

Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty

**Mehul Damani, Isha Puri, Stewart Slocum, Idan
Shenfeld, Leshem Choshen, Yoon Kim, Jacob Andreas**

Rui Sun @ REFORM

Motivation

- “LLMs tend to become overconfident following RL training.”

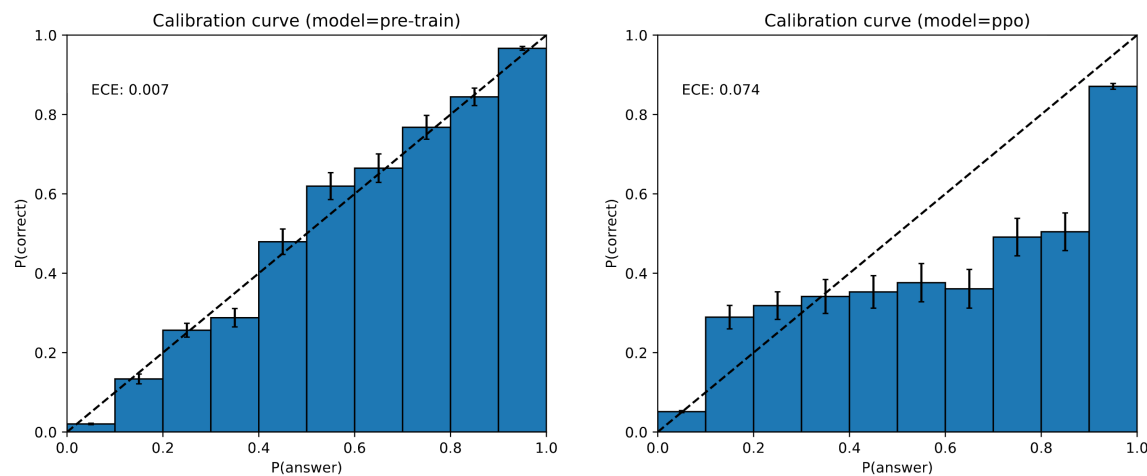
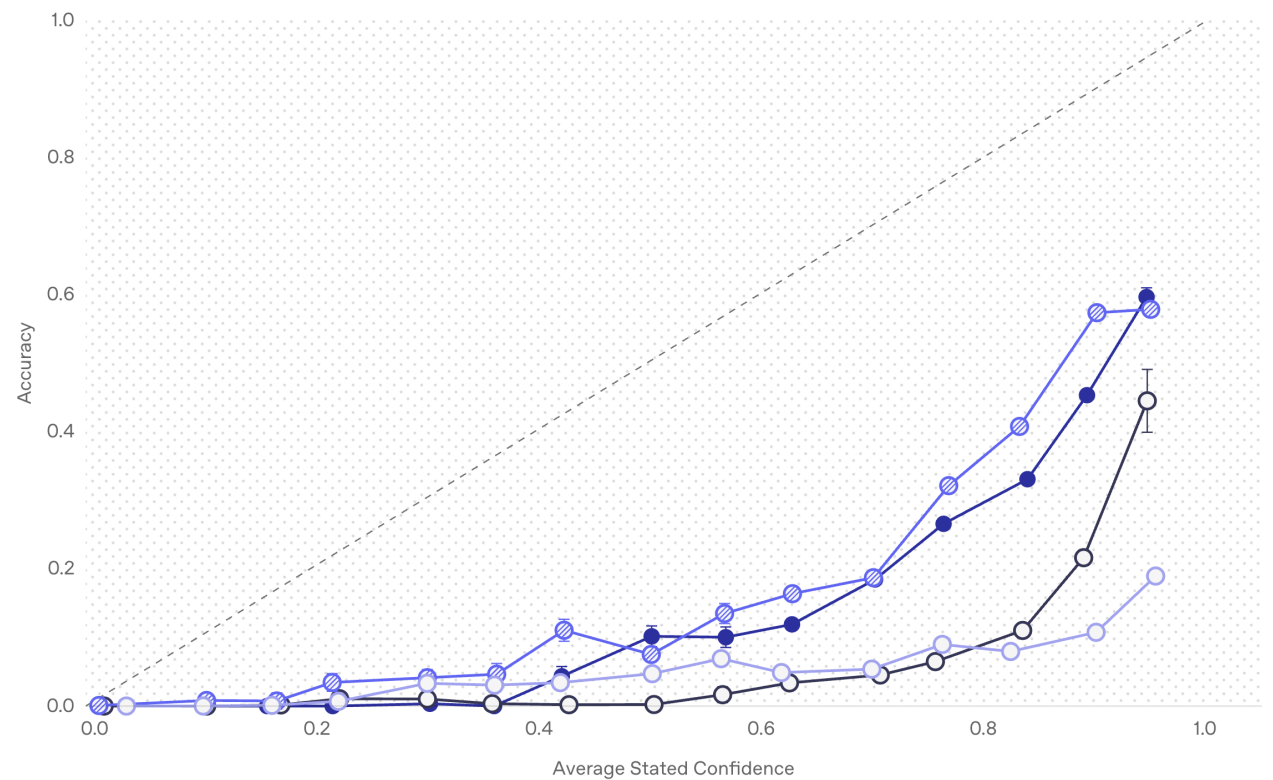


Figure 8. Left: Calibration plot of the pre-trained GPT-4 model on a subset of the MMLU dataset. On the x-axis are bins according to the model’s confidence (logprob) in each of the A/B/C/D choices for each question; on the y-axis is the accuracy within each bin. The dotted diagonal line represents perfect calibration. Right: Calibration plot of the post-trained GPT-4 model on the same subset of MMLU. The post-training hurts calibration significantly.

Calibration (Uniform)

● GPT-4o ○ GPT-4o-mini ● o1-preview ○ o1-mini ○ Perfect Calibration



Goals:

This paper aims to address these limitations by answering two questions:

- (1) Can reasoning models be optimized for both correctness and calibration?
- (2) Can the contents of reasoning chains themselves improve calibration?

Takeaway:

Modify the reward function to include the verbalized confidence.

$$\begin{aligned} R_{\text{RLCR}}(y, q, y^*) &= R_{\text{correctness}}(y, y^*) + R_{\text{Brier}}(y, q, y^*) \\ &= \mathbb{1}_{y=y^*} - (q - \mathbb{1}_{y=y^*})^2 . \end{aligned} \tag{1}$$

Method

optimizing a combined reward function

$$\arg \max_{\theta} \mathbb{E}_{(x, y^*) \sim D, y \sim \pi_{\theta}(\cdot | x)} R(y, y^*)$$

$$\begin{aligned} R_{\text{RLCR}}(y, q, y^*) &= R_{\text{correctness}}(y, y^*) + R_{\text{Brier}}(y, q, y^*) \\ &= \mathbb{1}_{y \equiv y^*} - (q - \mathbb{1}_{y \equiv y^*})^2. \end{aligned}$$

Theorem 1. Suppose, for any possible prediction y , that the success indicator $\mathbb{1}_{y \equiv y^*}$ is drawn from a distribution $\text{Bernoulli}(p_y)$.

Then R_{RLCR} in Eq. (8) satisfies two properties:

1. **Calibration incentive.** For any y , the expected reward $\mathbb{E}_{\mathbb{1}_{y \equiv y^*}} R_{\text{RLCR}}(y, q, y^*)$ is maximized when $q = p_y$.
2. **Correctness incentive.** Among all calibrated predictions (y, p_y) , expected reward is maximized by the prediction whose success probability p_y is greatest.² **monotonic in p_y**

That's why we use square loss instead of cross-entropy loss.

Experiments

1. **Base:** Qwen2.5-7B

2. **RLVR:** trained using $R_{\text{correctness}}$

3. **RLVR+BCE Classifier:** after **RLVR**, train another model $f_{\theta}(x, y)$

$$\mathcal{L}_{\text{BCE}}(\theta) = -\mathbb{E}_{(x, y, \mathbb{1}_{y=y^*})} [\mathbb{1}_{y=y^*} \log f_{\theta}(x, y) + (1 - \mathbb{1}_{y=y^*}) \log(1 - f_{\theta}(x, y))]$$

4. **RLVR+Brier Classifier:** after **RLVR**, train another model $f_{\theta}(x, y)$

$$\mathcal{L}_{\text{Brier}}(\theta) = \mathbb{E}_{(x, y, \mathbb{1}_{y=y^*})} \left[(f_{\theta}(x, y) - \mathbb{1}_{y=y^*})^2 \right]$$

5. **RLVR+Probe:** $f_{\theta}(x, y) = \log \sigma(\theta^{\top} \phi(x, y))$

6. **Answer Probability:** after **RLVR**, $\text{AnswerProb}(y) = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} P_{\theta}(y_t \mid y_{<t}, x)$

7. **RLCR (theirs)**

(a) Example Output

<think> The question asks for the song with which Lulu represented the UK in the 1969 Eurovision Song Contest. [...] I need to recall the specific song that Lulu performed for the UK in 1969. </think>

<answer> To Sir With Love </answer>

<analysis> There is a high level of uncertainty in this answer because while Lulu did represent the UK in the 1969 Eurovision Song Contest, the specific song she performed is not widely known or easily recalled. [...] but without more specific information, it is difficult to pinpoint the exact song.} </analysis>

<confidence> 0.3 </confidence>

Experiments

(a) Models Trained on HotpotQA								
Method	HotpotQA				O.O.D			
	Acc. (↑)	AUROC (↑)	Brier (↓)	ECE (↓)	Acc. (↑)	AUROC (↑)	Brier (↓)	ECE (↓)
Base	39.7%	0.54	0.53	0.53	53.3%	0.54	0.41	0.40
RLVR	63.0%	0.50	0.37	0.37	53.9%	0.50	0.46	0.46
RLVR + BCE Classifier	63.0%	0.66	0.22	0.07	53.9%	0.58	0.27	0.24
RLVR + Brier	63.0%	0.65	0.22	0.09	53.9%	0.60	0.32	0.33
RLVR + Probe	63.0%	0.55	0.24	0.10	53.9%	0.53	0.38	0.38
Answer Prob	63.0%	0.72	0.36	0.36	53.9%	0.60	0.42	0.42
RLCR (ours)	62.1%	0.69	0.21	0.03	56.2%	0.68	0.21	0.21

(b) Models Trained on Big-Math								
Method	Math				O.O.D Averaged			
	Acc. (↑)	AUROC (↑)	Brier (↓)	ECE (↓)	Acc. (↑)	AUROC (↑)	Brier (↓)	ECE (↓)
Base	56.1%	0.56	0.40	0.39	47.8%	0.53	0.46	0.45
RLVR	72.9%	0.47	0.28	0.26	52.5%	0.52	0.49	0.49
RLVR +Classifier	72.9%	0.78	0.15	0.10	52.5%	0.55	0.34	0.33
RLVR +Brier-Classifier	72.9%	0.78	0.15	0.10	52.5%	0.57	0.28	0.27
RLVR +Probe	72.9%	0.65	0.19	0.13	52.5%	0.53	0.33	0.30
Answer Prob	72.9%	0.52	0.26	0.26	52.5%	0.52	0.44	0.43
RLCR (ours)	72.7%	0.67	0.17	0.10	50.9%	0.60	0.28	0.25
SFT+RLCR (ours)	72.2%	0.78	0.14	0.08	43.8%	0.66	0.24	0.18

Table 1: Accuracy and calibration metrics for models trained on HotpotQA and Big-Math.

Experiments

training curves:

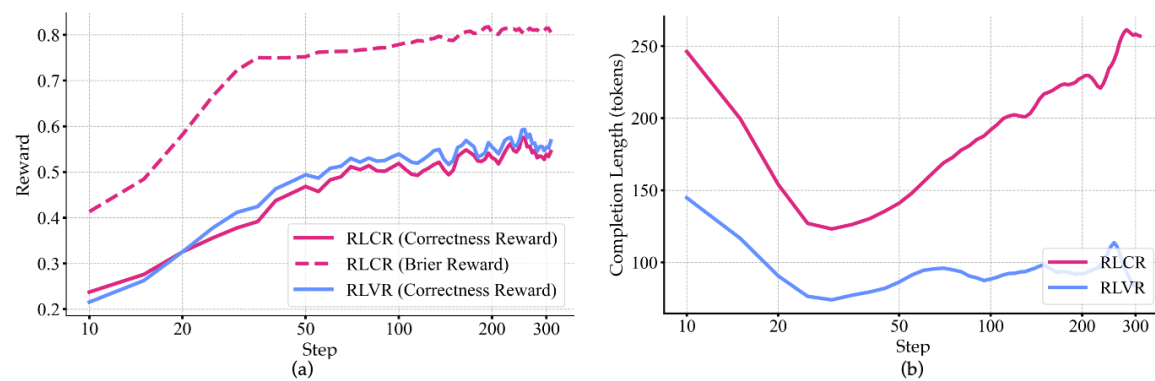


Figure 3: **(a) Reward curves for *RLCR* (ours) and *RLVR*.** Both correctness and calibration rewards improve under our method, demonstrating simultaneous gains in correctness and calibration. The Brier reward is shifted upward by 1 for clarity. **(b) Completion lengths during training.** The completion lengths of our method gradually increase during training as uncertainty reasoning improves.

test-time scaling:

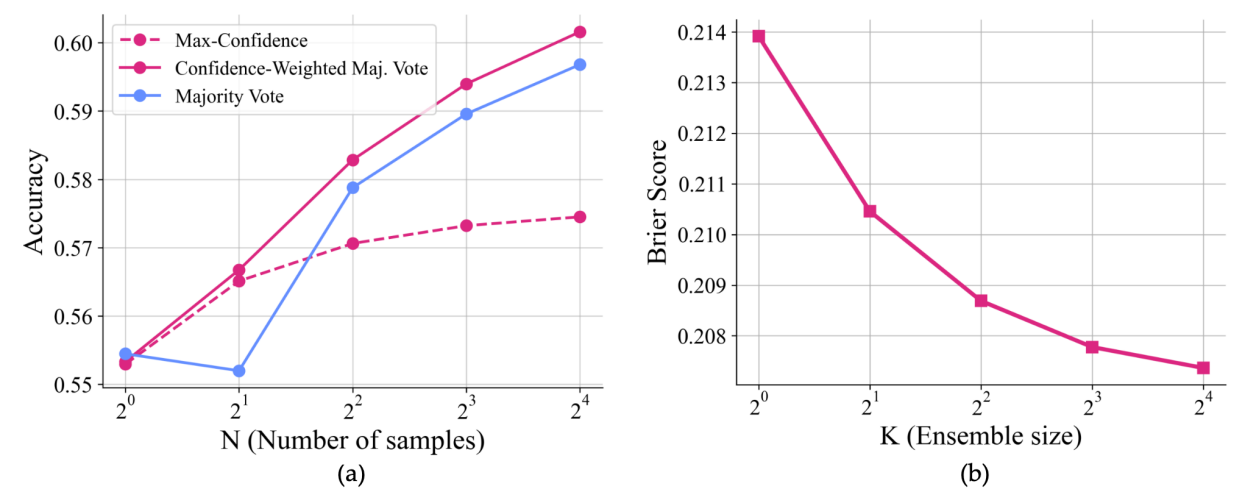


Figure 4: **Test-time scaling curves. (a) Accuracy vs Number of Samples (N).** Accuracy improves

Experiments

Does reasoning improve calibration?

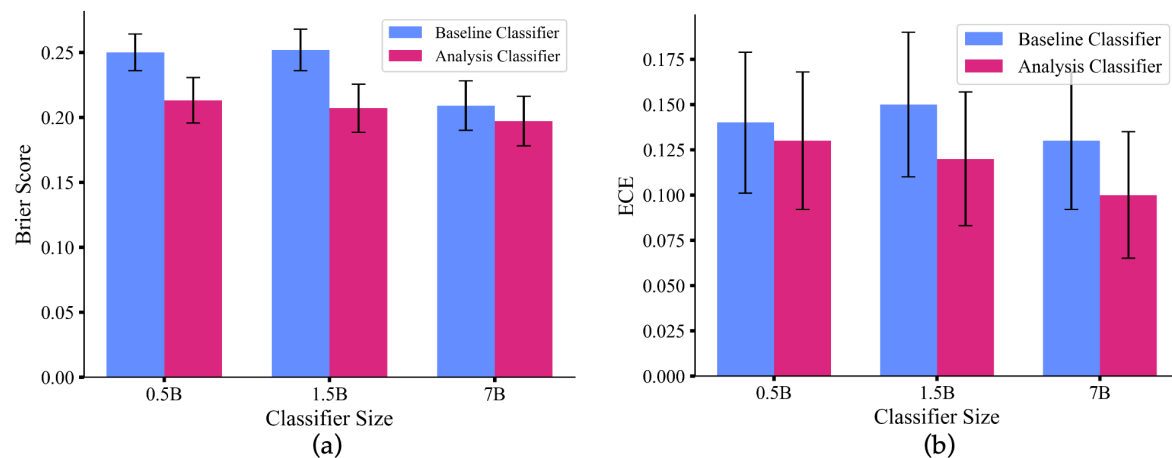


Figure 5: **Brier scores (a) and ECE (b) of baseline / analysis classifiers on HotPotQA-Modified across three model sizes.** Analysis classifiers outperform baselines at smaller sizes, suggesting that uncertainty CoT is essential for better calibration when capacity is limited.

Are confidences self-consistent?

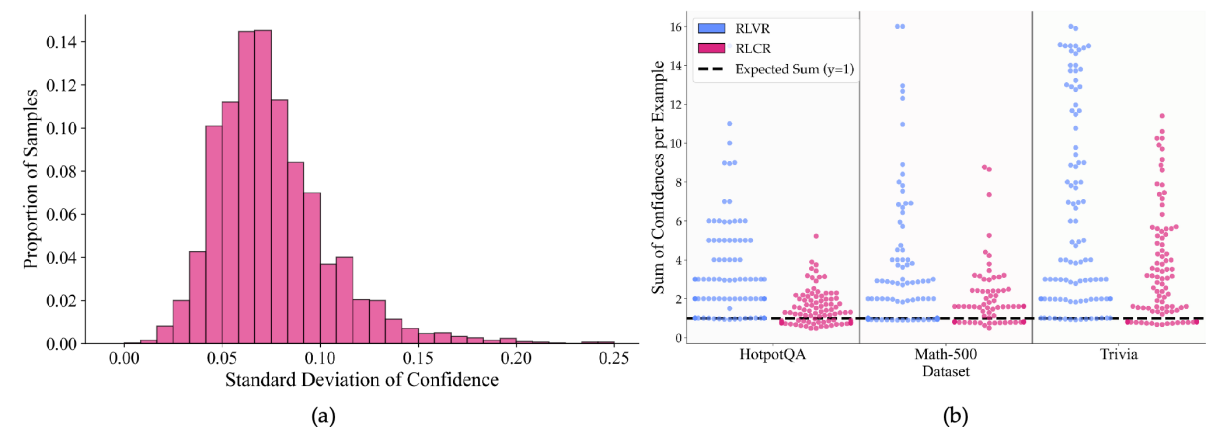


Figure 6: **(a): Distribution of standard deviation in confidence across multiple uncertainty reasoning chains for the same solution/answer.** Most samples exhibit low deviation, indicating that the model's confidence estimates are self-consistent. **(b) Swarm plot of confidence sums across 3 datasets.** RLCR consistently remains closer to the ideal sum of 1. Nonetheless, overconfidence remains, suggesting room for further improvement.

Experiments

SIMPLEQA EXAMPLE

Question

Who was awarded the Oceanography Society's Jerlov Award in 2018? (Ground Truth Answer: Annick Bricaud)

RLCR (ours)

<think> The question asks who was awarded the Oceanography Society's Jerlov Award in 2018. The Jerlov Award is given by the Oceanography Society, and it is named after the late Professor Gunnar Jerlov, a pioneer in the field of ocean optics. To find the winner of the 2018 award, I would need to look up the most recent information on the Oceanography Society's website or other reliable sources. </think>

<answer> Dr. David W. Schiedek </answer>

<analysis> The Oceanography Society's website or other reliable sources do not provide information on the winner of the 2018 Jerlov Award. It is possible that the information is not readily available or that the award was not given in 2018. Additionally, the name "Dr. David W. Schiedek" is not a well-known figure in the field of oceanography, which raises uncertainty about the accuracy of this answer. </analysis>

<confidence> 0.25 </confidence>

RLVR

<think> The question asks for the person who was awarded the Oceanography Society's Jerlov Award in 2018. The Jerlov Award is given by the Oceanography Society to recognize outstanding contributions to the field of oceanography. The recipient for 2018 was Dr. Laurence C. Smith. </think>

<answer> Laurence C. Smith </answer>

<confidence> 100. </confidence>

Qwen-7B-Base

<think> To find out who was awarded the Oceanography Society's Jerlov Award in 2018, I will search for the most recent information on the award winners from the Oceanography Society's official website or a reliable source that covers scientific awards. </think>

<answer> The Oceanography Society's Jerlov Award in 2018 was awarded to Dr. Lisa Levin. </answer>

<confidence> 0.95 </confidence>

Conclusion

6 CONCLUSION

We show that incorporating proper scoring rules into RL, via an objective we call RLCR, enables reasoning models to improve both accuracy and calibration. Our approach trains models to reason about and verbalize uncertainty, preserving task performance while significantly improving calibration in- and out-of-distribution. We demonstrate that reasoning about uncertainty improves calibration, and that our method improves the self-consistency of confidence, and improves with test-time scaling. However, there remains significant room for improvement—even after RLCR, out-of-domain calibration error is often high in an absolute sense, and models may still assign high confidence to multiple contradictory answers. Nevertheless, these results suggest a path toward reasoning systems that are not only accurate, but reliably reason about and communicate uncertainty.

Discussion

Calibration $\neq$ Hallucination

- People do not usually explicitly prompt for a confidence
- “IDK”, abstain when uncertain [KNVZ25]
- Hallucinating arbitrary facts $\neq$ hallucinating reasoning?
- “sub-claim” confidence [MH24,CGC24]

With probability 63% everything kept below is true

The shingles vaccine is typically recommended for adults aged 50 and older. The vaccine is given in two doses, with the second dose administered 2 to 6 months after the first dose. **It is currently recommended that individuals receive the shingles vaccine once in their lifetime.** However, it is always best to consult with a healthcare provider for personalized recommendations.

Discussion

Why model Hallucinate?

Table 2: Summary of evaluation benchmarks analyzed in this work and their treatment of abstentions. “Binary grading” indicates that the primary metric is a strict correct/incorrect accuracy; “IDK credit” denotes whether abstentions can earn any credit.

Benchmark	Scoring method	Binary grading	IDK credit
GPQA	Multiple-choice accuracy	Yes	None
MMLU-Pro	Multiple-choice accuracy	Yes	None
IFEval	Programmatic instruction verification	Yes ^a	None
Omni-MATH	Equivalence grading*	Yes	None
WildBench	LM-graded rubric*	No	Partial ^b
BBH	Multiple-choice / exact-match	Yes	None
MATH (L5 split)	Equivalence grading*	Yes	None
MuSR	Multiple-choice accuracy	Yes	None
SWE-bench	Patch passes unit tests	Yes	None
HLE	Multiple-choice / equivalence grading*	Yes	None

* Grading is performed using language models, hence incorrect *bluffs* may occasionally be scored as correct.

^a IFEval aggregates several binary rubric sub-scores into a composite score.

^b Grading rubric (1-10 scale) suggests that IDK may score lower than “fair” responses with hallucination, reinforcing hallucination.