REINFORCEMENT LEARNING DONE RIGHT FOR EXPLICIT CONFIDENCE TARGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can produce plausible but incorrect answers when they are uncertain, which can mislead users and undermine model reliability. Training models to abstain with reinforcement learning is a natural approach, but most existing methods use a fixed error penalty and therefore cannot adapt their behavior to different confidence requirements at inference time. We study *explicit confidence targets*, where a user specifies a target p in the prompt and the model directly chooses whether to answer or abstain, with the goal of achieving an accuracy of at least p among answered questions. However, learning to respond to different confidence targets with a single model has proven difficult, and prior attempts have failed to distinguish different targets. We show that this failure comes from two problems: the reward signal that distinguishes different values of p is often lost during optimization, while the original error penalty grows without bound as p approaches one. We propose **RLEC** (**R**einforcement **L**earning with **E**xplicit **C**onfidence Targets), which preserves the learning signal needed to distinguish confidence targets, introduces a stable reward family that keeps the desired answer condition unchanged, and uses reward scheduling to balance effective learning and training stability. Together, these designs enable the first successful training of a single LLM for explicit confidence target control. Extensive experiments across diverse benchmarks and model scales show that RLEC largely preserves general capabilities and achieves competitive calibration performance. RLEC also supports fine-grained target control without an external decision system or threshold calibration after training. By making explicit confidence targets part of the trained policy, RLEC opens a new and more direct path toward behavioral calibration. Our code is available at: `https://anonymous.4open.science/r/RLEC-ICLR-48E5`.

1 INTRODUCTION

Large language models (LLMs) Achiam et al. (2023); Comanici et al. (2025); Team et al. (2026) have shown strong abilities in knowledge-intensive and reasoning tasks, but they can still generate plausible yet incorrect responses, commonly known as hallucinations Ji et al. (2023); Huang et al. (2025). Such errors reduce the reliability of LLMs and may mislead users, posing particular risks when incorrect answers are costly. Rather than guessing when uncertain, a reliable model should know when not to answer Kamath et al. (2020); Kadavath et al. (2022); Kalai et al. (2025).

Kalai et al. (2025) argue that common training reward guessing over acknowledging uncertainty. A natural solution is to train abstention with reinforcement learning Wei et al. (2025); Mohamadi et al. (2025); Jha et al. (2026); Pan et al. (2026), rewarding correct answers, penalizing incorrect answers, and assigning no reward to abstentions. For example, TruthRL Wei et al. (2025) uses such a reward with a fixed error penalty. However, a fixed penalty represents only one risk preference, while different applications may re-

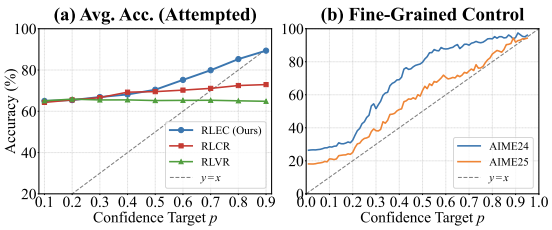


Figure 1: **Explicit Confidence Target control.** (a) RLEC's average accuracy across eight benchmarks increases with p specified in the prompt. (b) RLEC's fine-grained control on AIME with p.

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.

quire different levels of reliability. To represent these requirements, we study an *explicit confidence target*: the user specifies a target p in the prompt, and the model directly chooses whether to answer or abstain. The desired behavior is

$$\mathrm{Acc}(p) = \Pr(\text{correct} \mid \text{answer}, p) \geq p, \tag{1}$$

and a single model should respond to different values of p without retraining or an external decision system. However, learning such behavior has proved difficult. Wu et al. (2025) propose Explicit Risk Thresholding (ERT), which trains on randomly sampled values of p using GRPO Shao et al. (2024); Guo et al. (2025), assigning rewards of 1, $-p/(1-p)$, and 0 to correct answers, incorrect answers, and abstentions, respectively. However, they report that the model's hallucination rates remain insensitive to the input target. It attributes this failure to the noisy optimization problem created by training across many thresholds. Other work Mohamadi et al. (2025) further argues that limited model capacity may force a policy to specialize in one risk preference. These findings leave it unclear whether the difficulty is inherent to a single model or caused by the training method.

We show that this failure does not mean that the task is beyond the capacity of a single model. Instead, it can be explained by two problems in the training setup. First, ERT, like many recent RL methods for LLMs, uses GRPO. We observe that most of its rollout groups contain only one or two distinct reward values. Groups with a single reward value yield zero advantages, while in groups with two values, standard deviation normalization removes the effect of the reward scale induced by p on the advantages. The resulting advantages depend only on the fraction of the two values, rather than their magnitudes. As a result, reward magnitudes provide little direct signal for distinguishing confidence targets. Second, the original error reward grows without bound as p approaches one. Once the reward magnitudes directly affect the advantages, samples with large p can dominate the gradient and destabilize training. Based on these findings, we introduce **RLEC** (**R**einforcement **L**earning with **E**xplicit **C**onfidence Targets). RLEC removes division by the group standard deviation while retaining mean centering, thereby preserving the effect of the reward scale induced by p on the advantages. It further introduces a reward family controlled by a stability coefficient b, which controls reward magnitudes at large p without changing the desired decision rule. Finally, we observe that fixed values of b tend to trade off the strength of the learning signal against training stability. We therefore introduce a schedule for b that combines a strong signal for learning different confidence targets with stable training.

We evaluate RLEC using Qwen3-4B-Base and Qwen3-30B-A3B-Base Yang et al. (2025) across eight benchmarks covering mathematical reasoning, knowledge and hallucination, and scientific reasoning. Our results show that a single model can successfully learn different behaviors for different explicit confidence targets, as shown in Figure 1(a). We also compare RLEC with a standard RLVR Guo et al. (2025) baseline, showing that RLEC largely preserves the baseline's general capabilities at both model scales. For additional context, we further compare RLEC with RLCR Damani et al. (2026), which focuses on confidence reporting rather than direct abstention control. RLEC achieves calibration performance comparable to RLCR at 4B and obtains a lower overconfidence rate and a higher overall score at 30B. RLEC also provides several practical advantages. Methods based on verbalized confidence Tian et al. (2023); Damani et al. (2026) often require an external rule to convert their confidence estimates into answer or abstention decisions, while other selective prediction methods Wang et al. (2026); Kamath et al. (2020) further rely on separate calibrators or held-out calibration sets. In contrast, RLEC receives p in the prompt and directly produces the final answer or abstention without external processing. It also supports fine-grained control over answer or abstention, with smooth changes across densely spaced values of p, as illustrated in Figure 1(b).

Our main contributions are as follows:

- We identify how the optimization algorithm and reward design used in prior work weaken the confidence target signal, providing a concrete explanation for the previously reported failure.
- We propose RLEC, which preserves the confidence target signal, introduces a stable reward family that keeps the desired answer condition unchanged, and uses reward scheduling to balance effective learning and training stability. To our knowledge, RLEC thereby enables the first successful training of a single LLM for explicit confidence target control.
- Extensive experiments across diverse benchmarks and model scales show that RLEC largely preserves general capabilities and achieves competitive calibration performance. It also requires no external decision system or calibration after training and supports fine-grained target control.

2 BACKGROUND

2.1 FIXED-PENALTY ABSTENTION

To mitigate hallucinations Ji et al. (2023); Huang et al. (2025), a reliable model should know when to answer and when to abstain. Given a question x, the model chooses an action $a \in \{\text{ANS}, \text{ABS}\}$, where ANS denotes answering and ABS denotes abstaining. We primarily evaluate on verifiable question answering tasks following prior work Wei et al. (2025); Damani et al. (2026). This setup yields three outcomes: a correct answer, an incorrect answer, or an abstention. We treat an incorrect answer as a hallucination. We define accuracy among answered questions as

$$\text{Acc} = \Pr(\text{correct} \mid a = \text{ANS}). \tag{2}$$

Recent work Wei et al. (2025); Mohamadi et al. (2025); Jha et al. (2026) has used reinforcement learning with verifiable rewards (RLVR) Shao et al. (2024); Guo et al. (2025) to directly train LLMs to decide whether to answer or abstain . A common formulation assigns a positive reward to a correct answer, a neutral reward to abstention, and a fixed negative reward to an incorrect answer:

$$R_\lambda(a, y) = \begin{cases} 1, & a = \text{ANS and } y \text{ is correct}, \\ 0, & a = \text{ABS}, \\ -\lambda, & a = \text{ANS and } y \text{ is incorrect}, \end{cases} \qquad \lambda \geq 0. \tag{3}$$

TruthRL Wei et al. (2025) optimizes this reward with $\lambda = 1$ using Group Relative Policy Optimization (GRPO) Shao et al. (2024), whereas Reinforced Hesitation Mohamadi et al. (2025) trains separate models for different values of λ. Let $q(x)$ be the true conditional probability that the model's intended answer to x is correct. Under Equation 3, the expected reward of answering is

$$\begin{aligned} \mathbb{E}[R_\lambda \mid x, a = \text{ANS}] &= q(x) - \lambda\big(1 - q(x)\big) \\ &= (1 + \lambda)q(x) - \lambda. \end{aligned} \tag{4}$$

Since abstention yields zero reward, the policy answers when

$$q(x) > \frac{\lambda}{1 + \lambda}, \tag{5}$$

abstains when the inequality is reversed. Equivalently, λ determines the fixed confidence threshold

$$p_\lambda = \frac{\lambda}{1 + \lambda}. \tag{6}$$

Under the idealized assumptions that $q(x)$ gives the true probability of correctness and that the policy exactly follows Equation 5, its selective accuracy satisfies

$$\begin{aligned} \Pr(\text{correct} \mid a = \text{ANS}) &= \mathbb{E}[q(x) \mid a = \text{ANS}] \\ &= \mathbb{E}[q(x) \mid q(x) > p_\lambda] \geq p_\lambda. \end{aligned} \tag{7}$$

This characterizes idealized behavior, not an empirical guarantee, since the learned policy may not optimize or generalize perfectly and its uncertainty may not match the true probability of correctness.

Training with a fixed penalty provides an interpretable trade-off between accuracy and coverage, but each λ represents only one operating point. Adapting to different error costs therefore generally requires retraining the model Mohamadi et al. (2025); Jha et al. (2026) or selecting a different model.

2.2 EXPLICIT CONFIDENCE TARGETS

Explicit confidence targets Kalai et al. (2025) extend fixed-penalty abstention by allowing users to specify the required accuracy in the prompt. Let $p \in [0, 1)$ denote this target. Setting

$$\lambda(p) = \frac{p}{1 - p} \tag{8}$$

in Equation 3 gives

$$R_p = \begin{cases} 1, & \text{correct answer}, \\ -\dfrac{p}{1 - p}, & \text{incorrect answer}, \\ 0, & \text{abstention}. \end{cases} \tag{9}$$

Suppose the probability of answering correctly is $q(x)$. The expected reward for answering is

$$\begin{aligned}\mathbb{E}[R_p \mid x, \text{ANS}] &= q(x) - \frac{p}{1-p}\big(1 - q(x)\big) \\ &= \frac{q(x) - p}{1 - p}.\end{aligned} \tag{10}$$

Abstention receives zero reward. A policy that maximizes expected reward therefore answers when $q(x) > p$, abstains when $q(x) < p$, and is indifferent when $q(x) = p$. Thus, p directly controls the decision between answering and abstaining. When $p = 0$, answering is at least as good as abstaining for every question and is strictly better whenever $q(x) > 0$. We use this endpoint to represent the setting in which the user requires the model to answer, with ties resolved in favor of answering.

Our goal is to train a single policy $\pi_\theta(\cdot \mid x, p)$ that follows arbitrary confidence targets without retraining. For every p at which the model answers with nonzero probability, the desired behavior is

$$\text{Acc}(p) := \Pr(\text{correct} \mid \text{ANS}, p) \geq p. \tag{11}$$

The model should answer when doing so yields positive expected reward, rather than defaulting to abstention on every question. Importantly, p is a target specified by the user in the prompt, and the model directly chooses whether to answer or abstain. Inference therefore requires no external system to compare a reported confidence score Damani et al. (2026) with a threshold.

Wu et al. (2025) propose *Explicit Risk Thresholding* (ERT) to study this approach by sampling p during training, including it in the prompt, and using the reward in Equation 9. However, their hallucination rates remain largely unchanged across different p, showing that the model does not adapt its behavior to the target. They attribute this failure to a noisy optimization problem caused by sampling targets across the full range, making it difficult to learn a consistent strategy for different p. Mohamadi et al. (2025) also argues that a single model cannot support all risk preferences because limited model capacity prevents it from optimizing every error penalty simultaneously. Together, these findings illustrate the difficulty of training one model to follow arbitrary confidence targets.

3 METHOD

3.1 SIGNAL PRESERVATION

We first identify a key problem in prior work: most of the direct reward signal distinguishing different confidence targets is lost during training. ERT uses GRPO to optimize the reward in Equation 9. For each prompt, GRPO samples G responses and estimates their advantages from relative rewards. Let R_i denote the reward of response i. The corresponding advantage is

$$A_i = \frac{R_i - \overline{R}}{\sigma_R}, \qquad \overline{R} = \frac{1}{G}\sum_{j=1}^{G} R_j, \qquad \sigma_R = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_j - \overline{R})^2}. \tag{12}$$

However, this standardization can remove the effect of the reward scale induced by p on the advantages. At a fixed p, consider a rollout group with two distinct reward values $r_{\text{H}} > r_{\text{L}}$. Let $\rho \in (0, 1)$ denote the fraction of responses receiving r_{H}. The group mean and standard deviation are

$$\overline{R} = \rho r_{\text{H}} + (1 - \rho) r_{\text{L}}, \qquad \sigma_R = (r_{\text{H}} - r_{\text{L}})\sqrt{\rho(1 - \rho)}. \tag{13}$$

Substituting Equation 13 into Equation 12 gives

$$A_{\text{H}} = \sqrt{\frac{1-\rho}{\rho}}, \qquad A_{\text{L}} = -\sqrt{\frac{\rho}{1-\rho}}. \tag{14}$$

The resulting advantages are independent of p, so the training signal carried by the reward scale of p disappears. For example, in a group with two distinct reward values, a higher p imposes a larger reward penalty on an incorrect answer, but does not make its standardized advantage more negative. Also, for a group with one reward value, all advantages in Equation 12 are computed as zero in our implementation, so the group provides no reward signal. Only groups containing all three outcomes can retain the learning signal from changes in the reward scale induced by p.

Figure 2 shows how the three group types evolve during training. Groups with three reward values are rare and become less frequent, while groups with one value eventually dominate and most others contain two. Thus, only a small fraction of groups retains a direct signal from reward magnitudes that vary with p, explaining why the policy remains insensitive to the confidence target.

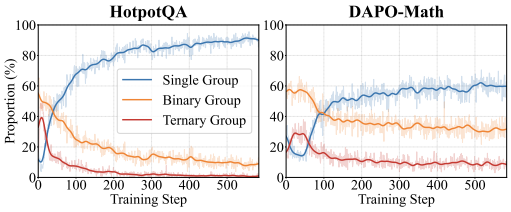


Figure 2: **Fractions of rollout groups.**

To preserve the signal, we remove the group standard deviation while retaining mean centering:

$$\widetilde{A}_i = R_i - \overline{R}. \tag{15}$$

Without standard deviation normalization, a reward gap that varies with p produces advantages that vary accordingly, preserving the direct reward signal across confidence targets. This also exposes the large penalty in Equation 9, which we address next.

3.2 Stable Rewards

Removing the group standard deviation preserves the effect of reward magnitudes on the advantages, but also makes policy updates sensitive to those magnitudes. In Equation 9, the rewards for a correct answer and abstention remain fixed at 1 and 0, while the reward for an incorrect answer is $-p/(1-p)$, which tends to $-\infty$ as p approaches one. Groups containing both incorrect answers and other outcomes can therefore produce large advantages. This creates two related problems.

Gradient dominance. Because the policy shares parameters across confidence targets, such groups at large p can dominate the overall gradient, reducing the influence of lower targets.

Training instability. The number and composition of such groups vary across batches. At large p, their centered advantages can cause sharp fluctuations in gradient norms and destabilize training.

Multiplying all rewards by a single constant scales the centered advantages equally across confidence targets, leaving their relative influence unchanged. We instead use a positive factor that decreases with p to reduce the penalty at large p without changing the theoretical decision to answer or abstain.

A simple way to bound the error penalty is to multiply all rewards by $1-p$:

$$(1-p)R_p = \begin{cases} 1-p, & \text{correct answer,} \\ -p, & \text{incorrect answer,} \\ 0, & \text{abstention.} \end{cases} \tag{16}$$

The magnitude of the error penalty is now bounded by one. This scaling also changes the reward for a correct answer to $1-p$, while abstention remains at zero.

To control how strongly rewards at large p are reduced, we introduce a family of rewards:

$$R_p^{(b)} = (1-p)^b R_p = \begin{cases} (1-p)^b, & \text{correct answer,} \\ -p(1-p)^{b-1}, & \text{incorrect answer,} \\ 0, & \text{abstention,} \end{cases} \qquad b \geq 0. \tag{17}$$

Here, b is a stability coefficient. For a fixed $p \in (0,1)$, increasing b reduces the magnitude of the error penalty and the reward for a correct answer. In particular, $b=0$ recovers Equation 9, while $b=1$ recovers Equation 16. Since the same scaling applies to all responses in a rollout group, it also scales their centered advantages:

$$\widetilde{A}_i^{(b)} = (1-p)^b \widetilde{A}_i^{(0)}. \tag{18}$$

Thus, larger values of b reduce the relative influence of groups with large p.

Importantly, scaling all rewards at a given p by the same positive factor does not change the theoretical decision to answer or abstain. Suppose the model answers correctly with probability $q(x)$. Its expected reward for answering is

$$\begin{aligned} \mathbb{E}\Big[R_p^{(b)} \mid x, \text{ANS}\Big] &= q(x)(1-p)^b - \big(1-q(x)\big)p(1-p)^{b-1} \\ &= (1-p)^{b-1}\big(q(x)-p\big). \end{aligned} \tag{19}$$

Since $(1-p)^{b-1} > 0$ for $p < 1$, answering has positive expected reward exactly when $q(x) > p$. Thus, every value of b preserves the same theoretical answer condition.

3.3 Reward Scheduling

The stability coefficient b balances the strength of the learning signal against training stability. For any fixed $p \in (0, 1)$, the centered advantages in Equation 18 become smaller in magnitude as b increases. This reduction is strongest at large p, where the original error penalty is largest. It can improve stability, but it also weakens the signal for learning when to abstain by reducing the error penalty relative to abstention. Our ablations suggest that keeping b fixed at a relatively large value, such as $b = 1$, can make it harder for the model to learn when to abstain and to vary its behavior across confidence targets. Details are shown in Section 4.4.

By contrast, a smaller b, such as $b = 0$, preserves stronger centered advantages at large p, helping the model learn when to abstain and distinguish confidence targets more effectively early in training. However, keeping b small throughout training can lead to gradient dominance and training instability, as discussed in Section 3.2. These effects may harm final performance and general capabilities.

Guided by this trade-off, we linearly increase b from 0 to 1 over T training steps. At step t, we set

$$b_t = \frac{t}{T}, \qquad 0 \le t \le T. \tag{20}$$

This schedule preserves a stronger learning signal early while improving training stability later.

4 Experiment

In this section, we conduct experiments to answer the following research questions:

- **(RQ1)** Can RLEC successfully distinguish different explicit confidence targets?
- **(RQ2)** How does RLEC compare with prior methods in calibration and general capabilities?
- **(RQ3)** What enables RLEC to distinguish targets while preserving general capabilities?
- **(RQ4)** What other properties and practical advantages does RLEC offer?

4.1 Experimental Setup

Implementation Details. Our main experiments and ablation studies are based on Qwen3-4B-Base Yang et al. (2025), and we further conduct experiments on Qwen3-30B-A3B-Base to evaluate scalability. Our training data combine DAPO-Math Yu et al. (2026) and a modified version of HotpotQA Yang et al. (2018); Damani et al. (2026). We randomly sample 15,000 examples from each dataset, resulting in 30,000 training examples, and train for five epochs. We set $p = 0$ with probability 10%. Otherwise, we uniformly sample p from the interval $[0, 0.9]$ and round it to two decimal places. During evaluation, we test $p \in \{0, 0.1, \ldots, 0.9\}$ and report results averaged over multiple sampled generations. Additional details are provided in Appendix B.1.

Benchmarks. Our evaluation covers eight benchmarks from three domains and different difficulty levels, including **(1) mathematical reasoning:** GSM8K Cobbe et al. (2021), MATH-500 Hendrycks et al. (2021); Lightman et al. (2024), and AIME 2024; **(2) knowledge and hallucination:** HotpotQA test set Yang et al. (2018); Damani et al. (2026), AA-Omniscience Jackson et al. (2025), and SimpleQA-Verified Wei et al. (2024); Haas et al. (2025); and **(3) scientific reasoning:** MMLU-Pro Wang et al. (2024) and GPQA Rein et al. (2023). Details are provided in Appendix B.2.

Baselines. We compare RLEC with three types of methods: **(1) RLVR:** standard RLVR with a binary reward; **(2) abstention methods:** methods that use a fixed error penalty or fail to learn behavior that varies with p, including TruthRL Wei et al. (2025) and ERT Wu et al. (2025), compared in RQ1; and **(3) confidence reporting methods:** RLCR Damani et al. (2026), included in RQ2 as an additional reference for calibration performance under a different formulation.

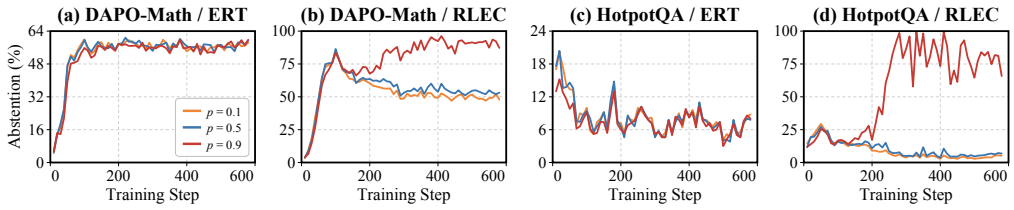


Figure 3: **Abstention rates on held-out validation sets during training.** Each curve evaluates the same training checkpoints with a fixed confidence target p. ERT produces nearly overlapping curves for different targets, while RLEC learns higher abstention rates for higher values of p.

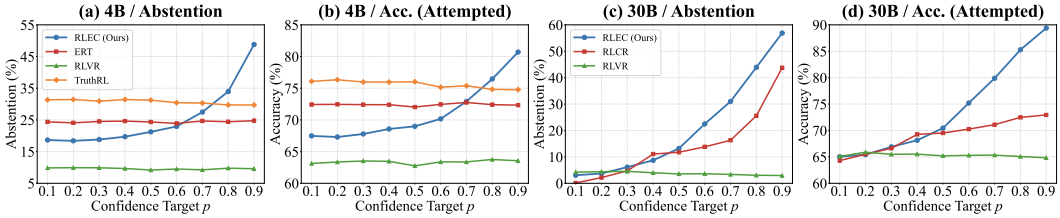


Figure 4: **Adaptation to explicit confidence targets.** (a,b) Direct decision methods with Qwen3-4B-Base. (c,d) RLEC, RLCR, and RLVR with Qwen3-30B-A3B-Base.

4.2 Adaptation to Confidence Targets

To answer **RQ1**, we study whether models can change their behavior according to p. We first compare the training behavior of RLEC and ERT, and then compare the final accuracy and abstention rates of RLEC, RLVR, ERT, and TruthRL across all evaluation benchmarks.

Training Curves. Figure 3 shows the abstention rate on held-out validation sets during training, evaluated with $p \in \{0.1, 0.5, 0.9\}$. On both HotpotQA and DAPO-Math, our reproduction of ERT shows the same failure reported in the original study Wu et al. (2025): its curves for different values of p almost completely overlap. In contrast, RLEC produces a clear dependence on p, with higher confidence targets generally leading to higher abstention rates. These results show that RLEC successfully learns different behaviors for different confidence targets.

Test Results. Figure 4(a,b) reports average abstention rates and answer accuracy across all benchmarks using Qwen3-4B-Base. Among methods that directly decide whether to answer, only RLEC changes clearly with p: as p increases from 0.1 to 0.9, its abstention rate rises from approximately 19% to 49%, while its accuracy among answered questions rises from approximately 68% to 81%. RLVR, ERT, and TruthRL remain nearly unchanged. Thus, RLEC successfully learns to adapt its decision to answer or abstain across confidence targets, answering RQ1.

4.3 Comparison with Calibration Methods

As shown in Section 4.2, existing methods that directly decide whether to answer or abstain show little sensitivity to the confidence target. To compare calibration performance and answer **RQ2**, we therefore compare it with RLCR Damani et al. (2026), a different type of method that reports a numerical confidence c together with its answer. Since RLCR cannot directly follow a user provided target p, we apply an external rule that retains the answer when $c \geq p$ and replaces it with an abstention otherwise, whereas RLEC conditions its answer or abstention decision directly on p.

Qualitative Analysis. Figure 4(c,d) compare RLEC, RLCR, and RLVR across confidence targets using Qwen3-30B-A3B-Base. RLVR changes little with p, whereas RLEC and RLCR increase their abstention rates as p increases. RLEC responds more strongly, especially at high p, and its accuracy

Table 1: Forced answer accuracy, average overconfidence (AOC), and average score across model scales. The best Micro and Macro results within each model scale are shown in bold.

			Mathematical Reasoning			Knowledge & Hallucination			Scientific Reasoning		Avg.	
Model	Metric	Method	GSM8K	MATH	AIME	Hotpot	OmniSci	SimpleQA	MMLU-Pro	GPQA	Micro	Macro
Qwen3-4B-Base	Acc. ↑	RLVR	93.33	69.70	20.83	69.50	13.75	5.20	68.47	35.91	**58.18**	**47.09**
		RLCR	93.10	73.90	16.25	66.30	15.58	4.75	66.94	34.00	57.35	46.35
		RLEC	92.15	70.50	19.79	67.60	14.17	4.75	67.37	36.24	57.40	46.57
	AOC ↓	RLVR	0.000	0.027	0.258	0.033	0.292	0.389	0.032	0.175	0.049	0.151
		RLCR	0.090	0.090	0.090	0.098	0.229	0.396	0.090	0.169	0.090	0.156
		RLEC	0.000	0.000	0.090	0.010	0.259	0.381	0.015	0.157	**0.013**	**0.114**
	Score ↑	RLVR	0.480	0.261	-0.162	0.237	-0.173	-0.213	0.237	-0.090	0.167	0.072
		RLCR	0.482	0.342	0.053	0.255	-0.004	-0.094	0.285	0.069	**0.235**	**0.173**
		RLEC	0.477	0.333	0.056	0.264	-0.048	-0.107	0.262	-0.016	0.214	0.153
Qwen3-30B-A3B-Base	Acc. ↑	RLVR	95.07	77.10	26.88	68.90	18.17	9.40	76.25	46.53	63.97	52.29
		RLCR	94.73	77.70	26.25	66.20	19.08	8.90	77.22	46.76	**64.24**	52.10
		RLEC	94.77	78.60	27.29	66.60	20.00	8.30	76.73	47.09	64.14	**52.42**
	AOC ↓	RLVR	0.000	0.014	0.221	0.038	0.265	0.372	0.020	0.108	0.045	0.130
		RLCR	0.000	0.003	0.000	0.038	0.270	0.367	0.000	0.057	0.025	0.092
		RLEC	0.000	0.003	0.000	0.009	0.209	0.368	0.002	0.072	**0.001**	**0.083**
	Score ↑	RLVR	0.504	0.329	-0.169	0.222	-0.198	-0.309	0.300	0.026	0.195	0.088
		RLCR	0.500	0.360	0.096	0.218	-0.173	-0.311	0.355	0.123	0.236	0.146
		RLEC	0.504	0.353	0.091	0.258	-0.015	-0.132	0.336	0.112	**0.261**	**0.188**

among answered questions follows the target boundary more closely. These results suggest that directly optimizing explicit confidence targets enables clearer adaptation to the requested target.

Quantitative Analysis. We use three metrics to evaluate answer ability, target violations, and utility. *Forced Accuracy* measures answer accuracy when abstention is disabled. We set $p = 0$ for RLEC and retain all RLCR answers, testing whether calibration training harms answer ability.

Average Overconfidence (AOC) measures the average magnitude of target violations:

$$\mathrm{AOC} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \max\left\{0,\ p - \mathrm{Acc}(p)\right\}, \tag{21}$$

where lower values indicate better compliance with the requested accuracy. Since AOC can be reduced by abstaining, it does not capture utility. We therefore also report

$$\mathrm{Score} = \frac{1}{N|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{i=1}^{N} z_{i,p}\left(y_{i,p} - p\right), \tag{22}$$

where $z_{i,p}$ indicates whether the model answers and $y_{i,p}$ indicates whether the answer is correct. A correct answer contributes $1 - p$, an incorrect answer contributes $-p$, and an abstention contributes zero. A higher Score indicates greater utility by accounting for both correctness and coverage.

For RLEC, each Score contribution is exactly its reward with $b = 1$. For RLCR, twice the corresponding score integrated over all confidence targets equals its reward, as proved in Appendix D.

Table 1 reports Micro results pooled over all examples and Macro results averaged equally across benchmarks. Micro reflects aggregate performance and is less affected by difficult benchmarks such as SimpleQA-Verified, whereas Macro prevents larger benchmarks from dominating. RLEC largely preserves the answer ability obtained by RLVR, with comparable Forced Accuracy at both scales. Compared with RLCR, the 4B model achieves lower AOC but also lower Score, indicating a trade-off between target compliance and utility. At 30B, RLEC achieves both lower AOC and higher Score, outperforming RLCR in both respects. These results demonstrate the feasibility of direct training with explicit confidence targets and show that RLEC remains effective at the larger scale.

4.4 Ablation Studies

To answer **RQ3**, we study two main design choices in RLEC: omitting the reward standard deviation from group advantage estimation and selecting the stability coefficient b.

Standard Deviation. As shown in Figure 5(b), adding group standard deviation normalization produces nearly overlapping abstention curves across different p. This supports our analysis that

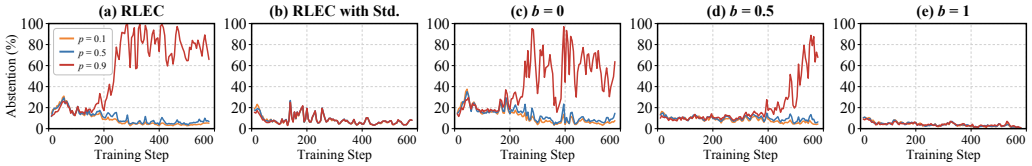


Figure 5: **Ablation training curves.** Abstention rates at $p \in \{0.1, 0.5, 0.9\}$ for (a) RLEC, (b) RLEC with standard deviation normalization, and fixed (c) $b = 0$, (d) $b = 0.5$, and (e) $b = 1$.

normalization weakens the reward scale signal induced by p. Also, Appendix C.2 shows that removing the group standard deviation does not consistently improve or degrade answer ability.

Choice of Stability Coefficient. Panels (c), (d), and (e) of Figure 5 illustrate the different behaviors induced by fixed choices of b. With $b = 0$, the model distinguishes the confidence targets relatively early, but its abstention rate fluctuates sharply. With $b = 1$, the curves nearly overlap, and the model fails to distinguish the targets. The intermediate choice $b = 0.5$ produces separation only late in training, indicating less efficient learning and potentially less reliable target differentiation.

We also evaluate the different choices on the test benchmarks. Forced Accuracy, AOC, and Score remain broadly consistent, demonstrating that the proposed reward family is robust to the specific choice of b. The scheduled configuration achieves the highest aggregate Forced Accuracy while retaining competitive calibration performance. Because we consider preserving answer ability a primary requirement for behavioral calibration, we select it as our default configuration. Complete results and further discussion are provided in Appendix C.1.

4.5 Additional Properties

To answer **RQ4**, we examine several additional properties of RLEC. We first evaluate fine-grained confidence control on AIME24 and AIME25 using targets spaced by 0.01 and extending beyond the training range to $p = 0.95$. Each target is evaluated with Avg@1024.

As shown in Figure 6, the abstention rate rises and attempted accuracy generally improves as p increases. Although small fluctuations occur between adjacent targets, both metrics follow stable overall trends across the dense target grid. Training uses targets only up to $p = 0.90$, yet these trends continue through $p = 0.95$ without an abrupt behavioral change, suggesting that RLEC remains responsive beyond the training range.

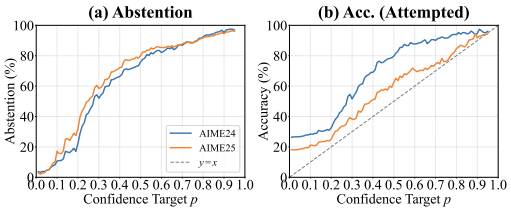


Figure 6: **Fine-grained confidence control.**

For comparison, the evaluated RLCR model produces confidence values only on a one-decimal grid. External thresholding therefore cannot recover operating points at the same resolution, whereas RLEC responds directly to targets spaced by 0.01. More detailed analysis of fine-grained confidence control is provided in Appendix C.3. In addition, we examine the implicit confidence distribution and risk–coverage behavior of RLEC, as detailed in Appendices C.6 and C.7.

5 Conclusion

We study explicit confidence targets as a more flexible and direct way to control LLM abstention. We show that the failure of prior work results from the loss of signals and the unstable scale of reward. Based on this analysis, we propose RLEC, which preserves the confidence target signal, introduces a stable reward family, and uses reward scheduling to balance effective learning and training stability. Experiments across diverse benchmarks and model scales show that explicit confidence target training is feasible. RLEC largely preserves general capabilities, achieves strong calibration, and supports fine-grained control without an external decision system. More broadly, RLEC opens a new and direct path toward behavioral calibration with explicit, user-specified confidence targets.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improve the clarity, organization, and presentation of the manuscript. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The method, reward design, and training objective are described in Section 3. Section 4.1 specifies the models, datasets, baselines, and evaluation protocol, while Appendix B provides further implementation details, prompt templates, benchmark descriptions, and additional results. We also provide our implementation through the anonymous link in abstract.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Michael Bereket and Jure Leskovec. Uncalibrated reasoning: Grpo induces overconfidence for stochastic outcomes. *arXiv preprint arXiv:2508.11800*, 2025.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Hui Chen, Antoine Didisheim, and Luciano A Somoza. Out of the black box: Uncertainty quantification for llms via conditional probabilities. Technical report, National Bureau of Economic Research, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yuyang Dai and Yuxia Wang. Rescaling confidence: What scale design reveals about llm metacognition. *arXiv preprint arXiv:2603.09309*, 2026.

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. In *International conference on learning representations*, volume 2026, pp. 59456–59493, 2026.

Cheng Ge, Caitlyn Heqi Yin, Hao Liang, and Jiawei Zhang. Why grpo needs normalization: A local-curvature perspective on adaptive gradients. *arXiv preprint arXiv:2601.23135*, 2026.

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

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on information systems*, 43(2):1–55, 2025.

Muhammed Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, and Samet Oymak. Learning to correct: Calibrated reinforcement learning for multi-attempt chain-of-thought. *arXiv preprint arXiv:2604.17912*, 2026.

Declan Jackson, William Keating, George Cameron, and Micah Hill-Smith. Aa-omniscience: Evaluating cross-domain knowledge reliability in large language models. *arXiv preprint arXiv:2511.13029*, 2025.

Abha Jha, Akanksha Mahajan, Ashwath Vaithinathan Aravindan, Praveen Saravanan, Sai Sailaja Policharla, and Sonal Chaturbhuj Gehlot. Rewarding intellectual humility learning when not to answer in large language models. *arXiv preprint arXiv:2601.20126*, 2026.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5684–5696, 2020.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Mohamad Amin Mohamadi, Tianhao Wang, and Zhiyuan Li. Honesty over accuracy: Trustworthy language models through reinforced hesitation. *arXiv preprint arXiv:2511.11500*, 2025.

Muyu Pan, Shu Zhao, Nan Zhang, Philip Shin, Varun Parekh, Vijaykrishnan Narayanan, and Rui Zhang. Tiar: Trajectory-informed advantage reweighting for llm abstention learning. *arXiv preprint arXiv:2605.25850*, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

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*, pp. 5433–5442, 2023.

Benjamin Turtel, Danny Franklin, Kris Skotheim, Luke Hewitt, and Philipp Schoenegger. Outcome-based reinforcement learning to predict the future. *arXiv preprint arXiv:2505.17989*, 2025.

Qingni Wang, Yue Fan, and Xin Wang. Safer: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations*, volume 2026, pp. 53723–53747, 2026.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

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

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, et al. Truthrl: Incentivizing truthful llms via reinforcement learning. *arXiv preprint arXiv:2509.25760*, 2025.

Jiayun Wu, Jiashuo Liu, Zhiyuan Zeng, Tianyang Zhan, Tianle Cai, and Wenhao Huang. Mitigating llm hallucination via behaviorally calibrated reinforcement learning. *arXiv preprint arXiv:2512.19920*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

APPENDIX

A Related Work **13**

A.1 Reinforcement Learning with Verifiable Rewards 13

A.2 Hallucination Mitigation and Behavioral Calibration 14

B Detailed Experimental Setup **14**

B.1 Implementation Details 14

B.2 Benchmarks 14

B.3 Baselines 15

C Additional Experimental Results **16**

C.1 Choice of Stability Coefficient 16

C.2 Standard Deviation 18

C.3 Fine-Grained Confidence Control 18

C.4 Increasing the Rollout Group Size 19

C.5 Scaling RLEC to Longer Contexts 20

C.6 Implicit Confidence Profile 20

C.7 Risk-Coverage Trade-off 21

D Relation Between Score and the RLCR Reward **22**

E Limitations and Future Work **22**

A RELATED WORK

A.1 REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Reinforcement learning with verifiable rewards (RLVR) Shao et al. (2024); Guo et al. (2025); Yu et al. (2026); Liu et al. (2025); Chen et al. (2025) has become a common approach for improving LLM reasoning. A widely used algorithm is Group Relative Policy Optimization (GRPO) Shao et al. (2024), which estimates advantages from multiple responses to the same prompt by subtracting the group mean and dividing by the group standard deviation. Prior work has examined the benefits and drawbacks of this normalization. Dr. GRPO Liu et al. (2025) removes it to avoid weighting questions according to their difficulty. Turtel et al. (2025) remove standard deviation normalization to preserve the effect of large prediction errors, while Bereket & Leskovec (2025) show that removing it reduces biased advantages and overconfidence. In contrast, Ge et al. (2026) view normalization as adaptive gradient scaling that can improve convergence and stability, while CAL-GRPO (Ildiz et al., 2026) also reports that retaining normalization improves sample efficiency across its GRPO variants.

These studies mainly examine how standard deviation normalization affects optimization and model performance. Unlike all of them, RLEC uses the reward scale to encode the confidence target p. We show that for groups with two distinct reward values and fixed proportions, normalization removes the effect of the reward scale induced by p on the advantages. Such groups are common during training, helping explain the failure to distinguish confidence targets in prior work Wu et al. (2025). RLEC retains this signal through mean centering without standard deviation normalization.

A.2 Hallucination Mitigation and Behavioral Calibration

Hallucination Ji et al. (2023); Huang et al. (2025) remains a major source of errors in LLMs. In *Why Language Models Hallucinate*, Kalai et al. (2025) argue that common training objectives encourage models to guess rather than admit uncertainty. This has motivated RL methods that reward correct answers, penalize errors, and allow abstention Wei et al. (2025); Mohamadi et al. (2025); Jha et al. (2026); Pan et al. (2026). However, these methods encode a fixed risk preference and cannot adapt their decision to answer or abstain to different user requirements at inference time. Wu et al. (2025) propose Explicit Risk Thresholding (ERT), which extends this setting to confidence targets specified in the prompt, but report that it fails to learn distinct behaviors across targets.

Another line of work trains or prompts models to express confidence Lin et al. (2022); Tian et al. (2023).For example, Damani et al. (2026) propose Reinforcement Learning with Calibration Rewards (RLCR), which trains models to produce an answer together with a calibrated confidence score. However, these confidence estimates typically need to be converted into decisions by an external rule or system, and some approaches further require a separate calibrator or calibration set Kamath et al. (2020); Wang et al. (2026). In contrast, RLEC enables models to learn how to directly adapt their answer or abstention decisions to explicit confidence targets, without an external rule for converting confidence estimates into decisions.

B Detailed Experimental Setup

B.1 Implementation Details

Training and Evaluation. We use `verl` for RL training Sheng et al. (2025), with the hyperparameters for all RL methods summarized in Table 2. We do not use Clip-Higher, entropy regularization, or a KL penalty. For RLCR, we determine abstention by applying the requested confidence target as an external threshold to its reported confidence. For all other methods, abstentions are identified by matching the designated abstention string. For $p = 0$, we use a binary reward and evaluation criterion: only a correct answer is treated as correct, whereas incorrect answers and abstentions are both treated as incorrect. For AOC, accuracy is set to zero when no answers are attempted, balancing calibration against utility. Due to limited computational resources, we do not conduct multiple training runs for each setting. Instead, we report Avg@16 for AIME and Avg@2 for all other benchmarks to reduce evaluation variance.

Table 2: **Hyperparameters.**

Hyperparameter	Value
Training	
`max_response_length`	4096
`train_batch_size`	256
`ppo_mini_batch_size`	128
`optim.lr`	1×10^{-6}
`rollout.temperature`	1.0
`rollout.n`	8
Evaluation	
`max_tokens`	4096
`temperature`	1.0
`top_p`	0.95

Prompts. The prompt templates used by RLEC and the baselines are shown in Figure 7. For other abstention methods, including ERT and TruthRL, we use the RLEC prompt during both training and evaluation. For RLVR, we use its standard prompt during training and at $p = 0$ to represent the conventional RL setting and compare answer ability. At $p > 0$, we use the RLEC prompt to demonstrate that standard RLVR training alone does not enable target-conditioned abstention. For SimpleQA-Verified and AA-Omniscience, we additionally use the grading prompt shown in Figure 8 and use Qwen3.6-27B as the judge. Since abstention is identified separately, this grader only determines whether the predicted answer is semantically equivalent to the reference answer.

B.2 Benchmarks

We use the standard versions of the evaluation benchmarks except for the following two settings.

HotpotQA. We use the modified version Damani et al. (2026) of HotpotQA Yang et al. (2018). Each example contains eight paragraphs, while zero, one, or both of the two relevant paragraphs are removed to test whether the model abstains when the available evidence is incomplete.

RLEC Prompt

Work through this problem step by step, then honestly assess how confident you are that your solution is correct. Only give a final answer if you are at least {p} confident that it is correct; otherwise do not guess and answer with \boxed{I am not sure.} instead. Put your final answer inside \boxed{}. If you are not confident enough, put \boxed{I am not sure.} instead.

RLCR Prompt

Work through this problem step by step, then honestly assess how confident you are that your solution is correct. Put your final answer inside \boxed{}, then output your confidence that the answer is correct as a number between 0 and 1 (inclusive) enclosed within <confidence> </confidence> tags.

RLVR Prompt

Work through this problem step by step. Put your final answer inside \boxed{}.

Figure 7: **Prompt templates used by RLEC and the baselines.**

Correctness-Grading System Prompt

You are a grading assistant. Decide whether the predicted answer is semantically equivalent to the reference gold answer for the given question. Ignore formatting differences, units notation, LaTeX wrappers, and trivial rephrasing. Reply with a single token: YES or NO.

Figure 8: **System prompt used to grade SimpleQA-Verified and AA-Omniscience.**

MMLU-Pro. We evaluate the four science subsets in MMLU-Pro Wang et al. (2024): biology, chemistry, physics, and health. Since the full benchmark contains substantially more examples than the other benchmarks, this selection prevents MMLU-Pro from dominating the aggregate results.

B.3 Baselines

RLVR. RLVR uses a binary correctness reward and does not explicitly reward abstention:

$$R_{\text{RLVR}} = \begin{cases} 1, & \text{correct,} \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

In the main experiments, we remove its group standard deviation normalization to align its advantage with RLEC and enable a closer comparison. Appendix C.2 examines the effect of this choice.

TruthRL. TruthRL Wei et al. (2025) uses a fixed ternary reward that penalizes incorrect answers while assigning zero reward to abstentions:

$$R_{\text{TruthRL}} = \begin{cases} 1, & \text{correct,} \\ -1, & \text{incorrect,} \\ 0, & \text{abstention.} \end{cases} \tag{24}$$

Our framework is compatible with TruthRL by setting $p = 0.5$, $b = 0$, and retaining group standard-deviation normalization.

Table 3: Forced answer accuracy, average overconfidence (AOC), and average score for RLVR and different stability coefficient and scheduling choices. The best Micro and Macro results among variants that distinguish confidence targets are shown in bold; RLVR and $b = 1$ are excluded from this comparison because they cannot distinguish among targets.

		Mathematical Reasoning			Knowledge & Hallucination			Scientific Reasoning		Avg.	
Metric	Method	GSM8K	MATH	AIME	Hotpot	OmniSci	SimpleQA	MMLU-Pro	GPQA	Micro	Macro
Acc. ↑	RLVR	93.33	69.70	20.83	69.50	13.75	5.20	68.47	35.91	58.18	47.09
	Ours	92.15	70.50	19.79	67.60	14.17	4.75	67.37	36.24	**57.40**	**46.57**
	$b = 0$	91.09	69.30	16.25	66.00	14.25	4.65	67.32	36.35	56.95	45.65
	$b = 0.5$	92.46	70.90	15.62	67.70	14.75	4.25	67.40	33.00	57.18	45.76
	$b = 1$	91.81	70.40	19.38	67.80	14.08	5.30	69.49	34.56	58.29	46.60
	LR Schedule	91.77	69.80	16.46	64.00	14.08	4.10	67.07	34.68	56.69	45.25
	Sine Schedule	93.03	69.60	20.00	67.80	13.75	4.45	65.77	35.12	56.63	46.19
AOC ↓	RLVR	0.000	0.027	0.258	0.033	0.292	0.389	0.032	0.175	0.049	0.151
	Ours	0.000	0.000	0.090	0.010	0.259	0.381	0.015	0.157	0.013	0.114
	$b = 0$	0.000	0.002	0.037	0.008	0.178	0.388	0.016	0.144	0.010	0.096
	$b = 0.5$	0.000	0.001	0.018	0.000	0.207	0.370	0.015	0.141	**0.008**	**0.094**
	$b = 1$	0.000	0.010	0.020	0.031	0.251	0.384	0.025	0.154	0.029	0.109
	LR Schedule	0.000	0.004	0.013	0.017	0.227	0.382	0.013	0.144	0.011	0.100
	Sine Schedule	0.000	0.003	0.009	0.016	0.293	0.399	0.015	0.148	0.022	0.111
Score ↑	RLVR	0.480	0.261	-0.162	0.237	-0.173	-0.213	0.237	-0.090	0.167	0.072
	Ours	0.477	0.333	0.056	0.264	-0.048	-0.107	0.262	-0.016	0.214	0.153
	$b = 0$	0.474	0.304	0.054	0.267	-0.010	-0.063	0.264	-0.001	**0.222**	0.161
	$b = 0.5$	0.477	0.334	0.048	0.276	-0.025	-0.061	0.258	0.015	**0.222**	**0.165**
	$b = 1$	0.474	0.305	0.073	0.241	-0.087	-0.129	0.247	-0.034	0.198	0.136
	LR Schedule	0.473	0.316	0.063	0.241	-0.011	-0.066	0.261	0.001	0.219	0.160
	Sine Schedule	0.478	0.330	0.074	0.257	-0.101	-0.180	0.264	-0.007	0.204	0.139

ERT. ERT Wu et al. (2025) samples a confidence target p and applies its corresponding error cost during standard GRPO training:

$$R_{\mathrm{ERT}}(p) = \begin{cases} 1, & \text{correct}, \\ -\dfrac{p}{1-p}, & \text{incorrect}, \\ 0, & \text{abstention}. \end{cases} \tag{25}$$

Our framework is compatible with ERT by setting $b = 0$ and retaining group standard-deviation normalization.

RLCR. RLCR Damani et al. (2026) trains the model to produce an answer together with a confidence score $c \in [0, 1]$. Let $y = 1$ if the answer is correct and $y = 0$ otherwise; its reward combines correctness with the Brier score:

$$R_{\mathrm{RLCR}}(y, c) = y - (c - y)^2. \tag{26}$$

For RLCR, we remove group standard deviation normalization to obtain better calibration Turtel et al. (2025); Bereket & Leskovec (2025). At evaluation time, we use an external decision system that retains the answer when $c \geq p$ and replaces it with an abstention otherwise.

C Additional Experimental Results

C.1 Choice of Stability Coefficient

We examine the choice of b within our proposed reward family, considering fixed values as well as linear and sinusoidal schedules. As shown in Equation 19, these choices preserve the same theoretical decision rule but induce different balances among signal strength, optimization stability, and answer ability. Table 3 reports their Forced Accuracy, AOC, and Score, reflecting answer ability, target compliance, and utility, respectively. We additionally include LR Schedule as a possible variant, which applies a global cosine learning-rate decay.

Forced Accuracy. Among the configurations that distinguish confidence targets, Forced Accuracy remains broadly stable, with the linear schedule obtaining the highest aggregate Forced Accuracy.

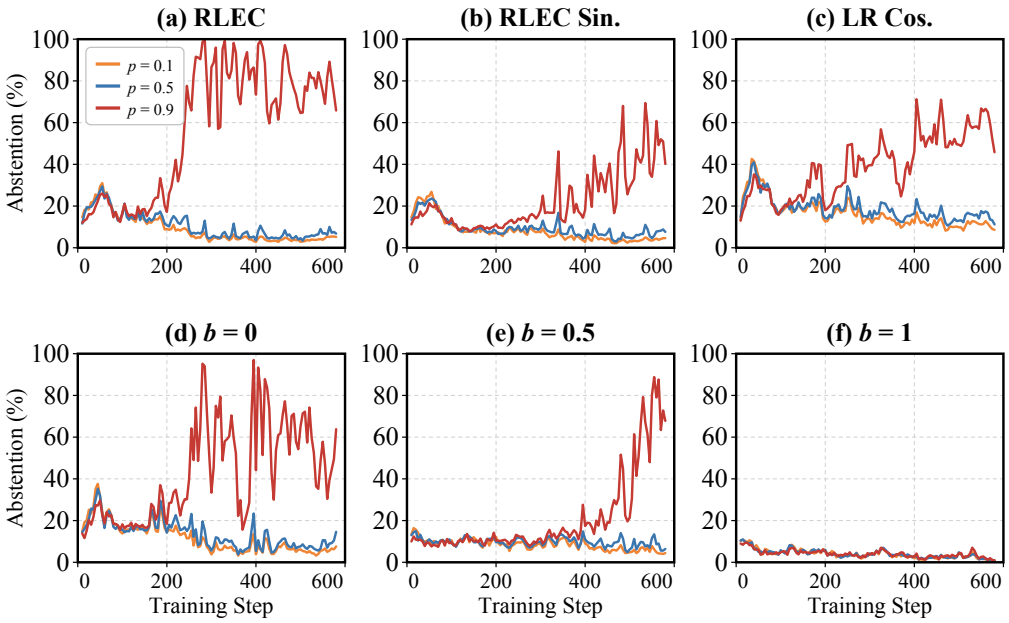


Figure 9: Training dynamics of the abstention rate on HotpotQA under different choices of the stability coefficient, scheduling strategy. RLEC uses the linear schedule of b, while RLEC Sin. uses a sinusoidal schedule. Curves correspond to target confidence levels $p \in \{0.1, 0.5, 0.9\}$.

The clearest differences appear on AIME, where fixed $b = 0$, fixed $b = 0.5$, and LR Schedule are all more than three points below the linear schedule. Because AIME is particularly difficult and benefits substantially from reinforcement learning, this result is consistent with training stability being especially important for preserving abilities acquired during this stage. Differences on most other benchmarks are smaller or less consistent, particularly on knowledge benchmarks such as SimpleQA-Verified and AA-Omniscience, where performance largely depends on knowledge acquired during pretraining. LR Schedule performs worse in aggregate Forced Accuracy, indicating that scheduling the relative reward scale across confidence targets is not equivalent to applying a global learning-rate decay.

Calibration Performance. All choices of b substantially improve calibration over RLVR, while the differences among these choices remain modest. In particular, $b = 0.5$ obtains the lowest aggregate AOC and the highest Macro Score, while retaining relatively strong Forced Accuracy.

Training Dynamics. The training dynamics in Figure 9, however, reveal an additional distinction that is not captured by the final aggregate results. Under the linear schedule, the behavior for $p = 0.9$ separates from the lower-confidence targets around the middle of training and remains distinctly ordered thereafter. Fixed $b = 0$ also distinguishes the targets relatively early, but exhibits larger fluctuations. With fixed $b = 0.5$, clear separation emerges only after approximately 400 steps despite its favorable final calibration and relatively strong Forced Accuracy, suggesting that it may not establish target conditioned behavior as reliably throughout optimization and may be more sensitive to the available training horizon. With $b = 1$, the curves remain nearly indistinguishable, consistent with the absence of a target-dependent reward gap. Overall, smaller values of b strengthen target-dependent learning but reduce stability, whereas larger values stabilize updates but weaken differentiation. The linear schedule balances these effects by establishing the distinctions early and stabilizing them later.

Final Choice. We treat preserving answer ability as a primary requirement for behavioral calibration, since calibration improvements are less desirable when accompanied by degradation in the model's underlying answer accuracy. Although the sinusoidal schedule achieves comparable Macro

Table 4: Forced answer accuracy for RLVR with and without group standard deviation normalization. The best result in each column is shown in bold.

	Mathematical Reasoning			Knowledge & Hallucination			Scientific Reasoning		Avg.	
Method	GSM8K	MATH	AIME	Hotpot	OmniSci	SimpleQA	MMLU-Pro	GPQA	Micro	Macro
RLVR w/o Std.	**93.33**	69.70	**20.83**	**69.50**	13.75	**5.20**	68.47	35.91	58.18	47.09
RLVR w/ Std.	92.57	**72.30**	19.17	68.30	**14.83**	5.15	**68.97**	**36.80**	**58.45**	**47.26**

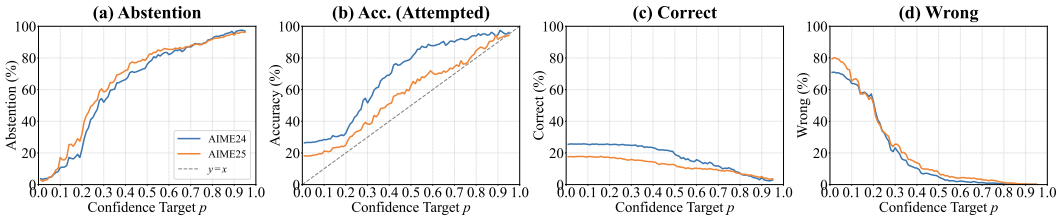


Figure 10: **Fine-grained confidence control on AIME24 and AIME25.**

results across all three metrics, it provides no consistent advantage over the linear schedule. We therefore adopt the linear schedule as our default configuration and use it to scale RLEC to the 30B model, because it achieves stronger aggregate Forced Accuracy, establishes earlier and more persistent target separation, and retains a simpler form than the sinusoidal alternative.

C.2 Standard Deviation

RLEC removes the group standard deviation from advantage estimation to preserve the reward signal. To enable a closer comparison, we also remove this normalization from RLVR in our main experiments. To further isolate its effect on RLVR's answer ability, we compare RLVR with centered advantages against standard RLVR under the same binary reward. As shown in Table 4, Their aggregate Forced Accuracy is nearly identical, and neither variant consistently outperforms the other across benchmarks. Thus, removing the group standard deviation provides no clear advantage or disadvantage in answer ability in our experiment setting.

C.3 Fine-Grained Confidence Control

We evaluate RLEC on AIME24 and AIME25 with confidence targets spaced at intervals of 0.01, extending up to $p = 0.95$. We choose these benchmarks because their difficulty provides an informative regime for studying fine-grained control: their relatively low accuracy avoids the saturation observed on easier benchmarks such as GSM8K, while a meaningful subset of problems remains within the model's reasoning capability. In contrast, errors on knowledge-intensive benchmarks such as SimpleQA more often reflect missing factual knowledge, leaving less room for confidence control to distinguish solvable from unsolvable examples. Each target is evaluated using Avg@1024. Figure 10 reports the abstention rate, attempted accuracy, and the proportions of all responses that are correct and wrong. Because training uses only targets up to $p = 0.90$, the interval $p \in (0.90, 0.95]$ additionally evaluates behavior beyond the training range.

Fine-Grained Control. As shown in Figure 10, the abstention rate rises steadily and attempted accuracy generally improves as p increases, demonstrating that RLEC responds to fine-grained changes in the target, including values such as $p = 0.01$ and $p = 0.02$. The curves are not strictly monotonic at every adjacent target, which we attribute to finite sample evaluation noise, but their overall trends remain stable. Attempted accuracy also generally meets or exceeds the target.

In contrast, although the RLCR prompt does not restrict the number of decimal places in its confidence output, as shown in Figure 7, all confidence values produced by the evaluated RLCR model lie on a one-decimal grid; we do not observe intermediate values such as 0.12. This observation is consistent with prior findings (Chen et al., 2026; Dai & Wang, 2026). Consequently, even if an external system applies arbitrary thresholds to these outputs, it cannot recover additional operating points

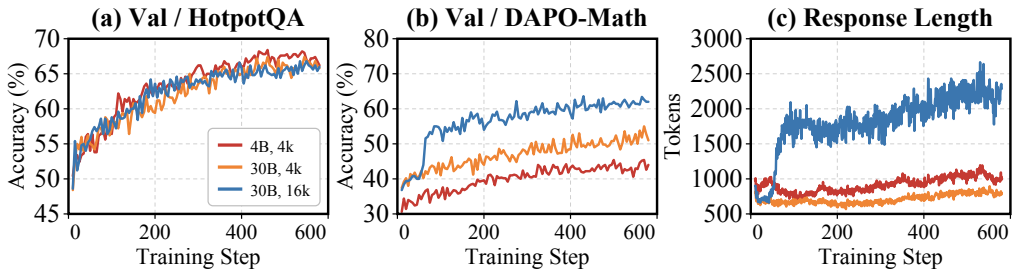


Figure 11: **Training dynamics under different model scales and context lengths.** We report validation accuracy on HotpotQA and DAPO-Math and the average response length during training.

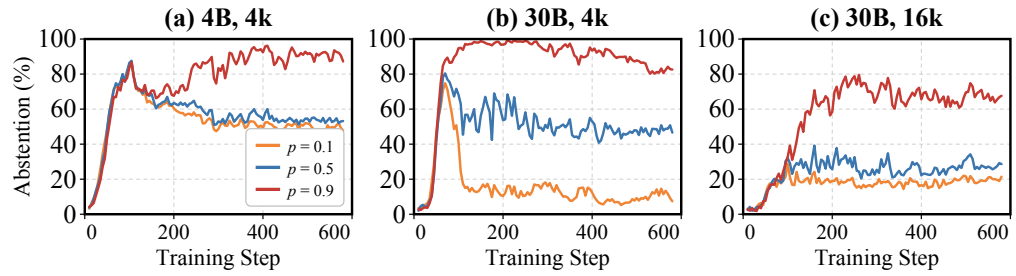


Figure 12: **Abstention dynamics on DAPO-Math for different model scales and context lengths.**

between adjacent confidence values. The evaluated RLCR model therefore provides substantially coarser control than directly conditioning RLEC on targets at a resolution of 0.01.

Selective Error Filtering. Panels (c) and (d) further show that the number of wrong answers decreases much faster than the number of correct answers as p increases. This effect is particularly clear at lower confidence targets, where the proportion of correct answers remains nearly unchanged while many wrong answers are filtered out. RLEC therefore initially improves attempted accuracy primarily by abstaining from likely errors rather than discarding correct answers, yielding a favorable trade-off between coverage and answer reliability.

Extrapolation Beyond the Training Range. RLEC is trained only with targets up to $p = 0.90$, whereas Figure 10 extends the evaluation to the unseen range $p \in (0.90, 0.95]$. Despite small local fluctuations and the reduced coverage at these high targets, all four metrics remain stable and continue their overall trends without an abrupt behavioral change. These results suggest that RLEC remains responsive even to confidence targets beyond the range observed during training.

C.4 Increasing the Rollout Group Size

We further examine whether increasing the rollout group size can help ERT preserve the confidence target signal. We retain the original reward and standard deviation normalization, so this setting differs from ERT only by increasing the number of responses per prompt from 8 to 16. This change approximately doubles the rollout generation cost.

As shown in Figure 13, increasing the rollout group size does not enable ERT to distinguish confidence targets. On HotpotQA, groups containing all three reward values rapidly disappear, and groups containing a single reward dominate later training. DAPO Math retains more groups containing all three rewards, but its abstention curves still largely overlap across different values of p. Thus, although a larger rollout group can increase the chance of observing all three outcomes, it does not resolve the loss of target information in groups containing only one or two reward values.

Table 5: Forced answer accuracy, average overconfidence (AOC), and average score for RLEC-30B with 4K and 16K contexts. The best Micro and Macro results for each metric are shown in bold.

		Mathematical Reasoning			Knowledge & Hallucination			Scientific Reasoning		Avg.	
Metric	Context	GSM8K	MATH	AIME	Hotpot	OmniSci	SimpleQA	MMLU-Pro	GPQA	Micro	Macro
Acc. ↑	4K	94.77	78.60	27.29	66.60	20.00	8.30	76.73	47.09	64.14	52.42
	16K	94.28	75.00	38.96	65.60	20.08	7.55	77.99	51.68	**64.86**	**53.89**
AOC ↓	4K	0.000	0.003	0.000	0.009	0.209	0.368	0.002	0.072	**0.001**	0.083
	16K	0.000	0.005	0.000	0.013	0.182	0.358	0.004	0.071	0.004	**0.079**
Score ↑	4K	0.504	0.353	0.091	0.258	-0.015	-0.132	0.336	0.112	0.261	0.188
	16K	0.502	0.329	0.157	0.251	0.003	-0.084	0.348	0.111	**0.272**	**0.202**

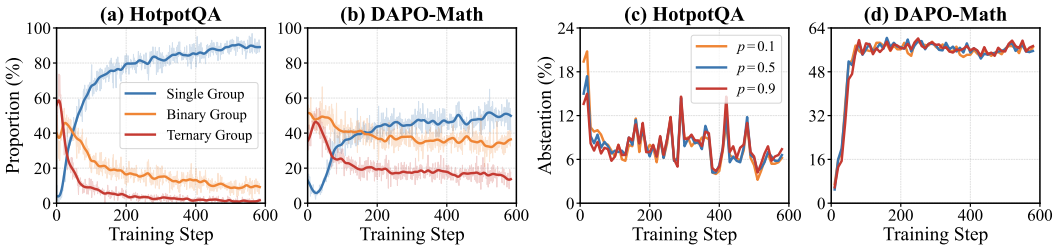


Figure 13: **Scaling ERT to a rollout group size of 16.** (a,b): group compositions during training. (c,d): abstention rates under different confidence targets.

These results indicate that increasing rollout size alone cannot replace preserving the target signal during advantage estimation.

C.5 Scaling RLEC to Longer Contexts

Due to computational constraints, our main experiments primarily use a 4K context window for both training and inference. However, difficult reasoning tasks such as AIME may benefit from a larger token budget. To examine whether RLEC can be extended to longer contexts, we increase both the training and inference context windows to 16K while keeping all other hyperparameters identical to those in Table 2.

Training Dynamics. As shown in Figure 11, increasing the context window from 4K to 16K leads to substantially longer responses from the 30B model. This additional budget produces a clear improvement in validation reward on DAPO-Math, whereas the HotpotQA trajectories remain largely unchanged. These results suggest that the longer context primarily benefits reasoning-intensive mathematical problems rather than uniformly improving all tasks. The stronger reasoning performance also changes the model's abstention behavior on DAPO-Math. As shown in Figure 12, the 16K model abstains less than its 4K counterpart at higher confidence targets, particularly at $p = 0.9$, while preserving the ordering across targets. This behavior is consistent with the stronger model answering more questions even under stringent confidence requirements.

Evaluation Results. Table 5 reports the complete evaluation results. As expected, the largest improvement appears on AIME, where Forced Accuracy increases from 27.29% with the 4K context to 38.96% with the 16K context. Results on the remaining benchmarks are mixed and generally show smaller differences. Consequently, aggregate Forced Accuracy improves slightly, largely driven by the gain on AIME, while AOC and Score remain broadly comparable. Overall, the longer context primarily benefits difficult mathematical reasoning without materially affecting behavior calibration.

C.6 Implicit Confidence Profile

RLEC directly accepts an explicit confidence target p and decides whether to answer, rather than outputting a confidence score. Nevertheless, to better understand its aggregate behavior and facilitate comparison with RLCR, we construct an operational approximation of its implicit confidence

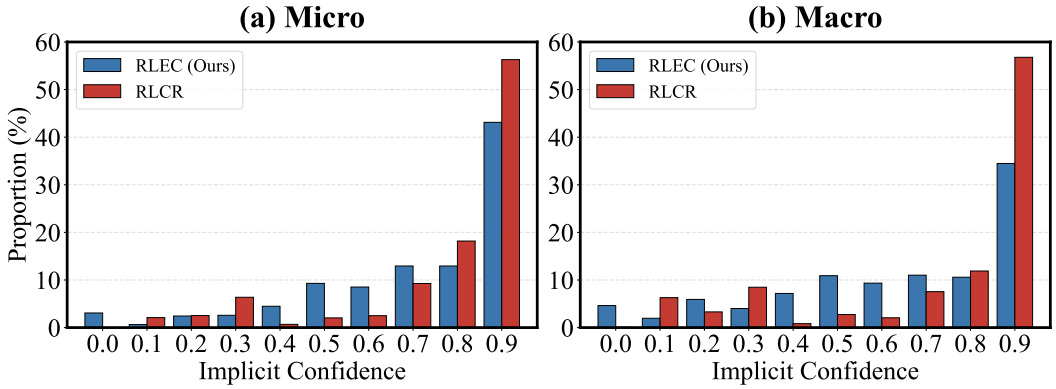


Figure 14: **Implicit confidence distributions.** The two methods exhibit broadly similar profiles, while RLEC is more dispersed, with less mass in the high interval and more in intermediate intervals.

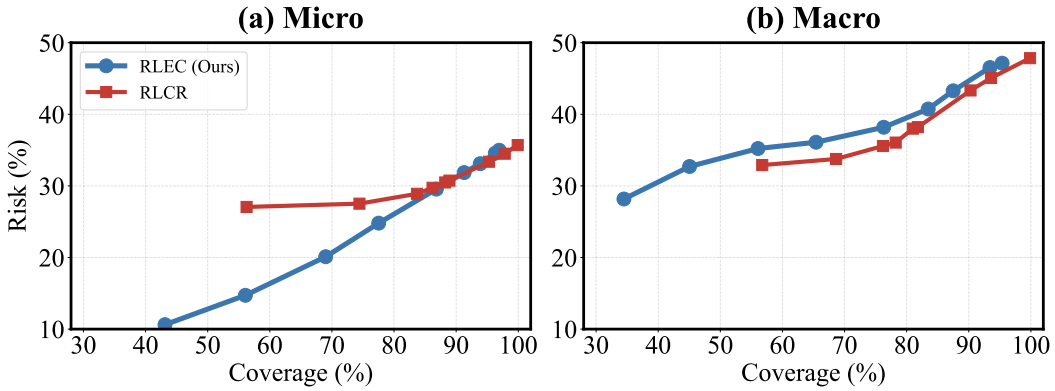


Figure 15: **Risk–coverage curves.** RLEC and RLCR exhibit broadly comparable trade-offs, while RLEC reaches lower-coverage operating points and supports more conservative behavior when high reliability is required, which is a potential practical advantage.

profile. We measure the abstention rate at $p \in \{0, 0.1, \ldots, 0.9\}$, treat it as an estimate of the cumulative mass below p, and use differences between neighboring targets as interval masses. This construction summarizes the aggregate behavior of the p-conditioned policies and does not assume a shared sample-level latent confidence variable. For RLCR, we directly use its reported confidence.

Figure 14 compares the resulting profiles on the 30B model. Under this construction, RLEC has a higher standard deviation (0.2335 versus 0.2126 for Micro and 0.2713 versus 0.2646 for Macro), with less mass in the highest interval and more in the intermediate intervals. These results suggest that RLEC supports a broader range of aggregate behavior across confidence targets.

C.7 Risk-Coverage Trade-off

Figure 15 plots risk, the error rate among answered questions, against coverage, the fraction of questions answered, on the 30B model. A lower curve is better, since it means fewer errors at the same coverage. Overall, the two methods perform similarly: RLEC is slightly better in the Micro setting, while RLCR is slightly better in the Macro setting. The main difference is in operating range. In both settings, RLEC extends to a lower coverage than RLCR, so it can abstain more and keep risk lower when the target calls for high reliability, whereas RLCR cannot reach the same conservative operating points. This is a potential practical advantage in applications that demand high reliability.

D Relation Between Score and the RLCR Reward

We show that the score used in our evaluation directly corresponds to the RLCR reward. Let $y \in \{0, 1\}$ indicate whether an answer is correct, and let $c \in [0, 1]$ be the confidence reported by RLCR. Given a confidence target p, the external decision system answers when $c \geq p$ and abstains otherwise. Its decision can therefore be written as

$$z_p = \mathbf{1}\{c \geq p\}. \tag{27}$$

The score at target p is

$$s_p = z_p(y - p) = \mathbf{1}\{c \geq p\}(y - p). \tag{28}$$

Integrating this score over all confidence targets gives

$$\int_0^1 s_p \, dp = \int_0^1 \mathbf{1}\{c \geq p\}(y - p) \, dp \tag{29}$$

$$= \int_0^c (y - p) \, dp \tag{30}$$

$$= cy - \frac{c^2}{2}. \tag{31}$$

RLCR uses the following correctness and calibration reward:

$$R_{\mathrm{RLCR}}(c, y) = y - (c - y)^2. \tag{32}$$

Because $y \in \{0, 1\}$, we have $y^2 = y$. The reward can therefore be rewritten as

$$R_{\mathrm{RLCR}}(c, y) = y - c^2 + 2cy - y^2 \tag{33}$$

$$= 2cy - c^2 \tag{34}$$

$$= 2\left(cy - \frac{c^2}{2}\right) \tag{35}$$

$$= 2\int_0^1 s_p \, dp. \tag{36}$$

Thus, when the confidence reported by RLCR is converted into an answer or abstention through thresholding, twice the score integrated over all confidence targets is exactly the RLCR reward.

In our experiments, we evaluate a finite set of confidence targets $\mathcal{P}$ and compute

$$\widehat{\mathrm{Score}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} s_p. \tag{37}$$

For uniformly spaced targets over $[0, 1)$, this is a Riemann approximation to the integral in Equation 31. The approximation approaches the exact relation in Equation 36 as the interval between evaluated targets decreases.

The same score also directly corresponds to the RLEC reward. For a given p,

$$z_p(y - p) = \begin{cases} 1 - p, & \text{correct answer,} \\ -p, & \text{incorrect answer,} \\ 0, & \text{abstention,} \end{cases} \tag{38}$$

which is exactly $R_p^{(1)}$, the RLEC reward with $b = 1$. Therefore, Score provides a common measure for comparing RLEC and RLCR under explicit confidence targets.

E Limitations and Future Work

Under idealized assumptions, our reward objective induces the decision rule $q(x) > p$ and consequently the target $\mathrm{Acc}(p) \geq p$. However, empirical compliance depends on how well the trained policy realizes this rule. First, limited model capabilities may restrict the ability to estimate correctness and adjust decisions across confidence targets. This is especially challenging on difficult

tasks, where even extensive abstention may not yield the requested accuracy. Second, generalization depends on the difficulty match between training and evaluation tasks. RLEC achieves better target compliance on reasoning benchmarks than on SimpleQA-Verified and AA-Omniscience. One plausible explanation is that DAPO-Math provides sufficiently challenging training examples to support generalization even to difficult benchmarks such as AIME. In contrast, HotpotQA is substantially easier than SimpleQA-Verified and AA-Omniscience, creating a larger difficulty gap that may limit generalization. Third, finite rollouts and stochastic optimization may not reach the optimal policy. Together, these factors can create a gap between the theoretical objective and empirical behavior.

Empirically, RLEC exhibits clear adaptation to confidence targets and strong aggregate target compliance. The remaining gaps are most visible with the 4B model and on particularly difficult benchmarks such as SimpleQA-Verified and AA-Omniscience. The stronger results at 30B suggest that increased model capacity can mitigate these gaps, although broader evaluation across model families and scales is still needed. Future work can explore stronger optimization objectives for reducing target violations, especially on difficult tasks and under distribution shifts.

Also, we primarily study verifiable question answering tasks to facilitate controlled training and reliable evaluation. This setting allows us to evaluate confidence target following precisely, but covers only part of real world model use. Future work should explore less structured tasks, such as open ended generation, where correctness is more difficult to define and verify.

We also acknowledge that RLEC achieves calibration broadly comparable to, rather than consistently better than, RLCR: their results are mixed at 4B, while RLEC performs better at 30B. Our primary contribution is therefore not a uniform improvement in aggregate calibration metrics, but a different form of control. Unlike RLCR, RLEC directly conditions its behavior on an explicit confidence target from user prompt, supports fine-grained control, and reaches more conservative operating points when high reliability is required. Future work may explore objectives that further improve absolute calibration while preserving these properties.