	.455±.018
	Frozen Probe*	—	**.022±.012**	**.898±.016**	**.109±.011**
	Verbal-Iso	—	.267±.019	.812±.017	.229±.011
	Verbal-TS+bias	—	.285±.019	.812±.017	.238±.011
	Answer Prob.	—	.689±.026	.724±.027	.680±.026
	Surface-Feat	—	.340±.021	.777±.020	.294±.012
	EA-SD	.288±.025	.215±.013	.842±.016	.211±.012
	SC-SD	**.296±.025**	.224±.016	.847±.015	.212±.011
	Probe-SD-LRM	.268±.024	.102±.014	.868±.016	.136±.010
	Probe-SD	.291±.025	**.090±.013**	**.881±.015**	**.130±.010**
NQ-open	Verbal	.440±.027*	.436±.029	.705±.021	.416±.025
	Frozen Probe*	—	**.074±.018**	**.817±.021**	**.181±.013**
	Verbal-Iso	—	.245±.027	.705±.021	.276±.017
	Verbal-TS+bias	—	.250±.027	.705±.021	.275±.016
	Answer Prob.	—	.500±.027	.557±.029	.498±.025
	Surface-Feat	—	.251±.026	.707±.024	.286±.016
	EA-SD	**.472±.026**	.273±.019	.751±.019	.283±.017
	SC-SD	.470±.026	.251±.020	.748±.018	.273±.015
	Probe-SD-LRM	.440±.026	.170±.020	**.797±.021**	.218±.014
	Probe-SD	.471±.026	**.137±.018**	.796±.020	**.207±.013**
SimpleQA	Verbal	.066±.012*	.661±.016	.627±.042	.532±.013
	Frozen Probe*	—	**.089±.014**	.670±.052	**.089±.010**
	Verbal-Iso	—	.310±.016	.626±.043	.203±.010
	Verbal-TS+bias	—	.348±.016	.627±.042	.223±.009
	Answer Prob.	—	.897±.014	.477±.047	.880±.013
	Surface-Feat	—	.405±.016	.608±.042	.276±.010
	EA-SD	.074±.012	.193±.011	.648±.034	.165±.010
	SC-SD	**.076±.012**	.234±.014	.665±.031	.180±.009
	Probe-SD-LRM	.066±.011	.134±.011	.658±.042	.108±.009
	Probe-SD	**.076±.012**	**.113±.011**	**.671±.037**	**.105±.009**

Table 19: Ministral-3-8B, complete per-dataset results. Conventions as in Table 18.

Dataset	Method	Acc ↑	ECE ↓	AUROC ↑	Brier ↓
TriviaQA	Verbal	.664±.025*	.247±.024	.669±.015	.263±.021
	Frozen Probe*	—	**.019±.008**	**.915±.013**	**.106±.009**
	Verbal-Iso	—	**.020±.014**	.669±.015	.188±.009
	Verbal-TS+bias	—	.067±.016	.669±.015	.193±.008
	Answer Prob.	—	.302±.025	.569±.027	.306±.025
	Surface-Feat	—	.042±.010	.770±.016	.174±.009
	EA-SD	.696±.024	.123±.015	.826±.014	.160±.012
	SC-SD	**.698±.025**	.052±.007	.809±.015	.153±.010
	Probe-SD-LRM	.644±.026	.064±.011	**.904±.012**	.128±.008
	Probe-SD	.694±.025	.044±.009	.895±.014	**.120±.008**
EntityQ	Verbal	.207±.023*	.627±.022	.734±.016	.564±.019
	Frozen Probe*	—	**.027±.012**	**.901±.017**	**.099±.010**
	Verbal-Iso	—	.320±.020	.734±.016	.258±.009
	Verbal-TS+bias	—	.340±.021	.734±.016	.271±.009
	Answer Prob.	—	.761±.025	.562±.028	.755±.024
	Surface-Feat	—	.331±.019	.790±.019	.267±.010
	EA-SD	.216±.023	.219±.014	.836±.017	.207±.012
	SC-SD	**.220±.023**	.150±.014	.821±.018	.174±.010
	Probe-SD-LRM	.197±.022	.108±.014	.875±.017	.127±.009
	Probe-SD	**.220±.023**	**.088±.014**	**.883±.015**	**.125±.008**
NQ-open	Verbal	.370±.026*	.513±.026	.643±.018	.490±.022
	Frozen Probe*	—	.114±.019	.805±.021	**.195±.013**
	Verbal-Iso	—	.225±.025	.643±.018	.274±.011
	Verbal-TS+bias	—	.236±.025	.643±.018	.277±.010
	Answer Prob.	—	.589±.027	.542±.023	.587±.026
	Surface-Feat	—	.253±.025	.687±.021	.283±.011
	EA-SD	**.410±.026**	.280±.019	.757±.017	.286±.014
	SC-SD	.405±.026	.174±.017	.732±.017	.237±.010
	Probe-SD-LRM	.366±.025	.163±.019	**.811±.019**	.204±.011
	Probe-SD	.406±.026	**.138±.018**	**.811±.018**	**.198±.011**
SimpleQA	Verbal	.061±.011*	.761±.014	.531±.040	.671±.012
	Frozen Probe*	—	.159±.014	.615±.050	.126±.010
	Verbal-Iso	—	.403±.014	.531±.040	.268±.008
	Verbal-TS+bias	—	.446±.014	.531±.040	.293±.007
	Answer Prob.	—	.909±.012	.457±.041	.895±.012
	Surface-Feat	—	.403±.014	.542±.041	.279±.008
	EA-SD	**.070±.011**	.253±.011	.610±.038	.210±.010
	SC-SD	**.070±.012**	.183±.011	.612±.033	.165±.008
	Probe-SD-LRM	.056±.010	**.156±.010**	.647±.042	**.111±.008**
	Probe-SD	**.070±.012**	.161±.011	**.683±.040**	.120±.008

The frozen Probe provides a lower-ECE reference than the distilled student. On the three held-out datasets, Qwen's frozen Probe has macro ECE .062 and AUROC .795, compared with .113 and .783 for Probe-SD. Ministral has .100 and .774, compared with .130 and .792. The student therefore does not fully match the teacher's calibration profile, although it remains substantially better calibrated than the verbal baseline. The two families agree: Probe-SD is best on most dataset–metric cells, and SC-SD occasionally overtakes it on accuracy (EntityQ, SimpleQA). The same ordering survives the alternative missing-score conventions of Appendix E.

H TRAINING-SET CONSTRUCTION AND DOWNSAMPLING

This appendix supports the data-downsampling step of Section 5.1. The probe-coherent selection and downsampling pipeline is shared across the two models: 15,000 questions $\times$ 20 samples $=$ 300,000 candidates, of which the coherence filter keeps one trajectory per question (Ministral 13,840, Qwen 14,951), which is then downsampled to 10,000 training trajectories.

Probe-score distribution before and after selection. Table 20 reports the probe-score distribution of the candidate pool and the retained set. The pool statistics are over a random sample of 15,000 scored candidates, so its size is directly comparable to the retained set's one trajectory per question. The filter biases the retained set toward mid-to-high confidence: on Ministral the retained probe mean rises from 0.636 to 0.672, whereas on Qwen it is nearly unchanged (0.722 to 0.723).

Table 20: Probe-score distribution of the candidate pool and the coherently selected set.

Model	Set	n	mean	p25	p50	p75
Ministral-3-8B	Pool (random 15k)	15,000	0.636	0.317	0.777	0.943
	Retained	13,840	0.672	0.495	0.782	0.930
Qwen3-14B	Pool (random 15k)	15,000	0.722	0.460	0.936	0.992
	Retained	14,951	0.723	0.485	0.915	0.989

Why the shift is large on Ministral. The filter examines candidates in order of $|p_i - \bar{p}_q|$ and keeps the first one whose relabeled triple is judged coherent. Candidates with multiple or inconsistent in-chain confidence numbers are rejected before the coherence check (Section 4.3). Ministral's low-probe trajectories fail this stage disproportionately: trajectories with probe below 0.2 pass at only $\sim$19%, because Ministral's uncertain answers tend to have long chains of thought (median 2,064 tokens for probe$<$0.2 versus 399 for probe$\geq$0.8) in which the model re-evaluates and inserts several, often conflicting, confidence numbers, and such trajectories are discarded by the multiple-score check. Format failures likewise concentrate in long, low-probe trajectories. The average position of the first passing candidate is therefore much higher on Ministral (3.11) than on Qwen (1.43), and the retained set shifts toward mid-to-high confidence. This makes Ministral the harder case for testing whether the shift matters.

Downsampling variants. To assess how the selection shift affects performance, we downsample the retained trajectories to 10,000 under four policies: uniform random (*rand_sub*), a quota by difficulty decile replicating the pool's difficulty mix (*diffdist_rp*), the difficulty quota plus a correct/incorrect quota aligning the training accuracy to the global accuracy (*calibrated*), and the difficulty quota plus a within-decile sliding window aligning each decile's probe mean (*probe_align*). Table 21 reports their 4-dataset macro calibration (3 seeds; Δpp is the largest difficulty-decile deviation from the pool and Train Acc the aligned training accuracy), together with the Qwen calibrated-versus-random comparison discussed below.

Table 21: Downsampling policies across models, 4-dataset macro (3 seeds). Δpp is the largest difficulty-decile deviation from the pool and Train Acc the aligned training accuracy, both from the Ministral variant study.

Model	Downsampling	ECE $\downarrow$	AUROC $\uparrow$	Brier $\downarrow$	Sample Acc	Δpp	Train Acc
Qwen3-14B	calibrated (Probe-SD)	0.091	0.819	0.134	0.397	—	—
	random (*rand_sub*)	0.094	0.820	0.134	0.394	—	—
Ministral-3-8B	random	0.123	0.808	0.150	0.346	4.26	70.0%
	diffdist_rp	0.112	0.813	0.144	0.347	0.01	65.5%
	calibrated	**0.105**	**0.815**	**0.142**	0.348	1.17	61.8%
	probe_align	0.118	0.807	0.149	0.348	0.01	66.2%

Robustness to the selection shift. The alignment matters only where the filter induces a large distribution shift. On Qwen, whose selection barely changes the probe distribution, random subsampling (*rand_sub*) gives macro ECE/AUROC/Brier of 0.094/0.820/0.134, essentially indistinguishable from the calibrated Probe-SD (0.091/0.819/0.134). On Ministral, whose shift is large, cali-

brated improves ECE from 0.123 to 0.105 (Table 21). The gain from alignment is thus modest and confined to the large-shift case, and the method is overall robust to the coherence filter's distribution shift. Sample Acc is essentially unchanged in either model.

I SUPPORTING DATA FOR THE MECHANISTIC ANALYSIS

This appendix reports the teacher-to-student fidelity and label-source results underlying §6.1–§6.2. All experiments use Qwen3-14B with at most 10 samples per question. Probe-SD, SC-SD and Verbal-SD share the same source sampling, filtering pipeline and Base initialization, and differ only in the confidence target: the frozen probe probability, the self-consistency answer-cluster frequency, and the source model's own verbalized score. Question- level metrics average over the samples of a question first. Trace-level metrics are computed per trace. Probe is the L24 residual, C=0.005, three-seed ensemble used to generate training labels. Correlations and intervals use question-level bootstrap with B=5000.

Table 22: Per-decile values behind Figure 3(a): each source's mean, aggregated by decile of the source verbalized score q_{src} on TriviaQA (n=100 per decile). Deciles are equal-frequency in q_{src}, so that column is monotone by construction.

Decile	q_{src}	Probe	Observed acc	Probe-SD	SC-SD	Verbal-SD
1	.675	.191	.141	.295	.374	.756
2	.803	.337	.331	.424	.522	.817
3	.866	.461	.423	.537	.637	.867
4	.907	.601	.546	.645	.752	.906
5	.938	.836	.891	.865	.878	.936
6	.952	.900	.916	.904	.926	.945
7	.959	.935	.940	.957	.940	.950
8	.966	.953	.977	.967	.967	.955
9	.975	.977	.985	.984	.980	.960
10	.989	.986	.989	.992	.993	.972

The four lowest deciles are the region where the source is overconfident: in decile 1 the source reports .675 against an observed accuracy of .141, Probe reports .191, Probe-SD follows down to .295, and Verbal-SD still reports .756. In the high deciles all curves converge, because the two teachers already agree there.

Table 23: Per-quartile values behind Figure 3(b): mean absolute error of each student relative to Probe, by quartile of the overconfidence gap $q_{src} - p_{\text{probe}}$ on TriviaQA (n=250 per quartile). The last column is $|q_{src} - p_{\text{probe}}|$, the distance between the verbalized score and Probe with no distillation at all.

Quartile	Gap range	Probe-SD	SC-SD	Verbal-SD	q_{src} itself
Q0	$[-.560, -.011)$	.018	.034	.037	.030
Q1	$[-.011, +.044)$	.025	.030	.019	.013
Q2	$[+.044, +.373)$	.120	.134	.202	.188
Q3	$[+.373, +.832]$	**.164**	.263	.588	.575

Verbal-SD nearly coincides with the no-distillation reference at every quartile (.588 versus .575 in Q3), i.e. it simply reproduces the original verbalized score, whereas Probe-SD stays at .164 in the quartile of largest disagreement. The three students differ mainly in Q2–Q3 and are almost identical in Q0.

Table 24: Question-level Pearson correlation between each student's verbalized score and the two candidate teachers, Probe and the source verbalized score q_{src}. Bold marks, for each student, the teacher it correlates with more strongly.

	Probe-SD		SC-SD		Verbal-SD	
Dataset	Probe	q_{src}	Probe	q_{src}	Probe	q_{src}
TriviaQA	**.913**	.771	**.872**	.791	.767	**.822**
EntityQ	**.895**	.676	**.810**	.738	.658	**.803**
NQ-open	**.823**	.550	**.700**	.590	.526	**.725**
SimpleQA	**.762**	.494	**.642**	.641	.468	**.757**

All four datasets agree in direction: Probe-SD always correlates more strongly with Probe than with the source verbalized score, Verbal-SD does the reverse, and SC-SD is intermediate. The confidence structure a student inherits is determined by its training label, and this holds outside the training domain as well.

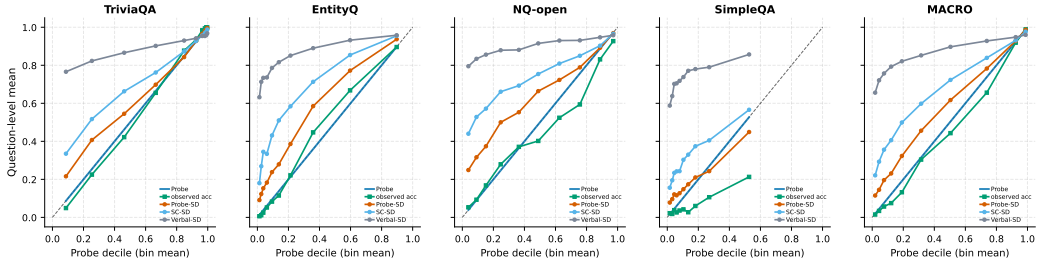

Figure 6: Probe, observed accuracy and the three students' verbalized confidence, conditioned on Probe decile rather than on q_{src}. Single datasets have n=100 per decile, MACRO n=400.

Binning by Probe rather than by q_{src} (Figure 6) makes the shape argument explicit: Probe-SD's output follows Probe's difficulty profile decile by decile, whereas Verbal-SD is a nearly flat high-confidence line almost independent of question difficulty—on EntityQ it rises only from .632 in the lowest decile to .958 in the highest, while Probe rises from .013 to .897 over the same range. If Probe-SD were merely applying a global downward shift to the verbalized score, its curve would run parallel to Verbal-SD. The measured difference in shape rules that explanation out.

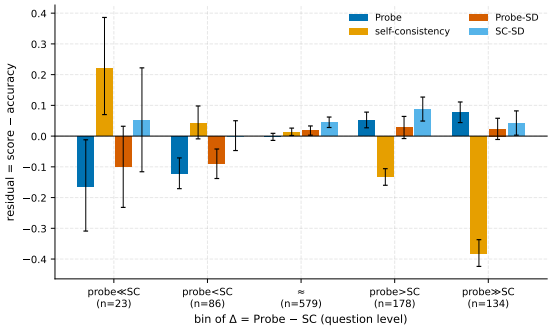

Figure 7: Calibration residuals on TriviaQA, binned by the disagreement between Probe and self-consistency (§6.2). Positive values indicate overconfidence, negative values underconfidence. Per-bin values are in Table 25.

Table 25: Per-bin values behind Figure 7: calibration residuals on TriviaQA (N=1,000 questions), binned by the question-level disagreement $\Delta = p_{\text{probe}} - f_{\text{SC}}$. Residual = score − corresponding accuracy, so positive values are overconfidence. Probe and SC are referenced to the source accuracy and each student to its own. Brackets give 95% question-level bootstrap intervals (B=5000).

Bin	n	acc_{src}	Probe	SC	Probe-SD	SC-SD
Probe≪SC	23	.635	−.165 [−.309, −.012]	+.219 [.070, .386]	−.100 [−.232, .032]	+.051 [−.116, .222]
Probe<SC	86	.700	−.123 [−.171, −.071]	+.041 [−.009, .098]	−.090 [−.138, −.042]	−.000 [−.047, .050]
≈	579	.739	−.003 [−.014, .009]	+.013 [.002, .026]	+.018 [.003, .033]	+.045 [.028, .062]
Probe>SC	178	.610	+.052 [.027, .078]	−.133 [−.160, −.106]	+.028 [−.008, .064]	+.087 [.049, .127]
Probe≫SC	134	.766	+.078 [.044, .111]	−.381 [−.424, −.337]	+.023 [−.011, .058]	+.041 [.003, .082]

The middle bin holds 57.9% of the questions. Both teachers are close to calibrated there and the students differ least, so the ECE gaps of Table 2 can be attributed bin by bin to the two disagreement

tails. SC-SD fails to reproduce its teacher's direction at the Probe≫SC end (SC is $-.381$, SC-SD is $+.041$). The same mismatch shows up as a distance to its own label: in the top quintile of $|\mathrm{Probe} - \mathrm{SC}|$, SC-SD's MAE to the SC label is .415 while its MAE to Probe is only .126. Answer frequency is too noisy a label in that region for the student to fit, which places a ceiling on the fidelity of SC distillation.

Table 26: Mean absolute error of each student relative to Probe, by quintile of $|\mathrm{Probe} - \mathrm{SC}|$ on TriviaQA (n=200 per quintile). This binning is by the magnitude of teacher disagreement and is not the same convention as Table 23.

Quintile	$\lvert\mathrm{Probe} - \mathrm{SC}\rvert$	mean Δ	Probe-SD	SC-SD	Verbal-SD
Q0	$[.000, .013)$	$-.004$	**.030**	.055	.115
Q1	$[.013, .044)$	$-.018$	**.053**	.079	.173
Q2	$[.044, .107)$	$-.023$	**.123**	.171	.327
Q3	$[.107, .242)$	$+.060$	**.103**	.146	.263
Q4	$[.242, .855]$	$+.295$	**.100**	.126	.179

Probe-SD's MAE to Probe stays in .030–.123 and is the lowest in every disagreement quintile, so the fidelity difference is not driven by the low-disagreement samples alone.

J THE SOURCE OF THE BASE-INITIALIZATION ACCURACY GAIN

§5.3 ruled out probe relabeling as the source of the small accuracy gain, which leaves the initialization. Factual priming explains it (Gekhman et al., 2026): a reasoning model gains on factual QA less by decomposing the task than by *retrieving into its own context*—stating relevant facts in the chain of thought lowers the threshold for the answer that follows. Correctness then turns on two things, whether the key fact surfaces and whether it is used.

We measure the first with the *gold recall hit rate* r: a trace counts as a hit when the gold answer or an alias appears in the chain of thought, by string match, with no LLM in the loop. With a_1 and a_0 the accuracies conditional on hit and on miss,

$$\mathrm{acc} = r\,a_1 + (1-r)\,a_0, \tag{3}$$

so a difference between checkpoints splits into a retrieval term (Δr) and a usage term (Δa_1). A separate audit labels every extracted fact for truth and for whether it bridges to the answer, and *retrieval difficulty* counts hits over a question's 10 samples: 10 is *easy*, 1–9 *medium*, 0 *hard*.

Qwen3-14B: retrieval accounts for nearly everything. SFT-Base beats Instruct by 2.36 points (.7508 versus .7272). Gold recall explains 2.21 of them, 93.7%, while the usage term contributes 0.8%: accuracy after a hit is identical (.8915 versus .8912) and the two checkpoints produce equally many true bridging facts (4.95 versus 4.98). The gap is about getting a stored fact into context, not about reasoning over it.

Ministral-3-8B: retrieval dominates but does not exhaust. SFT-Base leads by 3.56 points (.6995 versus .6639), of which retrieval supplies 2.60 (73.0%). Here accuracy after a hit still differs (.8889 versus .8791), the usage term takes another 19.9%, and Reasoning produces fewer true bridging facts (4.32 versus 4.81) while hallucinating slightly more (27.6% versus 26.7%)—the direction Gekhman et al. (2026) report for hallucinated facts inflating final error.

Difficulty transitions agree across families: exactly one question falls from *easy* to *hard* in each, so finetuning erases nothing firmly encoded, while the middle band moves up (net $+42$ questions for Qwen, $+52$ for Ministral). Post-training reduces the retrievability of marginal facts, not knowledge. Gold recall is a string metric and misses paraphrases outside the alias list, so its level is an underestimate. Both checkpoints are measured the same way, so Δr survives the bias, and the fact audit, which uses no string matching, points the same way.

The rest of this appendix gives the supporting data. We compare two pairs—Qwen3-14B's SFT-Base student against Instruct, and Ministral-3-8B's SFT-Base student against Reasoning—on the 1,000 questions of the TriviaQA validation split, with 10 symmetric samples per question and answers graded by Qwen3-32B. The chain-of-thought quality audit uses glm-5-2-260617 (GLM-5-Team et al., 2026) to extract facts from 200 traces (100 matched pairs) and label *truth* and *relevance*. Gold recall only checks whether the gold answer or one of its aliases occurs as a substring of the chain of thought and does not depend on any LLM judgment.

Table 27: Sample-level accuracy difference and its decomposition into a retrieval term, a usage (conditional accuracy) term, and a miss-branch term, for both families.

	Qwen3-14B: SFT vs Instruct	Ministral-3-8B: SFT vs Reasoning
Sample accuracy	.7508 vs .7272 (+.0236)	.6995 vs .6639 (+.0356)
Retrieval term	+.0221 **(93.7%)**	+.0260 **(73.0%)**
Usage term	+.0002 (0.8%)	+.0071 (19.9%)
Miss-branch term	+.0013 (5.5%)	+.0025 (7.0%)

Writing $\mathrm{acc} = r\,a_1 + (1-r)\,a_0$, the total difference is first split into a hit branch $\Delta(ra_1)$ and a miss branch $\Delta[(1-r)a_0]$, and the hit branch is then split into a retrieval term $\Delta r \cdot a_1^{(F2)}$ and a usage term $r^{(F1)} \cdot \Delta a_1$. Both families are dominated by the retrieval term, but the residual differs: for Qwen the usage term contributes only 0.8%, whereas for Ministral it accounts for a further 19.9%, so the Ministral mechanism cannot be reduced to a pure retrieval effect.

On Qwen the two checkpoints are tied on effective bridges: true∧bridge counts are 4.95 and 4.98, and the 2.1-point advantage of SFT in true-fact rate comes from context or tangential facts that do

Table 28: Chain-of-thought quality audit, 200 traces per family. Facts are counted per trace.

	Qwen3-14B		Ministral-3-8B	
Metric	SFT-Base	Instruct	SFT-Base	Reasoning
Facts per trace	9.24	9.02	9.33	9.29
truth: true	80.1%	78.0%	69.7%	67.8%
truth: false (hallucination)	13.5%	15.2%	26.7%	27.6%
relevance: bridge	5.83	5.97	6.13	5.57
true $\wedge$ bridge	4.95	4.98	4.81	4.32

not lead directly to the answer, while Instruct produces slightly more bridge facts overall without producing more *true* ones. There is thus no evidence on Qwen that better fact quality explains the accuracy gap. Ministral behaves differently: its SFT student is higher on true-fact rate, bridge coverage and true$\wedge$bridge alike. Conditioning on correctness, true$\wedge$bridge is 5.76 versus 5.43 on correct traces and 2.12 versus 2.06 on incorrect ones, and on incorrect traces Reasoning has the *higher* true-fact rate (52.0% versus 44.4%), indicating that part of its errors occur in the reasoning or synthesis step after a correct fact has already surfaced.

Table 29: Gold recall hit rate and conditional accuracy.

	Qwen3-14B			Ministral-3-8B		
Metric	SFT-Base	Instruct	Δ	SFT-Base	Reasoning	Δ
Gold recall hit rate r	**.7888**	.7640	+2.48pp	**.7294**	.6998	+2.96pp
Accuracy $\mid$ gold recalled	.8915	.8912	+0.03pp	**.8889**	.8791	+0.98pp
Accuracy $\mid$ gold not recalled	.2254	.1962	+2.92pp	.1888	.1622	+2.66pp
Sample accuracy	.7508	.7272	+2.36pp	.6995	.6639	+3.56pp

Once the gold fact has surfaced, Qwen's two checkpoints are indistinguishable (.8915 versus .8912). At the question level, recall on at least one of the 10 samples is .8660 versus .8530 for Qwen and .8370 versus .8080 for Ministral. Qwen's SFT student uniquely recalls 23 questions against Instruct's 10, with 843 recalled by both and 124 by neither, and Ministral's SFT student uniquely recalls 51 against 22, with 786 recalled by both and 141 by neither.

Table 30: Retrieval difficulty over 1,000 questions (left) and the difficulty transition matrix across the two checkpoints of the same family (right). Difficulty is defined by the number of gold-recall hits over the 10 samples of a question: *easy* = all 10, *medium* = 1–9, *hard* = never.

	Qwen3-14B		Ministral-3-8B			$\rightarrow$ SFT-Base		
Difficulty	SFT	Instr.	SFT	Reas.	Source side	easy	med.	hard
easy	687	663	582	563	*Qwen* Instruct easy (663)	635	27	**1**
medium	179	190	255	245	medium (190)	**51**	130	9
hard	134	147	163	192	hard (147)	1	22	124
					Min. Reasoning easy (563)	506	56	**1**
					medium (245)	**73**	151	21
					hard (192)	3	48	141

Firmly encoded knowledge is almost never lost: exactly one question moves from easy to hard in each family. The net flow is concentrated in the middle band—Qwen moves 51 medium questions to easy against 9 to hard (net $+42$) and Ministral 73 against 21 (net $+52$)—and the hard band largely persists, with an overlap of 124/147 (84%) for Qwen and 141/192 (73%) for Ministral. Ministral has both a larger hard band and a larger difference between checkpoints, consistent with more marginal facts changing retrievability under a change of training regime at the smaller scale. Taken together, both families show post-training acting mainly on the retrieval difficulty of marginal facts rather than systematically erasing firm knowledge: Qwen's $+2.36$pp is almost entirely gold recall with identical post-recall behavior, whereas Ministral's $+3.56$pp is still retrieval-dominated but includes fact quality and conditional reasoning as well. Gold surfacing is an auditable string metric but can

miss semantically equivalent phrasings outside the alias list, which is why we read it jointly with the fact-quality audit, the conditional accuracies, and the question-level transition matrix rather than treating any single metric as complete mechanistic evidence.

K TEST-TIME BEHAVIOR OF THE INTERNALIZED CONFIDENCE

We adapt two diagnostics that Damani et al. (2025) apply to confidence-reward models and ask whether analogous behaviors can arise from SFT alone. We train no RL baseline and our traces carry no separate `<analysis>` segment, so these diagnostics only approximate the original protocol. They test whether analogous behaviors can arise without online RL in our setting. They do not constitute a controlled comparison with RLCR.[1]

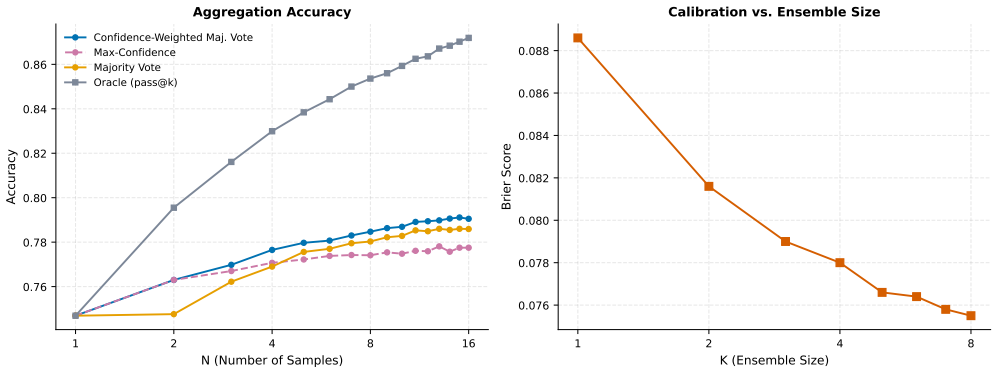

Figure 8: (a) Majority voting, max-confidence selection, confidence-weighted voting, and the oracle on TriviaQA as a function of the number of samples N. (b) Brier score of the confidence averaged over K traces of the same question that share an answer.

Confidence works better as a weight than as a target. On TriviaQA at $N{=}16$, majority voting, max-confidence selection, confidence-weighted voting and the oracle reach 78.59%, 77.75%, 79.05% and 87.19%, i.e. $+3.90$, $+3.06$, $+4.36$ and $+12.50$ points over $N{=}1$ (Figure 8a). Weighted voting leads plain vote counting. Picking the single most confident sample does not, and everything saturates past $N \geq 8$—the pattern §3.3 predicts, since confidence aggregates candidates and cannot invent ones the pool never produced.

Averaging the score over K traces that agree on an answer, across a fixed set of 900 questions, lowers Brier from 0.0886 at $K{=}1$ to 0.0755 at $K{=}8$, near the group-mean bound of 0.0748, while ECE holds at 0.033–0.036 (Figure 8b). Averaging removes variance, not within-bin bias. These traces were sampled independently and merely happened to agree, rather than being resampled with the answer fixed, so the check only approximates the original.

Two consistency checks close the analysis (Figure 9). Over 949 questions at $K{=}16$, 37.4% of scores vary by $\sigma < 0.02$ and 13.1% form a real long tail at $\sigma \geq 0.20$. Asked about mutually exclusive answers, the original Instruct model assigns confidences that sum to 2.287 on average and exceed 2 on 33.3% of questions, whereas Probe-SD sums to 1.383 on average and exceeds 2 on 13.3% of questions. Low variance alone proves nothing—saturated overconfidence is equally consistent—but together with the low ECE of §5 it rules that case out. These behaviors arise in our setting without online RL. We do not claim equivalence to an RL-trained model.

[1] RLCR (Qwen2.5-7B Base, trained on HotPotQA, so TriviaQA and SimpleQA are both OOD for it) reports TriviaQA Acc 60.8% / ECE 0.06 / AUROC 0.73 and SimpleQA Acc 12.1% / ECE 0.34 / AUROC 0.60. Our Probe-SD on Qwen3-14B reaches TriviaQA Acc 74.9% / ECE 0.024 / AUROC 0.928 and SimpleQA Acc 7.6% / ECE 0.113 / AUROC 0.671. Base model, training distribution, SimpleQA subset (RLCR uses all 4,326 questions, we use the 1,000-question Verified set) and grading protocol all differ, so these are order-of-magnitude reference points, not a controlled comparison.

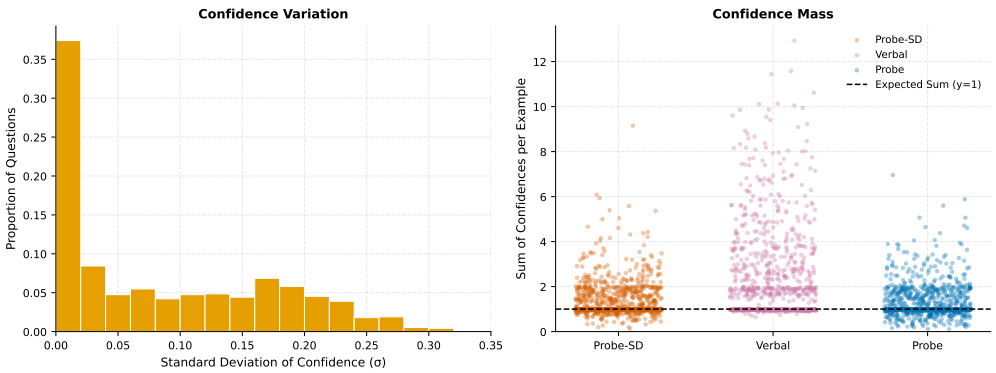

Figure 9: Self-consistency of the internalized confidence (§K). (a) Distribution of the per-question standard deviation of verbalized confidence over $K=16$ repeated samples of the same answer on TriviaQA. (b) Sum of the mean confidences assigned to mutually exclusive answers of the same question, for the original model and Probe-SD.

L LIMITATIONS

This appendix expands the limitations summarized in §7.

Scope. We validate on two open-weight reasoning families of at most 14B parameters and four factual QA benchmarks with short, checkable answers. Larger or closed models are untested, as is long-form or subjective generation—hard for a structural reason, since probe training needs stable trace-level correctness labels that such tasks do not supply. Replacing binary correctness with graded factuality is the natural next step. We also do not compare against RL methods on benefit or on training cost.

White-box access. The deployed student needs no probe, but constructing its training data requires hidden states for candidate traces and a per-model choice of layer and readout. The pipeline therefore cannot be built from a text-only interface.

Low base-rate tasks. As accuracy falls the probe keeps its probability scale but loses ranking power (§3.1), matching the general difficulty of uncertainty estimation on knowledge-intensive tasks (Tao et al., 2025). Calibration also acts only on candidates that were generated: if the pool contains no correct answer, no confidence estimate can create one. The ceiling on such tasks is set on the generation side, and calibration has to be combined with retrieval or stronger generation rather than used in its place.

Grader dependence. Probe labels and correctness metrics use Qwen3-32B-Instruct. A human audit confirmed 199/200 extractions and 198/199 answerable grading decisions. Correcting two labels in a separate 500-trace audit left Probe ECE unchanged and changed AUROC and Brier by less than 0.001 (Appendix D). These checks make large grading artifacts unlikely, but cannot rule out errors correlated with probe errors, such as shared failures on long-tail entities or ambiguous questions.