

CALIBER: Calibrating Confidence Before and After Reasoning in Language Models

Conor Finlay*, Joshua Kurien*, Saurabh Dash,
Marzieh Fadaee†, Beyza Ermis†

Cohere Labs

{conor.finlay, joshua.kurien, saurabh, marzieh, beyza}@cohere.com

Abstract

Reasoning language models are increasingly asked not only to answer difficult questions, but also to estimate their likelihood of success. Existing methods typically elicit confidence only once: either before thinking or after answering. We argue that confidence in reasoning models is state-dependent: before thinking, confidence should estimate whether the current policy will solve the prompt, while after thinking it should predict whether the realized answer is correct. This distinction determines the appropriate supervision target: prompt-level success should supervise confidence estimates made after seeing the prompt, while individual answer-level correctness should supervise confidence estimates made after answering. We introduce CALIBER (Calibration Before and After Reasoning), which elicits both estimates and supervises each with the target matched to its information state. Under this unified protocol, CALIBER reduces Expected Calibration Error (ECE) by 52.5% over the strongest single-confidence baseline on BigMathDigits for a 7B model, while achieving the best Brier score and AUROC, and remains within 2.0 points of the best accuracy. Out of distribution at this scale, it achieves the best ECE and Brier score on all three benchmarks, and ties for the best AUROC on SimpleQA. On a larger 30B model, CALIBER obtains the best ECE and Brier score on all four benchmarks and the best AUROC on three of the four. Ablations further show that this position–target alignment is most beneficial under distribution shift.

1 Introduction

Large language models increasingly solve complex tasks through extended thinking, and for many applications it is not enough that the model produces an answer; it should also provide a reliable estimate of whether its answer is correct. A well-calibrated

model signals to users how trustworthy its answers are, supports verification of low-confidence answers, and can be deployed more safely in settings where incorrect answers are costly (Geifman and El-Yaniv, 2017; Guo et al., 2017; Leng et al., 2025).

Prior work on LLM calibration explores several elicitation strategies, including token-probability calibration, sampling-based consistency, and explicit verbalized confidence (Kadavath et al., 2022; Lin et al., 2022; Kuhn et al., 2023; Xiong et al., 2024). We focus on verbalized confidence, where the model reports an explicit probability of correctness as part of its response, since such estimates can be easily extracted from reasoning models (Xu et al., 2024; Li et al., 2025; Hager et al., 2025).

Reasoning models follow an output structure that standard calibration methods are not designed to handle. They generate extended thinking traces before committing to a final response, so the model’s information state changes substantially between the start and end of generation. Calibration is conventionally defined as agreement between predicted probabilities and empirical correctness of a single answer (Guo et al., 2017), but for reasoning models the relevant probability concerns the correctness of a multi-step solution rather than a single next-token prediction (Xiong et al., 2024). Thus, confidence before and after thinking estimate distinct quantities: the former estimates whether the current policy is likely to solve the prompt, while the latter estimates whether the realized trace and answer are correct. Because these estimates are produced with access to different information, they require different supervision targets.

Recent methods address calibration in reasoning models through reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024; Wen et al., 2026), but differ in how they pair confidence positions with calibration targets. Methods vary along two axes: the *position* of the confidence estimate, either before thinking (Li et al., 2026) or after

*First authors.

†Principal senior advisors.

thinking (Damani et al., 2026; Zhang et al., 2026); and the *calibration target*, either instance-wise correctness (Damani et al., 2026), group-wise success across rollouts of the same prompt (Zhang et al., 2026; Li et al., 2026), or a hybrid of the two (Ma et al., 2026). Existing methods choose one position and one target, but do not ask whether that pairing matches the information state at which confidence is elicited.

We argue that the appropriate supervision target is determined by elicitation position: pre-thinking confidence, which conditions only on the prompt, should be supervised against the prompt-level success rate, while post-thinking confidence, which additionally conditions on the realized trace and answer, should be supervised against the correctness of that realized response. Conflating these introduces a mismatch between what the model is asked to estimate and what it is trained to predict.

We introduce CALIBER, a framework that elicits its confidence both before and after reasoning and supervises each estimate with the calibration target matched to its information state. Rather than forcing a single score to serve two roles, the framework lets each estimate specialize: pre-confidence captures policy-level solvability, while post-confidence captures answer-level correctness. The gap between them provides a per-example diagnostic of how thinking changes the model’s uncertainty.

Isolating the contribution of position–target alignment requires a controlled setting, since prior methods differ simultaneously in confidence position, calibration target, reward placement, and use of an accuracy reward. We therefore compare all combinations of position and target under a shared training and evaluation protocol. Across mathematical reasoning and question-answering benchmarks, CALIBER improves calibration while preserving task performance. On BigMathDigits, CALIBER reduces ECE relative to the strongest single-confidence baseline by 52.5% for the 7B model and achieves the best ECE for the 30B model, while staying within 2.0 percentage points of the best observed accuracy. These gains transfer out of distribution, where CALIBER obtains the best ECE and Brier score on all three OOD benchmarks at both scales.

Our work makes the following contributions:

- We frame confidence in reasoning models as state-dependent: pre-thinking and post-thinking confidence correspond to distinct un-

certainty quantities, policy-level solvability and answer-level correctness, and naturally align with group-wise and instance-wise calibration targets respectively.

- We introduce CALIBER, which emits both pre- and post-confidence estimates and supervises each with the target matched to its information state.
- We present a controlled comparison against prior calibration strategies under a shared protocol, isolating the contribution of position–target alignment from confounding design choices.
- We propose confidence-update diagnostics that compare pre- and post-thinking confidence as a per-example signal of how the model’s uncertainty evolves.

2 Related Works

Verbalized confidence in LLMs. LLMs can express uncertainty through token probabilities, sampling-based consistency, internal-state probes, or explicit verbalized confidence. Token probabilities are easy to obtain, but they measure uncertainty over next-token prediction rather than answer-level correctness (Xiong et al., 2024). Verbalized confidence instead asks the model to report a scalar estimate of the probability that its answer is correct. Prior work shows that models can be trained to express uncertainty in words (Lin et al., 2022), with later methods improving confidence through supervised fine-tuning, self-reflective rationales, confidence tuning, or distillation (Xu et al., 2024; Li et al., 2025; Hager et al., 2025). Other work suggests that models may encode information about their own correctness internally, even when this information is not reliably surfaced in their responses (Kadavath et al., 2022; Kumaran et al., 2026).

RLVR for calibrated verbalized confidence. A current trend for training calibrated reasoning models is to modify standard correctness objectives in RLVR to include calibration incentives (Shao et al., 2024; Wen et al., 2026). Reinforcement Learning with Calibration Rewards (RLCR) augments the RLVR correctness reward with a Brier-score reward that encourages the model’s verbalized confidence for a rollout to match that rollout’s binary correctness (Damani et al., 2026). Collaborative REASONER (COREA) instead trains confidence to

match the average correctness of a rollout group for the same prompt (Zhang et al., 2026). Decoupled Calibration Policy Optimization (DCPO) adopts a hybrid approach, interpolating between group-wise and instance-wise targets to account for both prompt-level group accuracy and individual rollout correctness (Ma et al., 2026). Co-optimized Confidence and Answers (CoCA) elicits confidence before the model produces its answer and trains it with a group-wise correctness target (Li et al., 2026). Both DCPO and CoCA utilize segmented credit assignment to compute and apply correctness and calibration advantages exclusively to the tokens that contribute to the respective rewards.

Prior RLVR methods vary in where confidence is elicited and how it is supervised (Damani et al., 2026; Zhang et al., 2026; Li et al., 2026; Ma et al., 2026). However, these choices have largely been studied in single-confidence settings. The complementary setting where both pre- and post-thinking confidences are elicited, and each matched to the target appropriate given its information state has not been systematically studied. CALIBER addresses this gap.

3 Preliminaries

This section defines the notation, evaluation metrics, and calibration targets used throughout the paper. We keep the main definitions concise and defer additional discussion of metric limitations and response formats to Appendix A.1.

3.1 Calibration Metrics

Let $p_i \in [0, 1]$ denote the model’s confidence that answer i is correct, and let $c_i \in \{0, 1\}$ denote correctness. A confidence estimator is calibrated if

$$\Pr(c_i = 1 \mid p_i = p) = p. \quad (1)$$

We evaluate both *absolute calibration* and *failure prediction* (Xiong et al., 2024). For absolute calibration, we report Expected Calibration Error (ECE) (Guo et al., 2017),

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (2)$$

and Brier score,

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n (p_i - c_i)^2. \quad (3)$$

For failure prediction, we report AUROC, treating confidence as a score for answer correctness. We

report all three metrics because average calibration and error discrimination capture complementary properties.

3.2 Confidence Conditioning and RLVR Targets

The position of a verbalized confidence estimate determines what information it can condition on. Let x be the prompt, z the thinking trace, $\hat{y}$ the model’s answer, and $c \in \{0, 1\}$ correctness. A confidence estimate emitted before thinking is question-conditioned,

$$p_{\text{question}} = P(c = 1 \mid x), \quad (4)$$

whereas one emitted after thinking and answering is answer-conditioned,

$$p_{\text{answer}} = P(c = 1 \mid x, z, \hat{y}). \quad (5)$$

In RLVR training, the current policy samples N rollouts for each prompt x , and each rollout receives a correctness label c_i (Wen et al., 2026; Shao et al., 2024). Calibration rewards compare predicted confidence p_i to a target t_i , typically with squared error. We consider three target types:

$$t_i^{\text{inst}} = c_i, \quad (6)$$

$$t^{\text{group}}(x) = \frac{1}{N} \sum_{j=1}^N c_j, \quad (7)$$

$$t_i^{\text{hybrid}} = \lambda c_i + (1 - \lambda) \frac{1}{N} \sum_{j=1}^N c_j. \quad (8)$$

The instance-wise target supervises answer-conditioned confidence, the group-wise target estimates prompt-level success under the current policy, and the hybrid target mixes both (Ma et al., 2026). CALIBER keeps instance and group-wise targets separate and assigns each to the confidence position whose information state matches it.

4 Method

We introduce CALIBER, a framework for training reasoning models to emit two calibrated verbalized confidence estimates in a single response. The model first emits a *pre-confidence* estimate before thinking, then produces a thinking trace and answer, and finally emits a *post-confidence* estimate. Each estimate is supervised according to the information available when it is emitted: pre-confidence with a group-wise target that reflects expected policy

success on the prompt, and post-confidence with an instance-wise target that reflects correctness of the realized answer.

4.1 State-Dependent Calibration

The key principle of CALIBER is that confidence should depend on the model’s information state. Existing calibration objectives use group-wise, instance-wise, or hybrid targets for a single confidence score, often treating them as competing supervision signals for the same underlying quantity. We instead view these targets as appropriate for different stages of thinking. Before thinking, the model has not yet produced a trace or answer, so its confidence cannot depend on the quality of a realized solution; it should estimate the expected success of the current policy on the prompt. After thinking, the model has produced a specific trace and answer, so its confidence can condition on the realized response and should estimate whether that particular rollout is correct. A target that does not match the confidence elicitation position therefore creates a mismatch between what the model can estimate and what it is trained to predict.

4.2 Setup and Notation

Our method is designed for RLVR training with GRPO-style rollout groups. For each prompt x , we sample N rollouts from the current policy:

$$\{r_i\}_{i=1}^N \sim \pi_\theta(\cdot \mid x). \quad (9)$$

Each rollout r_i contains a pre-confidence estimate $p_{\text{pre},i}$, a thinking trace z_i , an answer $\hat{y}_i$, and a post-confidence estimate $p_{\text{post},i}$. Let y^* denote the reference answer and let $V(\hat{y}, y^*) \in \{0, 1\}$ denote the task verifier. We define binary correctness as

$$c_i = V(\hat{y}_i, y^*), \quad (10)$$

and estimate the empirical success rate of the current policy on prompt x as

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^N c_i. \quad (11)$$

This group-wise quantity is shared by all rollouts for the same prompt.

4.3 Output Format

The model is prompted to produce a structured response with confidence estimates wrapped in special tags:

```
<pre_confidence>p_pre</pre_confidence>
<thinking> ... </thinking>
<answer> ... </answer>
<post_confidence>p_post</post_confidence>
```

The pre-confidence estimate is emitted before thinking, so it can condition only on prompt x . The post-confidence estimate is emitted after the thinking trace z_i and answer $\hat{y}_i$, so it conditions on the realized response. This implements Section 4.1’s state-dependent calibration: pre-confidence estimates prompt-level solvability, while post-confidence estimates answer-level correctness.

To ensure the model follows the required format, we include a format reward $R_{\text{fmt},i}$ that is granted only when all required fields are present, the answer can be extracted, and both confidence values are valid numbers in $[0, 1]$. All other reward components are applied only when the rollout is well formatted.

4.4 Calibration Targets

Each confidence estimate is assigned the calibration target that matches its position in the response.

Pre-confidence target. The pre-confidence estimate is trained with the group-wise empirical correctness target:

$$t_i^{\text{pre}} = \hat{p}(x) = \frac{1}{N} \sum_{j=1}^N c_j. \quad (12)$$

Since this target averages over all rollouts in the group, it is not conditioned on any particular answer. It therefore estimates the current policy’s success probability on prompt x , rather than the correctness of any particular rollout. This makes it the appropriate target for pre-confidence: before thinking, the model should estimate its expected success under the policy it is about to execute. Correct and incorrect rollouts sampled from the same prompt therefore receive the same pre-confidence target, so the target supervises prompt-level solvability rather than the correctness of a specific generated answer.

Post-confidence target. The post-confidence estimate is trained with the instance-wise correctness target:

$$t_i^{\text{post}} = c_i. \quad (13)$$

This target supervises whether rollout i produced the correct answer, and is therefore appropriate for post-confidence, which can condition on the realized thinking trace and generated answer.

Two consequences of this assignment are worth noting. First, the two confidence scores are not constrained to agree: $p_{\text{post},i}$ may rise above $p_{\text{pre},i}$ when the model has evidence that its solution is reliable, or fall below it when the generated solution appears uncertain or incorrect. Second, the gap $p_{\text{post},i} - p_{\text{pre},i}$ is itself a quantity of interest, since it summarizes how the model’s expressed uncertainty changes after thinking.

4.5 Training Objective

The calibration objective consists of the format reward, the pre-confidence group-wise reward, and the post-confidence instance-wise reward. In addition, we include the binary correctness of the model’s answer as a separate reward component. The reward for rollout i is

$$R_i = R_{\text{acc},i} + R_{\text{fmt},i} + R_{\text{grp},i}^{\text{pre}} + R_{\text{inst},i}^{\text{post}} \quad (14)$$

where the terms denote binary answer correctness ($R_{\text{acc},i} = c_i$), format validity ($R_{\text{fmt},i} = 1[\text{response is well formatted}]$), pre-confidence calibration ($R_{\text{grp},i}^{\text{pre}} = 1 - (p_{\text{pre},i} - \hat{p}(x))^2$), and post-confidence calibration ($R_{\text{inst},i}^{\text{post}} = 1 - (p_{\text{post},i} - c_i)^2$). The first calibration term trains the pre-confidence to match the prompt-level success rate, the second trains the post-confidence to match the correctness of the realized answer.

We retain an accuracy reward because training only for calibration can admit degenerate solutions: a model can produce incorrect answers with low confidence and still receive a high calibration reward (Li et al., 2026; Damani et al., 2026). The accuracy reward also ensures that calibration gains are not obtained by sacrificing task performance; the calibrated model should remain competitive with one trained using standard RLVR. Together, these design choices keep the two confidence estimates optimized for complementary roles rather than collapsing them into a single mixed target.

5 Experimental Setup

Models and Training. We run experiments using a 7B-parameter model and a 30B-parameter model, each trained for 500 calibration RL steps. The 7B model is based on the Command R7B architecture¹, and the 30B model is based on the North Mini Code architecture². Before calibration training,

¹<https://huggingface.co/CoherentLabs/c4ai-command-r7b-12-2024>

²<https://huggingface.co/CoherentLabs/North-Mini-Code-1.0>

all methods are first trained using RL with only the confidence-format reward, so that the model reliably follows the required confidence format. All calibration experiments then start from this shared format-learned checkpoint and use the same GRPO-based algorithm (without a KL penalty), the same number of rollouts per prompt, and the same reward filters.

Training CALIBER. We train CALIBER in two stages. We start with a *warmup* stage where we optimize only the pre-confidence group-wise calibration reward for a fixed number of steps. In the second stage, we use the remaining calibration steps to jointly train with the pre-confidence group-wise reward, the post-confidence instance-wise reward, and the correctness reward. The format reward is present in both stages. This staged schedule addresses an optimization difficulty that arises in this multi-reward setting when the pre-confidence reward is combined with the other rewards from the start. Under GRPO, the rewards for each rollout are summed and converted into group-relative advantages, so an individual reward’s influence depends on how much it varies across a prompt’s rollouts. The group-wise target $\hat{p}(x)$ is shared across a prompt’s rollouts, so if the model’s pre-confidence estimates vary little across rollouts, so do the corresponding pre-confidence rewards. In contrast, the post-confidence and correctness rewards are tied to per-rollout binary correctness and vary much more. As a result, pre-confidence receives a weak training signal, and the estimates collapse to a handful of values early in training and never recover. Training the pre-confidence reward first encourages enough diversity in pre-confidence estimates for all the individual rewards to train effectively in the second stage. We use fixed warmup lengths of 400 steps for the 7B model and 100 steps for the 30B model, selected because the range of pre-confidence estimates was consistently high by these points.

Baselines. We compare CALIBER against baselines that isolate two design choices: confidence position and calibration target. To keep the comparison controlled, no method uses segmented credit assignment—correctness and calibration advantages are not computed separately or applied to different token subsets—and all calibration rewards use squared error between predicted confidence and target. Methods labeled *-lite* are simplified adaptations of prior work under this shared setup: they preserve the original confidence position and

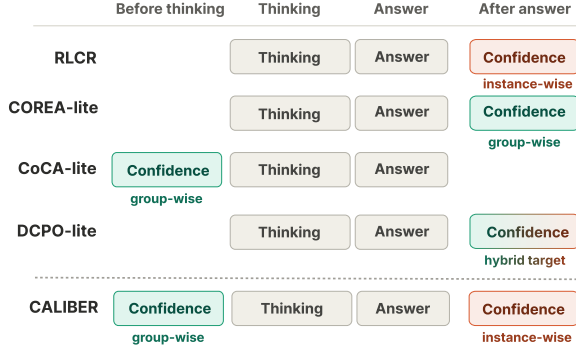


Figure 1: Confidence positions and calibration targets for the evaluated methods.

target choice, while omitting method-specific details that would confound comparison. Figure 1 summarizes the confidence format and calibration target for each method. **Base** is the format-trained checkpoint before calibration RL. **RLVR** is trained with an answer-correctness reward but no explicit calibration reward. **RLCR** adds post-confidence instance-wise calibration with an accuracy reward (Damani et al., 2026). **COREA-lite** uses post-confidence group-wise calibration with an accuracy reward (Zhang et al., 2026). **CoCA-lite** uses pre-confidence group-wise calibration with an accuracy reward (Li et al., 2026). **DCPO-lite** trains post-confidence calibration with a hybrid target that interpolates between group-wise and instance-wise supervision, alongside an accuracy reward (Ma et al., 2026). In our experiments, we set the interpolation weight to $\lambda = 0.5$, giving equal weight to both types of supervision.

Datasets. All methods are trained on the same data: a filtered subset of Big-Math (Albalak et al., 2025). Following Damani et al. (2026), we retain only problems with strictly numerical answers, enabling reliable verification, and further filter to problems that a LLaMA-8B model solves between 0% and 70% of the time. This ensures diverse correctness signals for calibration training. We refer to the resulting dataset as BigMathDigits.

We evaluate in-distribution performance on a held-out validation set from BigMathDigits. To assess out-of-distribution (OOD) generalization, we also evaluate on three question-answering datasets: TriviaQA (Joshi et al., 2017), SimpleQA (Wei et al., 2024), and GPQA (Rein et al., 2024). For TriviaQA, we use the 1000-question subset from Chhikara (2025); for SimpleQA, we use all 4326 samples and for GPQA, we use the 198-question

Diamond subset. For BigMathDigits, correctness is verified using heuristic and symbolic checks before falling back to an LLM-as-a-judge. GPQA is evaluated via regex matching against the correct multiple-choice letter, while SimpleQA and TriviaQA are evaluated using an LLM judge. We use Command A (Cohere et al., 2025) as our LLM judge and verify the robustness of this choice by re-evaluating our 7B SimpleQA and TriviaQA results with a different judge in Appendix A.7.

Evaluation Protocol. We evaluate checkpoints every 100 steps over the full 500-step training run and select the best-performing checkpoint for each method based on the BigMathDigits held-out validation set. Checkpoint selection uses the best combined rank across in-domain ECE and AUROC, with AUROC used to break ties. For methods that emit two confidence estimates, we evaluate only the post-thinking confidence estimate.

For each question, we sample responses using a temperature of 0.3 and top-p of 0.75. 3 responses are sampled per question for all datasets except GPQA, where we sample 12 responses with shuffled answer choices to cover different answer positions and reduce positional bias. Before computing metrics, we discard malformed generations, responses that do not follow the required confidence format, and responses with repetitive token loops.

Metrics. We report **ECE** for absolute calibration, **Brier score** as a proper scoring rule, **AUROC** for failure prediction (whether correct rollouts receive higher confidence than incorrect ones), and **Accuracy** to verify that calibration is not obtained at the cost of correctness. We consider these metrics jointly because a model can be well calibrated on average while still failing to distinguish correct answers from incorrect ones (Xiong et al., 2024). A prototypical example of this is where the model always predicts its dataset-average accuracy—all estimates land in one confidence bin, yielding near-perfect ECE but very poor AUROC.

6 Results

We evaluate CALIBER through comparisons designed to isolate the effects of confidence position, calibration target, and the two-confidence structure. These experiments address whether CALIBER improves calibration over pre-only or post-only confidence, whether the gains generalize out of distribution, whether matching targets to positions is

Method	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑
Base	0.208	0.208	0.644	0.759
RLVR	0.211	0.210	0.502	0.788
RLCR	0.092	0.142	0.753	0.785
COREA-lite	0.079	0.142	0.793	0.790
CoCA-lite	0.070	0.151	0.681	0.800
DCPO-lite	0.059	0.136	0.780	0.789
CALIBER	0.028	0.130	0.806	0.780

Table 1: BigMathDigits calibration and accuracy metrics for the 7B model.

Method	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑
Base	0.137	0.144	0.609	0.751
RLVR	0.162	0.166	0.528	0.829
RLCR	0.029	0.110	0.807	0.823
COREA-lite	0.039	0.114	0.787	0.827
CoCA-lite	0.049	0.143	0.656	0.818
DCPO-lite	0.027	0.115	0.811	0.817
CALIBER	0.019	0.108	0.821	0.820

Table 2: BigMathDigits calibration and accuracy metrics for the 30B model.

important, and whether the model meaningfully updates its uncertainty after observing its own thinking trace and answer.

6.1 In-Domain Calibration on BigMathDigits

We first evaluate on the held-out BigMathDigits split used for in-domain validation. As shown in Tables 1 and 2, our method CALIBER performs best across all calibration metrics (ECE, Brier, and AUROC) at both model scales, indicating both strong absolute calibration and discriminative power in its confidence estimates.

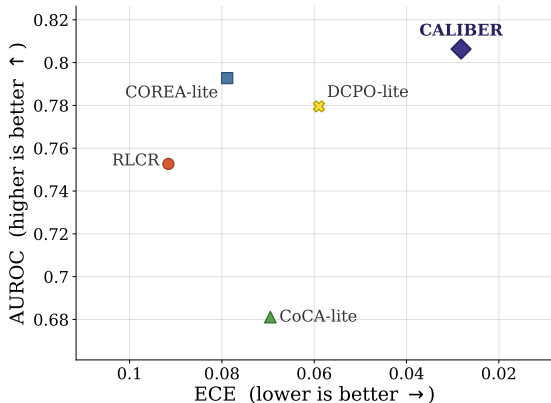


Figure 2: ECE–AUROC tradeoff on BigMathDigits. Lower ECE and higher AUROC are better. CALIBER achieves both the lowest ECE and highest AUROC among evaluated methods.

Figure 2 visualizes the ECE–AUROC tradeoff

for the 7B model: CALIBER occupies the best region, with both the lowest calibration error and highest failure-prediction performance. Importantly, these calibration gains do not come from substantially sacrificing task accuracy. Relative to standard RLVR, CALIBER’s accuracy changes by -0.8 and -0.9 percentage points for the 7B and 30B models. Eliciting confidence both before and after thinking thus improves absolute calibration while preserving competitive task performance.

When analyzing CALIBER’s confidence behavior, we find that it avoids overconfidence in-domain. The reliability diagrams (Figure 5 and Figure 6 in Appendix A.3) show that when CALIBER assigns high confidence to its answers, its empirical accuracy is usually well aligned with that confidence. For instance, in the highest-confidence bin at 7B, its mean confidence is 0.929 and its accuracy is 0.918.

Further, CALIBER uses its reasoning and answer to adjust its post-confidence estimate relative to its pre-confidence. Its post-confidence rises relative to pre-confidence by **4.96 percentage points** on average for correct responses and falls by **8.08 percentage points** for incorrect responses at 7B, demonstrating that the final confidence estimate is responsive to the realized thinking trace and answer. Appendix A.5 reports the full per-dataset breakdown (Table 9) and provides illustrative examples. We also evaluate the calibration of the pre-confidence estimate itself in Appendix A.4.

6.2 Out-of-Distribution Generalization

We next evaluate whether calibration gains transfer beyond the BigMathDigits training distribution. Results on GPQA, TriviaQA, and SimpleQA are reported in Table 3 for the 7B model and Table 4 for the 30B model, while Figure 4 summarizes the average ECE–AUROC tradeoff across the three OOD datasets for the 30B model.

CALIBER consistently improves absolute calibration under distribution shift, achieving the lowest ECE and Brier score on all three OOD datasets for both model sizes. The AUROC results show a tradeoff: COREA-lite or DCPO-lite sometimes obtain slightly stronger failure prediction on GPQA and TriviaQA at 7B, but at the cost of higher calibration error. CALIBER therefore provides the most reliable probability estimates on these OOD datasets while preserving competitive discrimination between correct and incorrect answers.

The TriviaQA reliability diagram for the 7B

Method	GPQA				TriviaQA				SimpleQA			
	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑
Base	0.419	0.415	0.584	0.357	0.276	0.269	0.763	0.629	0.720	0.585	0.592	0.051
RLVR	0.609	0.609	0.503	0.387	0.317	0.303	0.674	0.643	0.789	0.687	0.627	0.049
RLCR	0.149	0.265	0.601	0.386	0.161	0.198	0.814	0.635	0.546	0.399	0.655	0.049
COREA-lite	0.218	0.278	0.622	0.378	0.121	0.177	0.840	0.646	0.464	0.311	0.657	0.053
CoCA-lite	0.387	0.384	0.553	0.391	0.130	0.235	0.606	0.654	0.661	0.503	0.557	0.054
DCPO-lite	0.127	0.252	0.599	0.372	0.141	0.196	0.789	0.635	0.540	0.382	0.669	0.050
CALIBER	0.093	0.246	0.585	0.382	0.082	0.173	0.809	0.629	0.296	0.163	0.669	0.050

Table 3: Out-of-distribution calibration and accuracy metrics for the 7B model on GPQA, TriviaQA, and SimpleQA.

Method	GPQA				TriviaQA				SimpleQA			
	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑
Base	0.248	0.259	0.626	0.540	0.298	0.300	0.597	0.526	0.887	0.849	0.583	0.036
RLVR	0.400	0.399	0.503	0.589	0.418	0.418	0.509	0.564	0.942	0.936	0.506	0.050
RLCR	0.130	0.260	0.618	0.434	0.144	0.197	0.808	0.552	0.437	0.268	0.670	0.048
COREA-lite	0.090	0.212	0.733	0.587	0.183	0.197	0.868	0.561	0.492	0.295	0.683	0.048
CoCA-lite	0.186	0.272	0.547	0.575	0.213	0.289	0.535	0.554	0.727	0.577	0.494	0.050
DCPO-lite	0.100	0.232	0.697	0.557	0.119	0.177	0.839	0.549	0.374	0.193	0.599	0.048
CALIBER	0.052	0.200	0.757	0.562	0.089	0.164	0.858	0.555	0.342	0.172	0.707	0.049

Table 4: Out-of-distribution calibration and accuracy metrics for the 30B model on GPQA, TriviaQA, and SimpleQA.

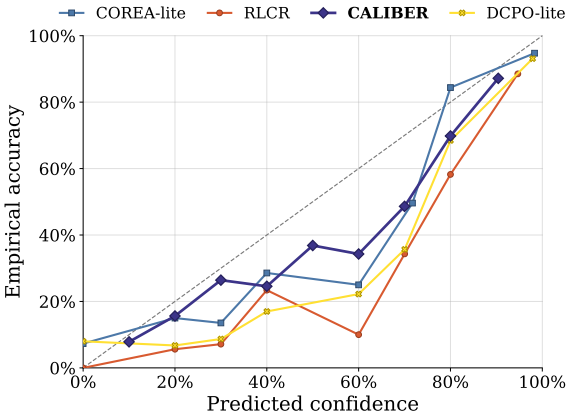


Figure 3: Reliability diagram for TriviaQA, omitting bins with fewer than 10 samples. CALIBER follows the diagonal more closely across much of the confidence range than the single-confidence baselines, consistent with its lower ECE and Brier score on TriviaQA.

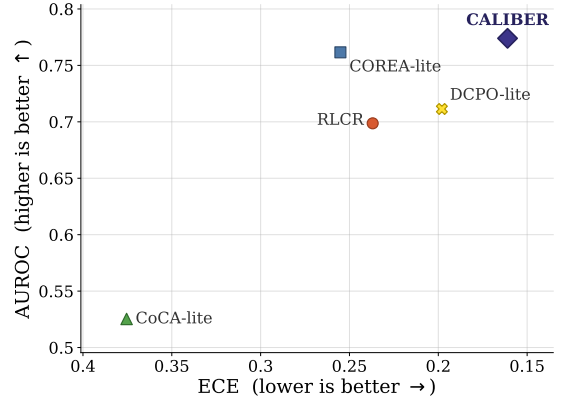


Figure 4: Average OOD ECE–AUROC tradeoff across GPQA, TriviaQA, and SimpleQA at 30B. Lower ECE and higher AUROC are better. CALIBER is best on both axes.

model in Figure 3 provides a complementary view of the aggregate calibration metrics. CALIBER’s curve follows the diagonal more closely across much of the confidence range than the single-confidence baselines, matching its lower ECE and Brier score in Table 3. At the same time, the remaining deviations from the diagonal show that CALIBER improves OOD calibration without fully eliminating miscalibration.

Figure 4 shows that calibration baselines often occupy different ends of the ECE–AUROC tradeoff, whereas CALIBER remains near the favorable re-

gion on both axes. Per-dataset paired comparisons against the strongest baseline on each metric, at both scales, are reported in Appendix A.2.

The confidence-update behavior observed in-domain is preserved out of distribution: CALIBER’s post-confidence rises by **4.97 percentage points** on average for correct responses and falls by **3.61 percentage points** for incorrect responses with the 7B model. See Appendix A.5, Table 9 for the per-dataset breakdown and illustrative qualitative examples.

Data	Method	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑
<i>7B</i>					
BMD	CALIBER	0.028	0.130	0.806	0.780
	Swapped	0.024	0.126	0.790	0.793
GPQA	CALIBER	0.093	0.246	0.585	0.382
	Swapped	0.123	0.241	0.600	0.354
TQA	CALIBER	0.082	0.173	0.809	0.629
	Swapped	0.128	0.193	0.800	0.650
SQA	CALIBER	0.296	0.163	0.669	0.050
	Swapped	0.529	0.354	0.640	0.046
<i>30B</i>					
BMD	CALIBER	0.019	0.108	0.821	0.820
	Swapped	0.054	0.154	0.735	0.809
GPQA	CALIBER	0.052	0.200	0.757	0.562
	Swapped	0.265	0.309	0.670	0.586
TQA	CALIBER	0.089	0.164	0.858	0.555
	Swapped	0.298	0.330	0.738	0.561
SQA	CALIBER	0.342	0.172	0.707	0.049
	Swapped	0.808	0.698	0.562	0.049

Table 5: Ablation comparing CALIBER and *Swapped* across all evaluation datasets for the 7B and 30B models. BMD, TQA, and SQA denote BigMathDigits, TriviaQA, and SimpleQA.

6.3 Ablation: Position–Target Alignment

To evaluate the role of position–target alignment, we run an ablation in which the confidence targets are swapped. In the swapped-rewards run, post-confidence is trained with the group-wise target and pre-confidence is trained with the instance-wise target. To mirror CALIBER’s schedule, we first train with the COREA-lite objective for the same warmup length used by CALIBER at each scale, which corresponds to applying a warmup to the component carrying the group-wise target, before training for the remaining steps with the full objective. Table 5 compares this swapped configuration against CALIBER across all evaluation datasets.

In-domain, the effect of alignment depends on model scale. For the 30B model, CALIBER substantially improves over Swapped, reducing ECE from 0.054 to 0.019, reducing Brier score from 0.154 to 0.108, and increasing AUROC from 0.735 to 0.821. For the 7B model, the result is more mixed: Swapped achieves slightly lower ECE and Brier score, while CALIBER achieves higher AUROC.

The clearest advantage of position–target alignment appears under distribution shift. Across all three OOD datasets and both model sizes, CALIBER achieves lower ECE than Swapped. The largest gaps appear for the 30B model, where alignment reduces ECE by 0.213 on GPQA, 0.209 on TriviaQA, and 0.466 on SimpleQA. The gains

also often extend beyond ECE: CALIBER improves Brier score on all OOD datasets except GPQA for the 7B model, and improves AUROC on all OOD datasets except GPQA for the 7B model. Overall, the ablation supports the main hypothesis that matching each confidence estimate to the target appropriate for its information state is especially important for OOD calibration.

7 Conclusion

Rather than treating confidence as a single scalar attached to an answer, CALIBER elicits confidence before and after thinking and supervises each estimate with the target appropriate to its information state. Pre-confidence is trained to estimate prompt-level success under the current policy, while post-confidence is trained to estimate the correctness of the realized answer. Across in-domain and out-of-distribution evaluations, CALIBER improves calibration while maintaining competitive task performance, and the pre/post confidence gap provides a diagnostic for how model uncertainty changes during thinking. The swapped-target ablation further supports the central claim that matching confidence position to supervision target is especially important under distribution shift.

Limitations

Baselines. Our study focuses on a controlled setting designed to isolate the role of confidence position and calibration target. This makes the comparison cleaner, but it also means that several method-specific enhancements from prior work, such as segmented credit assignment or alternative reward weighting schemes, are not included. Combining CALIBER with these techniques is a promising direction for future work.

Methodology. CALIBER requires a two-stage training schedule: pre-confidence training only for a warmup period (400 steps at 7B and 100 steps at 30B), followed by joint training with all rewards for the remaining steps. This was empirically necessary because joint training from step 0 yields uncalibrated pre-confidence, but the dynamics that lead to this failure mode are underexplored in our work. Additionally, our ablation isolates position–target alignment through a single alternative configuration (swapped targets) but other variants such as identical rewards at both positions, asymmetric reward weighting, or alternative warmup schedules, are not explored.

Evaluation. Verification on the OOD datasets, as well as the fallback on BigMathDigits, relies on a single LLM judge (Command A). LLM judges introduce verification noise, which we do not characterize. To assess robustness to the choice of judge, we repeat selected evaluations with an alternative judge in Appendix A.7. Although the results are consistent across judges, perfect verification cannot be guaranteed. Additionally, each method is trained once per model scale. Additional seeds were infeasible, as the 7B experiments alone required over 11,000 H100 GPU-hours. We characterize evaluation uncertainty in Appendix A.2 with confidence intervals computed by resampling prompts, but this does not account for variance across training runs.

Beyond Two-Point Elicitation. By eliciting confidence estimates before and after reasoning, CALIBER implements the simplest two-point realization of a more general principle: uncertainty in language models evolves throughout reasoning, and supervision should track this evolution. Future work should explore eliciting confidence at additional intermediate states, such as across a multi-turn interaction, and consider how the supervision should adapt in response.

References

- Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. 2025. [Big-Math: A large-scale, high-quality math dataset for reinforcement learning in language models](#). *Preprint*, arXiv:2502.17387.
- Prateek Chhikara. 2025. [Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models](#). *Transactions on Machine Learning Research*.
- Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammam, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, and 210 others. 2025. [Command A: An enterprise-ready large language model](#). *Preprint*, arXiv:2504.00698.
- Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2026. [Beyond binary rewards: Training LMs to reason about their uncertainty](#). In *The Fourteenth International Conference on Learning Representations*.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 4885–4894, Red Hook, NY, USA. Curran Associates Inc.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- Sophia Hager, David Mueller, Kevin Duh, and Nicholas Andrews. 2025. [Uncertainty distillation: Teaching language models to express semantic confidence](#). *Preprint*, arXiv:2503.14749.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations*.
- Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Veličković. 2026. [How do LLMs compute verbal confidence?](#) In *Forty-third International Conference on Machine Learning*.
- Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2025. [Taming overconfidence in LLMs: Reward calibration in RLHF](#). In *The Thirteenth International Conference on Learning Representations*.
- Changcheng Li, Jiancan Wu, Hengheng Zhang, Zhengsu Chen, Guo An, Junxiang Qiu, Xiang Wang, and Qi Tian. 2026. [Confidence before answering: A paradigm shift for efficient LLM uncertainty estimation](#). *Preprint*, arXiv:2603.05881.
- Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. 2025. [ConfTuner: Training large language models to express their confidence verbally](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 53484–53513. Curran Associates, Inc.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching models to express their uncertainty in words](#). *Transactions on Machine Learning Research*.

Zhengzhao Ma, Xueru Wen, Boxi Cao, Yaojie Lu, Hongyu Lin, Jinglin Yang, Min He, Xianpei Han, and Le Sun. 2026. [Decoupling reasoning and confidence: Resurrecting calibration in reinforcement learning from verifiable rewards](#). In *Forty-third International Conference on Machine Learning*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level Google-Proof Q&A benchmark](#). In *First Conference on Language Modeling*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *CoRR*, abs/2402.03300.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. [Measuring short-form factuality in large language models](#). *Preprint*, arXiv:2411.04368.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2026. [Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs](#). In *The Fourteenth International Conference on Learning Representations*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs](#). In *The Twelfth International Conference on Learning Representations*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaozhe Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. [SaySelf: Teaching LLMs to express confidence with self-reflective rationales](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

Chuang Zhang, Zizhen Zhu, Yihao Wei, Bing Tian, Junyi Liu, Henan Wang, Wang Xavier, and Yaxiao Liu. 2026. [Confidence-calibrated small-large language model collaboration for cost-efficient reasoning](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4480–4501, Rabat, Morocco. Association for Computational Linguistics.

A Appendix

A.1 Additional Preliminaries

This appendix provides additional detail on the calibration metrics, verbalized-confidence formats, and

RLVR calibration targets summarized in Section 3.

A.1.1 Verbalized Confidence Formats

Verbalized confidence methods ask a model to output an explicit probability of correctness together with its response. For reasoning models, the response is often represented as a structured sequence containing a thinking trace, an answer, and one or more confidence estimates. A standard post-answer confidence format is:

```
<thinking> ... </thinking>
<answer> ... </answer>
<confidence> ... </confidence>
```

In this format, the confidence estimate is emitted after the model has already produced both its thinking trace and final answer. In CALIBER, we instead elicit confidence before and after thinking, as described in Section 4.

A.1.2 Calibration Metrics

Expected Calibration Error (ECE) measures absolute calibration by binning predictions according to confidence and comparing each bin’s average confidence to its empirical accuracy. This is useful for measuring whether confidence values match empirical correctness, but ECE can be misleading on its own. For example, a model that predicts its dataset-average accuracy for every example can appear well calibrated while providing little information about which answers are likely to be wrong.

The Brier score is a proper scoring rule that measures squared error between confidence and correctness. AUROC measures failure prediction by treating confidence as a score for answer correctness and evaluating whether correct answers tend to receive higher confidence than incorrect ones. We report ECE, Brier, AUROC, and accuracy jointly because no single metric captures all relevant aspects of calibrated confidence.

A.1.3 RLVR, Rollouts, and Calibration Rewards

Reinforcement learning with verifiable rewards (RLVR) trains models on tasks where answer correctness can be checked automatically, such as mathematical reasoning (Wen et al., 2026). In GRPO-style training, for each prompt x , the current policy samples N responses, or *rollouts*, and computes rewards for each rollout relative to the group (Shao et al., 2024). We use this rollout group both to train answer correctness and to estimate prompt-level solvability.

For a prompt x , let $c_i \in \{0, 1\}$ indicate whether rollout i is correct. Calibration rewards compare a predicted confidence p_i to a target t_i , typically using a squared-error reward such as

$$R_{\text{cal},i} = 1 - (p_i - t_i)^2. \quad (15)$$

The instance-wise target $t_i^{\text{inst}} = c_i$ uses the correctness of the individual rollout and is appropriate for answer-conditioned confidence. The group-wise target

$$t^{\text{group}}(x) = \frac{1}{N} \sum_{j=1}^N c_j \quad (16)$$

uses the empirical accuracy of all rollouts for the same prompt and estimates the current policy’s success probability on that prompt. Hybrid targets interpolate between instance-wise and group-wise supervision:

$$t_i^{\text{hybrid}} = \lambda c_i + (1 - \lambda) \frac{1}{N} \sum_{j=1}^N c_j. \quad (17)$$

These targets differ in the information they condition on. CALIBER uses this distinction by assigning the group-wise target to pre-confidence and the instance-wise target to post-confidence.

A.2 Per-Dataset Comparison Against the Strongest Baseline

Tables 6 and 7 compare CALIBER to the strongest baseline on each dataset using paired prompt-level bootstrap differences, for the 7B and 30B models respectively. The strongest baseline is selected per metric and per dataset, so the comparator can differ across rows. At 7B, CALIBER achieves a statistically significant ECE improvement on all four benchmarks, while AUROC falls significantly below the strongest baseline on GPQA and TriviaQA and the difference is not statistically significant elsewhere. At 30B, CALIBER significantly improves ECE on all three OOD datasets and significantly improves AUROC on GPQA, with no significant AUROC or ECE loss anywhere.

A.3 BigMathDigits Reliability Diagrams

A.4 Pre-Confidence Calibration

The main results evaluate post-confidence, since this is the most precise estimate which takes into account the realized thinking trace and answer. However, to evaluate our dual-confidence formulation and verify that the group-wise target is doing any work, it’s important to observe the performance of CALIBER’s pre-confidence estimates.

Data	Metric	Comp.	Δ [95% CI]
BMD	ECE	DCPO	-0.031 [-0.047, -0.012]
	AUROC	COREA	+0.013 [-0.015, +0.043]
	Brier	DCPO	-0.006 [-0.016, +0.003]
	Acc.	CoCA	-0.020 [-0.032, -0.009] [†]
GPQA	ECE	DCPO	-0.034 [-0.059, -0.006]
	AUROC	COREA	-0.037 [-0.064, -0.009] [†]
	Brier	DCPO	-0.006 [-0.016, +0.003]
	Acc.	CoCA	-0.009 [-0.028, +0.011]
TQA	ECE	COREA	-0.040 [-0.062, -0.019]
	AUROC	COREA	-0.030 [-0.051, -0.010] [†]
	Brier	COREA	-0.004 [-0.014, +0.007]
	Acc.	CoCA	-0.025 [-0.041, -0.010] [†]
SQA	ECE	COREA	-0.168 [-0.175, -0.162]
	AUROC	DCPO	+0.000 [-0.029, +0.027]
	Brier	COREA	-0.148 [-0.154, -0.143]
	Acc.	CoCA	-0.004 [-0.008, +0.000]

Table 6: Paired bootstrap differences ($\Delta = \text{CALIBER} - \text{comparator}$) against the strongest baseline for the 7B model. The comparator is selected per metric per dataset, so it can differ across rows. Intervals are BCa 95% confidence intervals from prompt-level cluster-resampled replicates on the prompts shared by both methods. Negative is better for ECE and Brier; positive is better for AUROC and accuracy. **Bold** marks a significant result favoring CALIBER; [†] marks a significant result favoring the comparator.

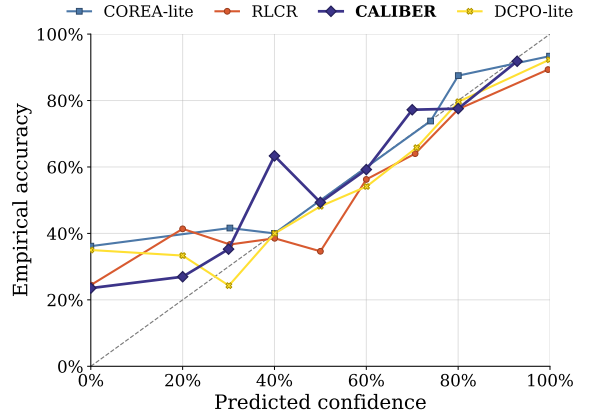


Figure 5: Reliability diagram for BigMathDigits at 7B, omitting bins with fewer than 10 samples. CALIBER tends to avoid overconfidence, especially in high-confidence bins, and achieves high accuracy when predicting high confidence in-domain.

Table 8 therefore reports ECE, Brier score, and AUROC for CALIBER’s pre-confidence at both scales. Since the existing baselines appear in Tables 1–4, we include only CALIBER and RLVR’s post-confidence estimates as reference points.

The table demonstrates that CALIBER’s pre-confidence estimates are substantially better calibrated than the RLVR baseline’s, showing that

Data	Metric	Comp.	Δ [95% CI]
BMD	ECE	DCPO	-0.008 [-0.023, +0.007]
	AUROC	DCPO	+0.010 [-0.014, +0.034]
	Brier	RLCR	-0.002 [-0.008, +0.003]
	Acc.	RLVR	-0.010 [-0.020, -0.000] [†]
GPQA	ECE	COREA	-0.037 [-0.065, -0.016]
	AUROC	COREA	+0.024 [+0.001, +0.046]
	Brier	COREA	-0.012 [-0.019, -0.004]
	Acc.	RLVR	-0.027 [-0.047, -0.007] [†]
TQA	ECE	DCPO	-0.032 [-0.056, -0.009]
	AUROC	COREA	-0.010 [-0.027, +0.007]
	Brier	DCPO	-0.013 [-0.019, -0.005]
	Acc.	RLVR	-0.008 [-0.022, +0.006]
SQA	ECE	DCPO	-0.031 [-0.036, -0.026]
	AUROC	COREA	+0.023 [-0.002, +0.049]
	Brier	DCPO	-0.020 [-0.022, -0.017]
	Acc.	RLVR	-0.001 [-0.005, +0.003]

Table 7: Paired bootstrap differences (Δ = CALIBER – comparator) against the strongest baseline for the 30B model. The comparator is selected per metric per dataset, so it can differ across rows. Intervals are BCa 95% confidence intervals from prompt-level cluster-resampled replicates on the prompts shared by both methods. Negative is better for ECE and Brier; positive is better for AUROC and accuracy. **Bold** marks a significant result favoring CALIBER; [†] marks a significant result favoring the comparator.

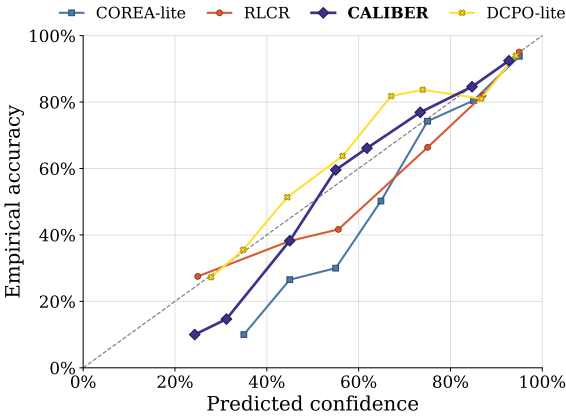


Figure 6: Reliability diagram for BigMathDigits at 30B, omitting bins with fewer than 10 samples. As at 7B, CALIBER avoids overconfidence, particularly in high-confidence bins.

the group-wise target for pre-confidence is indeed doing some work.

A.5 Confidence Updates

Training with CALIBER results in a model that increases its confidence more after reasoning on questions it ends up getting correct than on those it answers incorrectly. This demonstrates that the model is using the additional context of its thinking

Data	Confidence	ECE ↓	Brier ↓	AUROC ↑
7B				
BMD	RLVR	0.211	0.210	0.502
	CALIBER pre	0.037	0.149	0.742
	CALIBER post	0.028	0.130	0.806
GPQA	RLVR	0.609	0.609	0.503
	CALIBER pre	0.135	0.263	0.534
	CALIBER post	0.093	0.246	0.585
TQA	RLVR	0.317	0.303	0.674
	CALIBER pre	0.049	0.212	0.688
	CALIBER post	0.082	0.173	0.809
SQA	RLVR	0.789	0.687	0.627
	CALIBER pre	0.325	0.171	0.601
	CALIBER post	0.296	0.163	0.669
30B				
BMD	RLVR	0.162	0.166	0.528
	CALIBER pre	0.059	0.133	0.721
	CALIBER post	0.019	0.108	0.821
GPQA	RLVR	0.400	0.399	0.503
	CALIBER pre	0.095	0.231	0.681
	CALIBER post	0.052	0.200	0.757
TQA	RLVR	0.418	0.418	0.509
	CALIBER pre	0.198	0.265	0.686
	CALIBER post	0.089	0.164	0.858
SQA	RLVR	0.942	0.936	0.506
	CALIBER pre	0.585	0.396	0.557
	CALIBER post	0.342	0.172	0.707

Table 8: Calibration of CALIBER’s pre-confidence estimate, against RLVR and against CALIBER’s own post-confidence, at both model scales. BMD, TQA, and SQA denote BigMathDigits, TriviaQA, and SimpleQA.

trace and answer effectively to update its confidence.

This is shown in more detail in Table 9, which lists the CALIBER-trained 7B model’s average pre- and post-thinking confidence for each dataset, separated based on correctness. In all cases, the change in confidence after thinking is substantially more positive for correct answers than for incorrect ones.

We also list several illustrative examples of the 7B model updating its confidence in light of the realized thinking trace and answer – one correct and one incorrect response from each of BigMathDigits and TriviaQA. In the correct examples, the model increases its confidence after thinking through the problem and, conversely, in the incorrect examples, the model decreases its confidence in light of its thinking trace and answer. This represents the behavior we want to induce with CALIBER: updating confidence as the information state changes.

BigMathDigits

Data	Subset	Pre	Post	Δ
BMD	Correct	0.785	0.835	+0.050
	Incorrect	0.626	0.545	-0.081
GPQA	Correct	0.528	0.466	-0.062
	Incorrect	0.510	0.408	-0.102
TQA	Correct	0.641	0.813	+0.171
	Incorrect	0.514	0.540	+0.027
SQA	Correct	0.429	0.468	+0.040
	Incorrect	0.372	0.339	-0.033

Table 9: Mean pre- and post-thinking confidence for CALIBER on correct and incorrect responses at 7B, by dataset. BMD, TQA, and SQA denote BigMathDigits, TriviaQA, and SimpleQA.

Q. There are 20 points on a circle, paired and connected by 10 chords which do not share endpoints and do not intersect. In how many ways can this be done?

A. 16796 ✓

$p_{\text{pre}}=0.30 \rightarrow p_{\text{post}}=0.70$ ($\Delta=+0.40$)

“This seems like a combinatorial problem, maybe related to Catalan numbers? I remember that Catalan numbers count the number of non-crossing partitions, like parentheses or triangulations. [...] For n pairs, the n -th Catalan number gives the number of non-crossing chord diagrams. So $n=10$. Let me check: $C_{10} = \frac{1}{11} \binom{20}{10}$. $\binom{20}{10}$ is 184756, so 184756 divided by 11 is 16796. [...] But wait, I should make sure that this applies here. The problem states that the chords do not share endpoints and do not intersect. That’s exactly the condition for Catalan numbers. [...] So C_{10} is indeed 16796.”

Q. A bricklayer and an electrician together spend some hours working on a new house. The bricklayer earns \$12 per hour, the electrician earns \$16 an hour, and the owner pays them a total of \$1350 for their work. If each worker spends 22.5 hours on the house, how many hours do they spend working together?

A. 48.21 ✗

$p_{\text{pre}}=0.70 \rightarrow p_{\text{post}}=0.30$ ($\Delta=-0.40$)

“The total payment is \$1350. The bricklayer’s rate is \$12/hour, electrician’s is \$16/hour. Each worked 22.5 hours. [...] The total cost should be $(12 \cdot 22.5) + (16 \cdot 22.5) = 1350$. Let me calculate. $12 \cdot 22.5$ is 270, $16 \cdot 22.5$ is 360. $270 + 360 = 630$. Wait, that’s only \$630, but the owner paid \$1350. That doesn’t add up. So there’s a problem here. [...] Either the hours are different, or there’s a mistake in the problem.”

TriviaQA

Q. Which country will host the 2014 FIFA World Cup?

A. Brazil ✓

$p_{\text{pre}}=0.30 \rightarrow p_{\text{post}}=0.90$ ($\Delta=+0.60$)

“I remember that the 2014 World Cup was a recent event, maybe around 2014. Wait, wasn’t Brazil the host? Let me verify. The FIFA World Cup is held every four years, and the 2010 edition was in South Africa. Then 2014 would be the next one. Yes, Brazil hosted it. I’m pretty sure about that. [...] Brazil is a big football nation, and they had a lot of infrastructure for the tournament. I think that’s correct.”

Q. 1986 film starring Goldie Hawn as a high-school football coach.

A. Cousin Oliver ✗

$p_{\text{pre}}=0.80 \rightarrow p_{\text{post}}=0.30$ ($\Delta=-0.50$)

“Goldie Hawn was in several films in the 80s. One that comes to mind is ‘Cousin Oliver’ from 1987, but that’s after 1986. [...] ‘Private Benjamin’ was 1980. ‘Sleepless in Seattle’ was 1993. ‘Overboard’ 1987. ‘Fletch’ 1985. [...] If I can’t recall, maybe the answer is ‘Cousin Oliver’ from 1986, but I’m not entirely sure. [...] Let’s assume that the correct answer is ‘Cousin Oliver’ from 1986, even though I’m not 100% sure.”

A.6 Confidence Distribution

Compared to the other methods, CALIBER spreads its confidence distribution across a wider range of bins for both in-domain and out-of-domain datasets. For instance, in Figure 7, CALIBER allocates noticeably less mass to the highest-confidence bin than RLCR, and places nontrivial mass across all the bins between 20% and 90%.

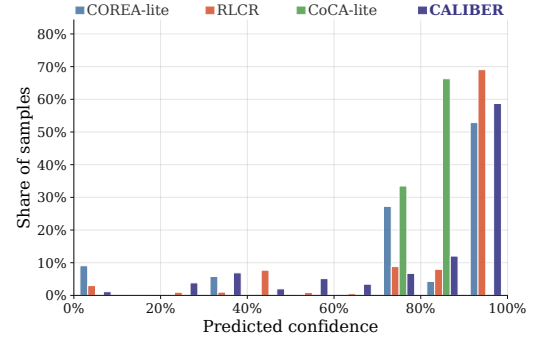


Figure 7: Confidence distribution of CALIBER and baselines on BigMathDigits at 7B.

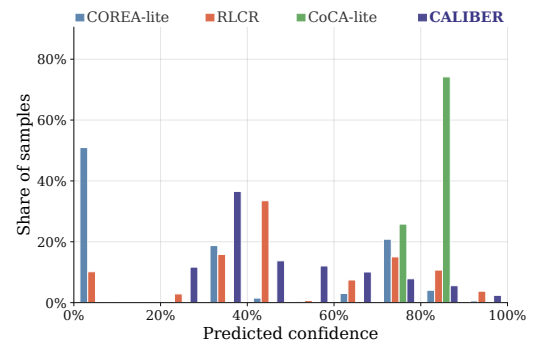


Figure 8: Confidence distribution of CALIBER and baselines on GPQA at 7B.

A.7 SimpleQA and TriviaQA Evaluations with an Alternative Judge

To demonstrate that our results are robust to the choice of judge, we re-evaluate the methods using the 7B model on TriviaQA and SimpleQA with

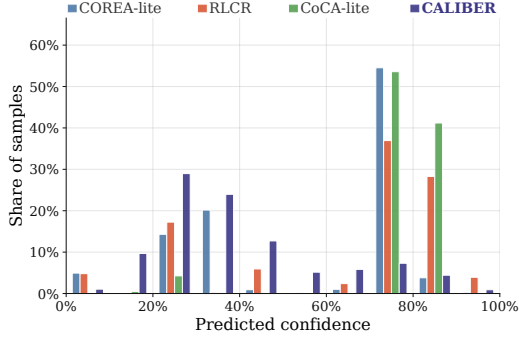


Figure 9: Confidence distribution of CALIBER and baselines on SimpleQA at 7B.

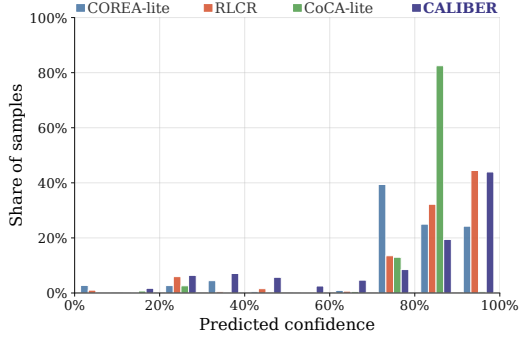


Figure 10: Confidence distribution of CALIBER and baselines on TriviaQA at 7B.

Gemma 4 31B as the judge. The results are shown in Tables 10 and 11, respectively. Consistent with our results with the Command A judge, CALIBER has the best ECE and Brier score on TriviaQA and the best ECE, Brier score, and AUROC on SimpleQA.

	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑
RLCR	0.165	0.197	0.823	0.634
COREA-lite	0.128	0.174	0.850	0.640
CoCA-lite	0.143	0.243	0.603	0.642
DCPO-lite	0.143	0.197	0.797	0.634
CALIBER	0.067	0.163	0.823	0.640

Table 10: TriviaQA results for the 7B model using Gemma 4 31B as the judge.

	ECE ↓	Brier ↓	AUROC ↑	Acc. ↑
RLCR	0.548	0.399	0.677	0.049
COREA-lite	0.470	0.316	0.651	0.051
CoCA-lite	0.661	0.503	0.563	0.055
DCPO-lite	0.542	0.382	0.675	0.048
CALIBER	0.294	0.160	0.684	0.051

Table 11: SimpleQA results for the 7B model using Gemma 4 31B as the judge.