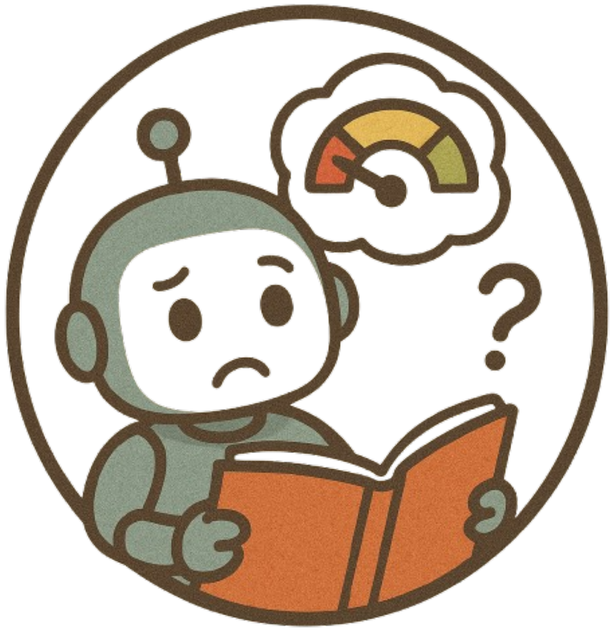
Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty

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

COLM SCALR Workshop — October 10th, 2025



*Equal Contribution

Standard Approaches to Reasoning Training

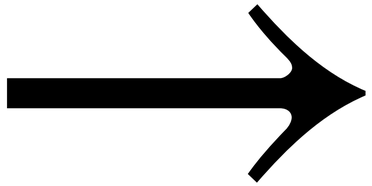
- **“Reasoning models” are trained via RL to “think out loud” before answering questions — these models do very well on math and programming**

- **Standard approach to reasoning training:**

Reinforcement Learning w/ Verifiable Rewards

Standard Approaches to Reasoning Training

Reinforcement
Learning
Verifiable
Rewards



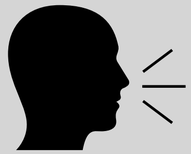
Binary Correctness Reward:

$$R_{correctness} = \begin{cases} 1 & \text{if answer is correct} \\ 0 & \text{if answer is wrong} \end{cases}$$

⚙️ ***awards correctness*** **- equivalent rewards are given whether models are confident or just guessing.**

Standard Approaches to Reasoning Training

Reinforcement
Learning
Verifiable
Rewards



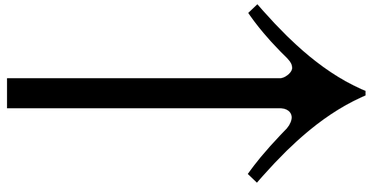
Hermione signed up for at least 12 classes this term. After dropping 5 and enrolling in 8 more, how many classes is she in?

“Hermione signed up for more than 12 classes. She dropped 5 of them and then added 8 more. So, the total number of classes she would have is (12 + 8) – 5.”

Question is ambiguous! “At least” 12 classes might mean 12, or it might mean 40. Model should not be confident!

Standard Approaches to Reasoning Training

Reinforcement
Learning
Verifiable
Rewards



Binary Correctness Reward:

$$R_{correctness} = \begin{cases} 1 & \text{if answer is correct} \\ 0 & \text{if answer is wrong} \end{cases}$$

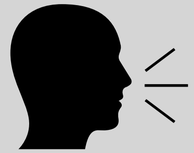
⚙️ ***awards only correctness*** **- equivalent rewards are given whether models are confident or just guessing.**

🚨 **This incentivizes guessing, undermining usefulness and trustworthiness**

- highly problematic in high stakes settings

Standard Approaches to Reasoning Training

Reinforcement
Learning
Verifiable
Rewards



Should I take aspirin before surgery?

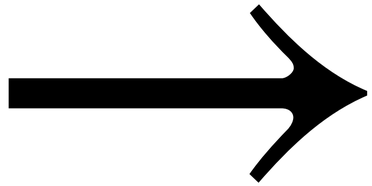
“Yes, taking aspirin before surgery can help prevent blood clots.”

While this might mostly be true, it is also potentially dangerous - it depends on the type of surgery, patient’s condition and medications. Model answer should not be confident!

How can we better calibrate our models without compromising on the strengths of RLVR?

RLCR

Reinforcement
Learning w/
Verifiable
Rewards



Reinforcement
Learning w/
Calibrated
Rewards

RLVR

RLCR

Question

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

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>

Question

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

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>

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>

Reason about uncertainty

<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>

Verbalize confidence

<confidence> 0.25 </confidence>

RLCR Reward

💡rewards *BOTH* *correctness* *AND* *calibration*

$$R_{RLVR} = \text{Correctness}$$

↓

$$R_{RLCR} = \text{Correctness} - (\text{Confidence} - \text{Correctness})^2$$

BRIER SCORE

Still incentivizes correctness!

incentivizes *verbalized confidence to be close to actual correctness*

RLCR - Overview

💡**Models reason about both correctness and uncertainty, and output both an answer and a verbalized confidence.**

💡***Awards BOTH correctness AND calibrated confidence***

💡***We want to both get the right answer BUT ALSO know how confident we are about it***

Sample Think, Answer, Analysis, and Confidence Tags of a Target Generation

<think> The question asks for the song with which Lulu represented the UK in the 1969 Eurovision Song Contest. Lulu is a well-known British singer, and the Eurovision Song Contest is an annual competition where countries submit songs to be performed and judged. I need to recall the specific song that Lulu performed for the UK in 1969. **</think>**

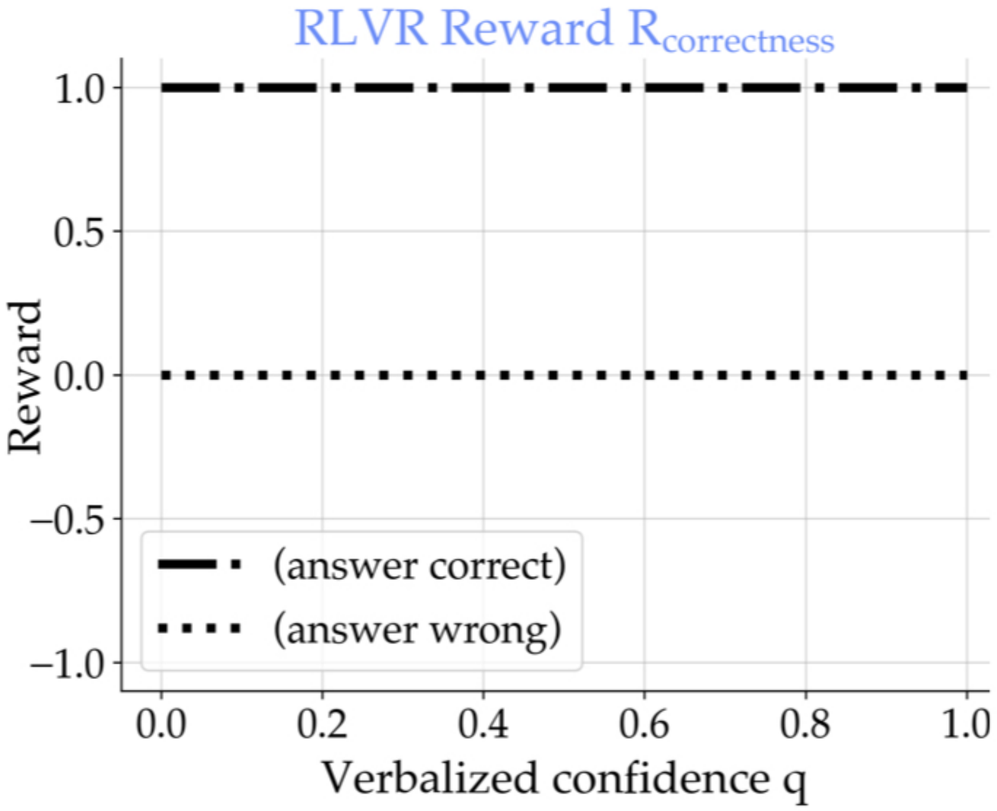
<answer> To Sir With Love **</answer>**

<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. The song "To Sir With Love" is a popular song by Lulu, but it is not commonly associated with her Eurovision performance. There might be other songs that are more closely linked to her Eurovision appearance, but without more specific information, it is difficult to pinpoint the exact song.} **</analysis>**

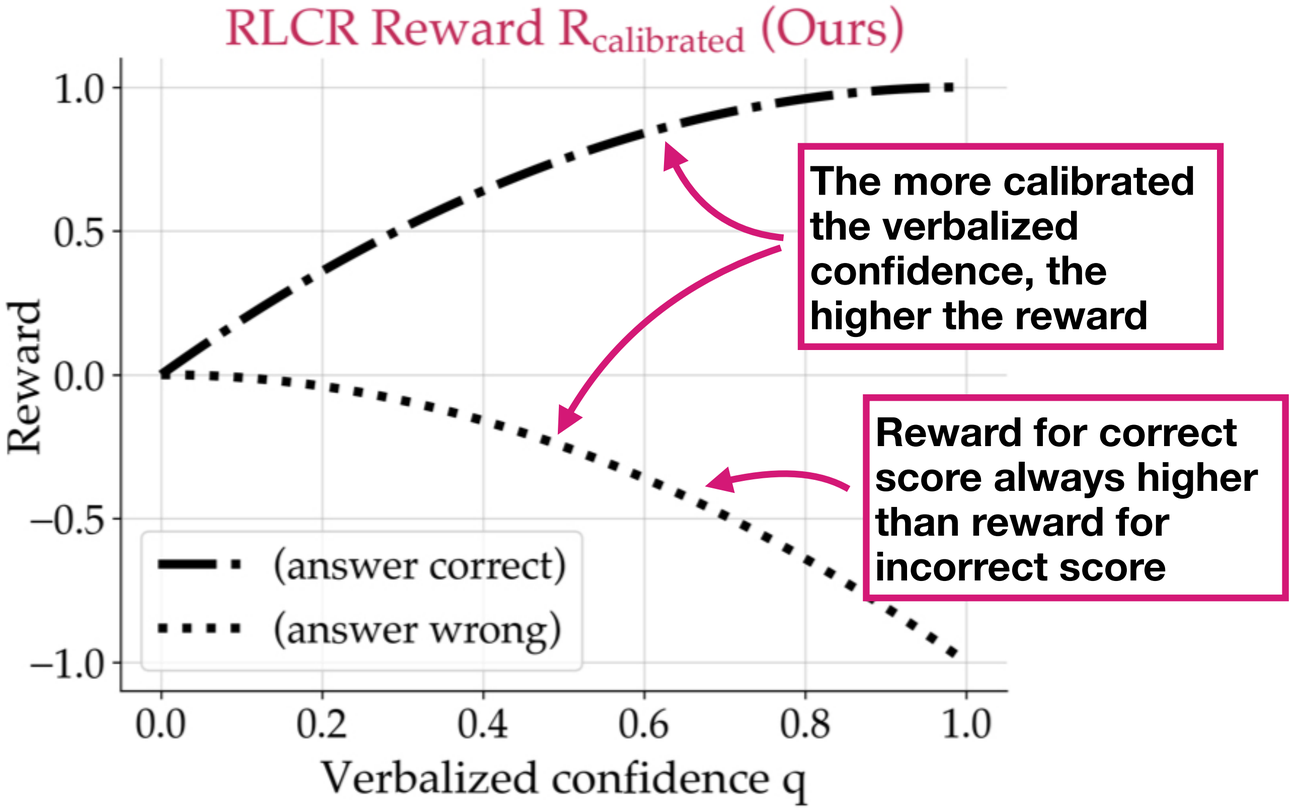
<confidence> 0.3 **</confidence>**

RLCR CoT

Reward Comparison



R_{RLVR} = correctness



R_{RLCR} (Ours) = correctness − (confidence − correctness)^2

But is this too good to be true? Is there any tradeoff when you optimize for both accuracy and calibration?

Theorem

RLCR provably incentivizes models to:

- **Report honest confidences *(calibration incentive)***
- **Output answers that maximize accuracy *(correctness incentive)***

Theorem 1. *Suppose, for any prediction y and verbalized confidence q, that the success indicator $\mathbb{1}_{y \equiv y^*}$ is distributed as Bernoulli(p_y). Then RLCR reward satisfies two properties:*

1. ***Calibration incentive.*** *For* any *y, the expected reward $\mathbb{E}_{\mathbb{1}_{y \equiv y^*}} R_{RLCR}(y, q, y^*)$ 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.*

Experimental Setup

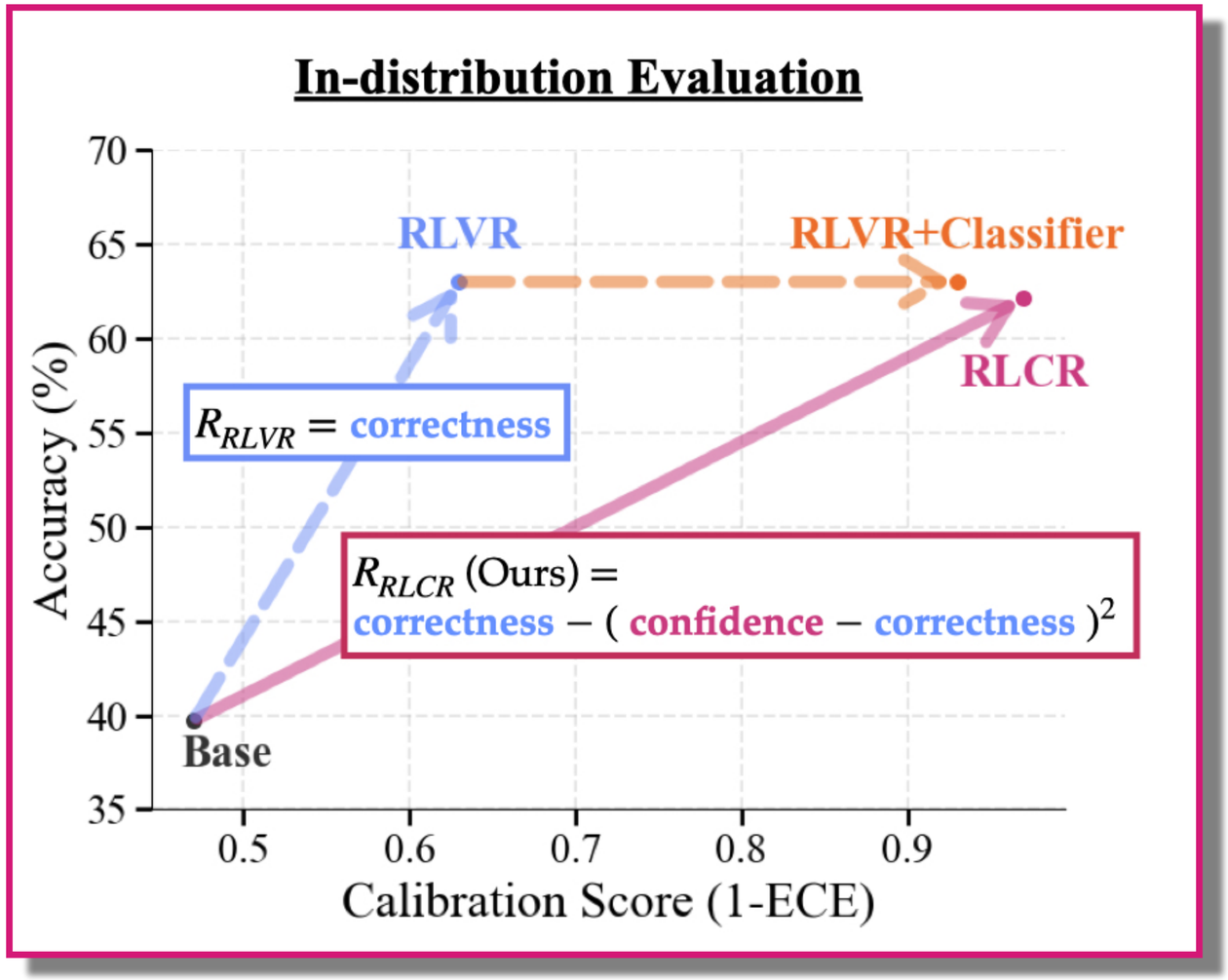
Initialize both RLVR and RLCR from Qwen-2.5-7B model and train using GRPO. Trained on HotPotQA and Math datasets (see paper).

We compare:

1. 💬 ***Base:*** *Model asked to output answers and verbalize confidence*
2. ⚙️ ***RLVR:*** *Model asked to output answers and verbalize confidence*
3. 🧠 ***RLVR+Classifier:*** *Train a separate classifier model (7B params) on* ***<think> <ans>*** *solutions from RLVR*
4. 🌟 ***RLCR****:* ***ours****! Train model to output answers and verbalized confidence in a single CoT!*

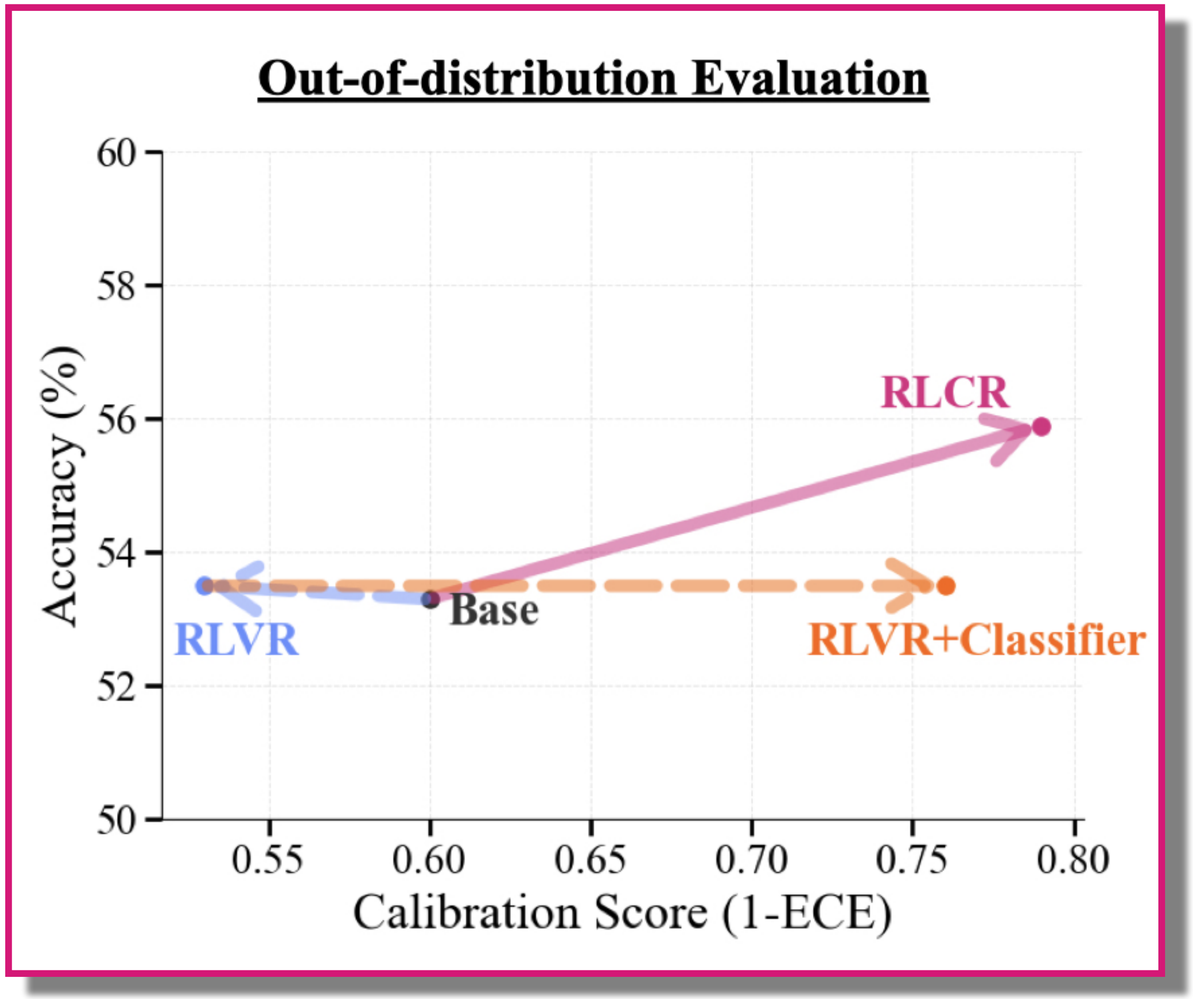
Results

- *Accuracy stays on par (or better) than RL baselines, with calibration error reduced by up to 90%.*
- *Outperforms post-hoc classifier on calibration.*



Results

While accuracy does not improve OOD for any methods,
***RLVR degrades calibration in OOD tasks, while RLCR significantly improves calibration*.**



Using Confidence for Test Time Scaling

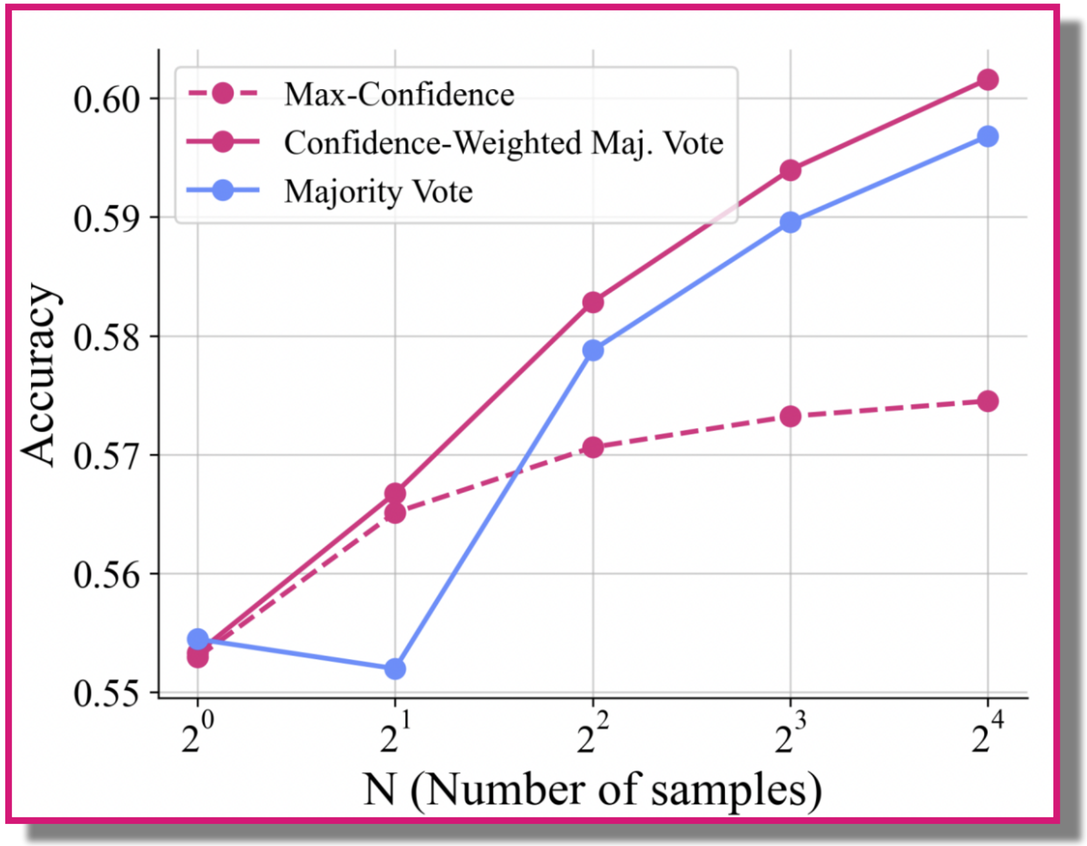
- ***Reward models are commonly used in test-time approaches like Best-of-N.***

- ***Insight: Model's confidence can be used as a proxy for reward! 2 simple algorithms:***

    1. ***Max-Confidence Selection:*** Choose the response with the highest self-reported confidence.

    2. ***Confidence-Weighted Majority Voting:*** *Aggregate multiple responses, weighting each vote by its confidence score.*

- RLCR enables easy test-time scaling without the need to train a reward model!



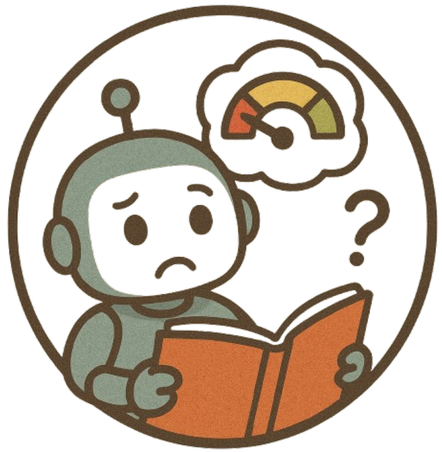
R L C R:

💡simple modification to RLVR that enables models to reason about their uncertainty.

💡provably incentivizes both accuracy and calibration

💡integrates into test-time scaling methods

Website (arxiv, code, models, slides):
rl-calibration.github.io



*Equal Contribution


Mehul Damani*


Isha Puri*


Stewart Slocum


Idan Shenfeld


Leshem Chosen

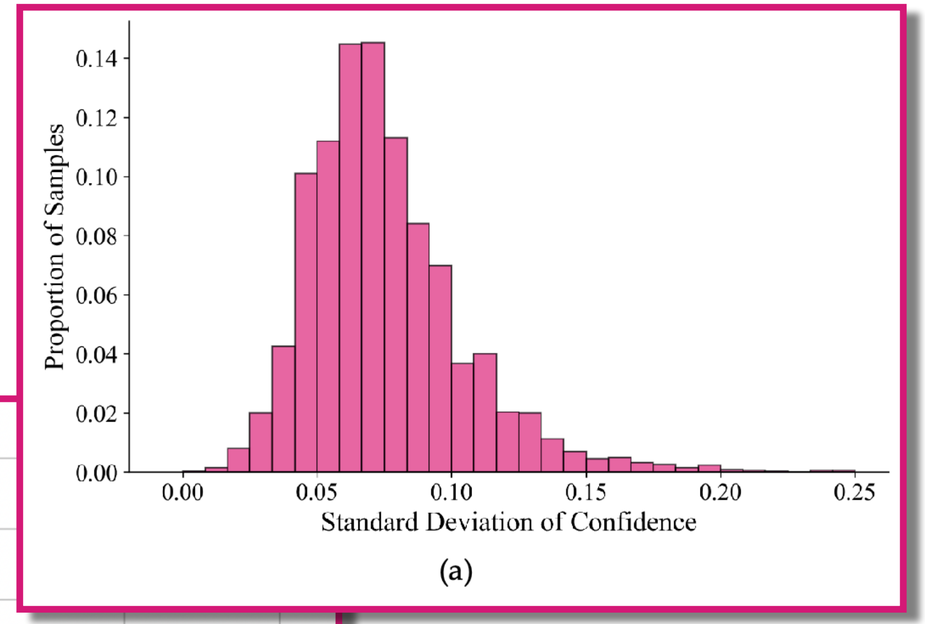

Yoon Kim

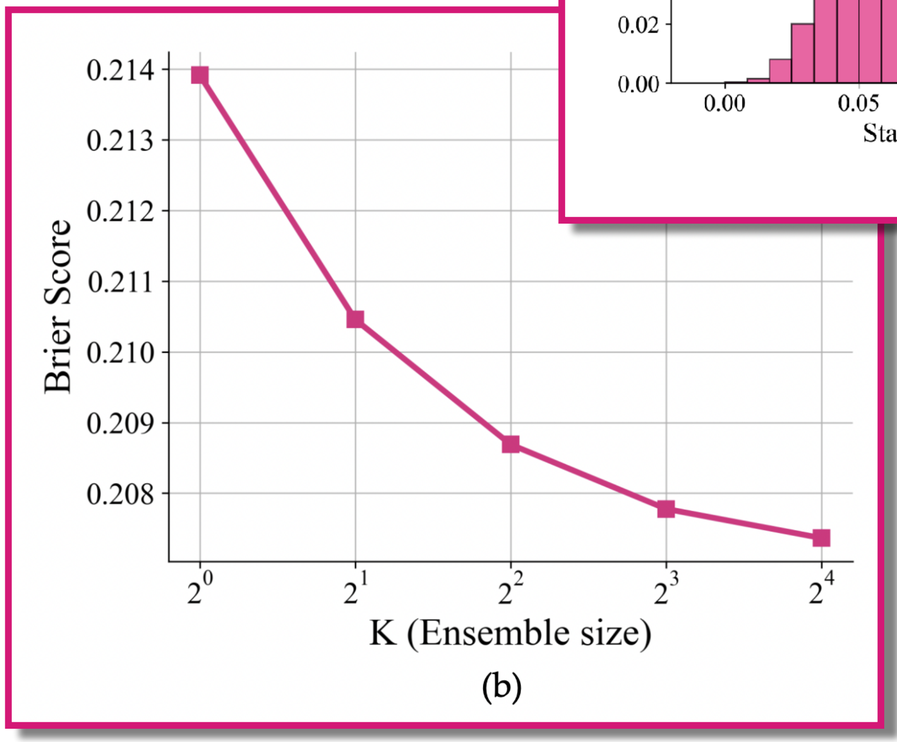

Jacob Andreas

Extra Slides

Uncertainty about Uncertainty

- ***Given a solution and answer, how much uncertainty is there about uncertainty?***
- *For fixed* **<think> <ans>** *, we sample multiple* **<analysis> <confidence>** *CoTs and compute:*
  1. *Variance of confidences.*
  2. *Mean confidence*

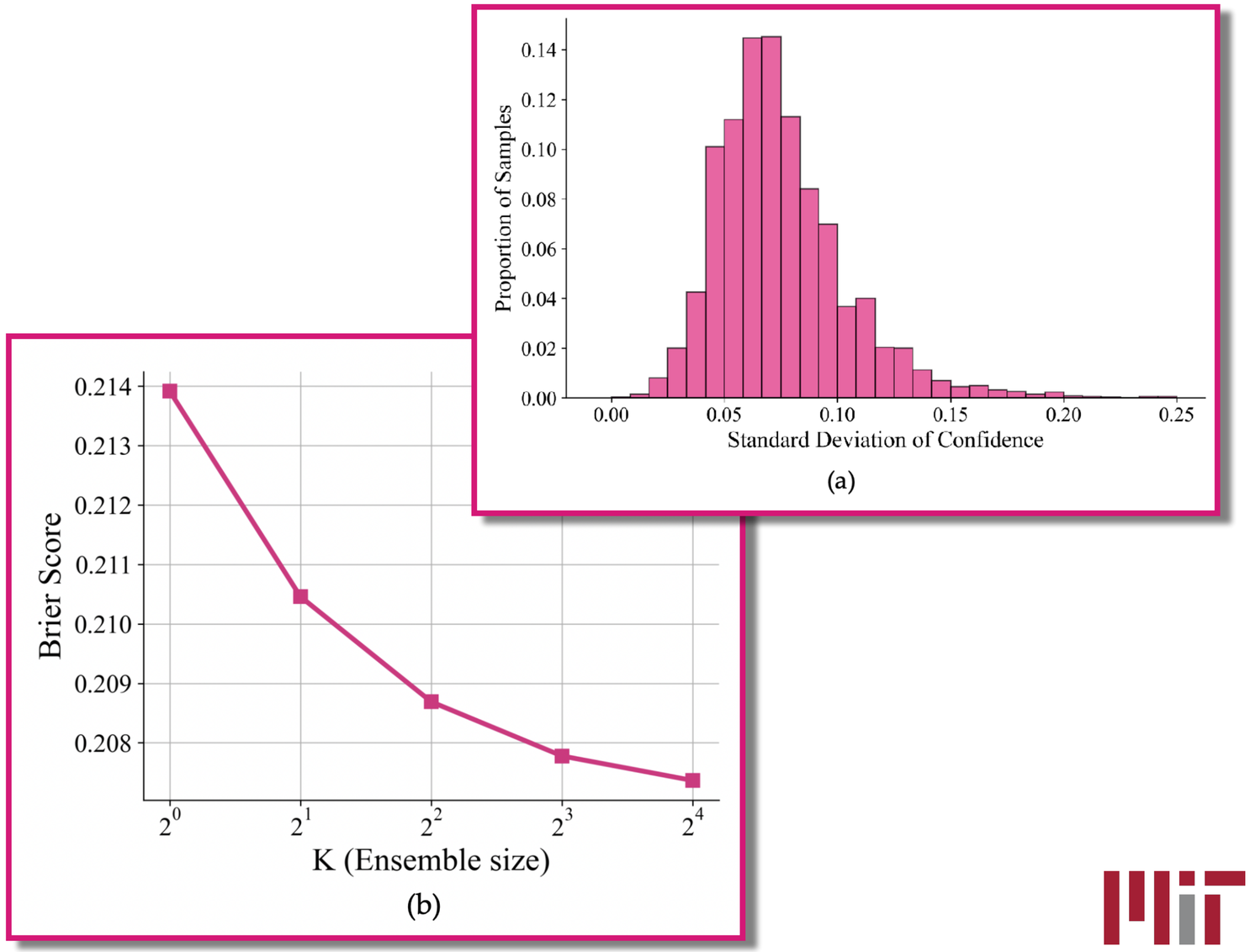




Uncertainty about Uncertainty

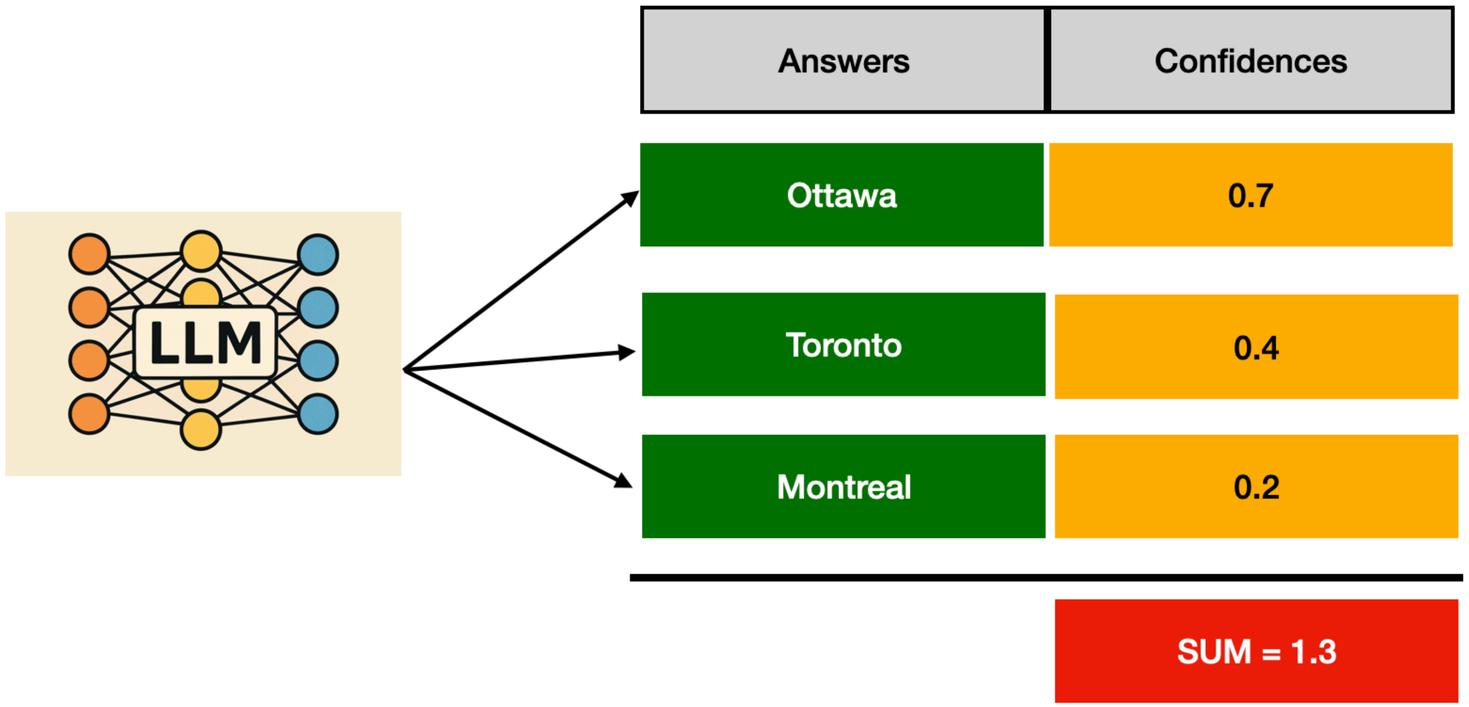
Low Variance
- *There is low "uncertainty about uncertainty"*

Mean aggregation improves calibration!
- *Calibration can also be improved by test-time scaling, although gains are modest.*



Inter Solution Consistency

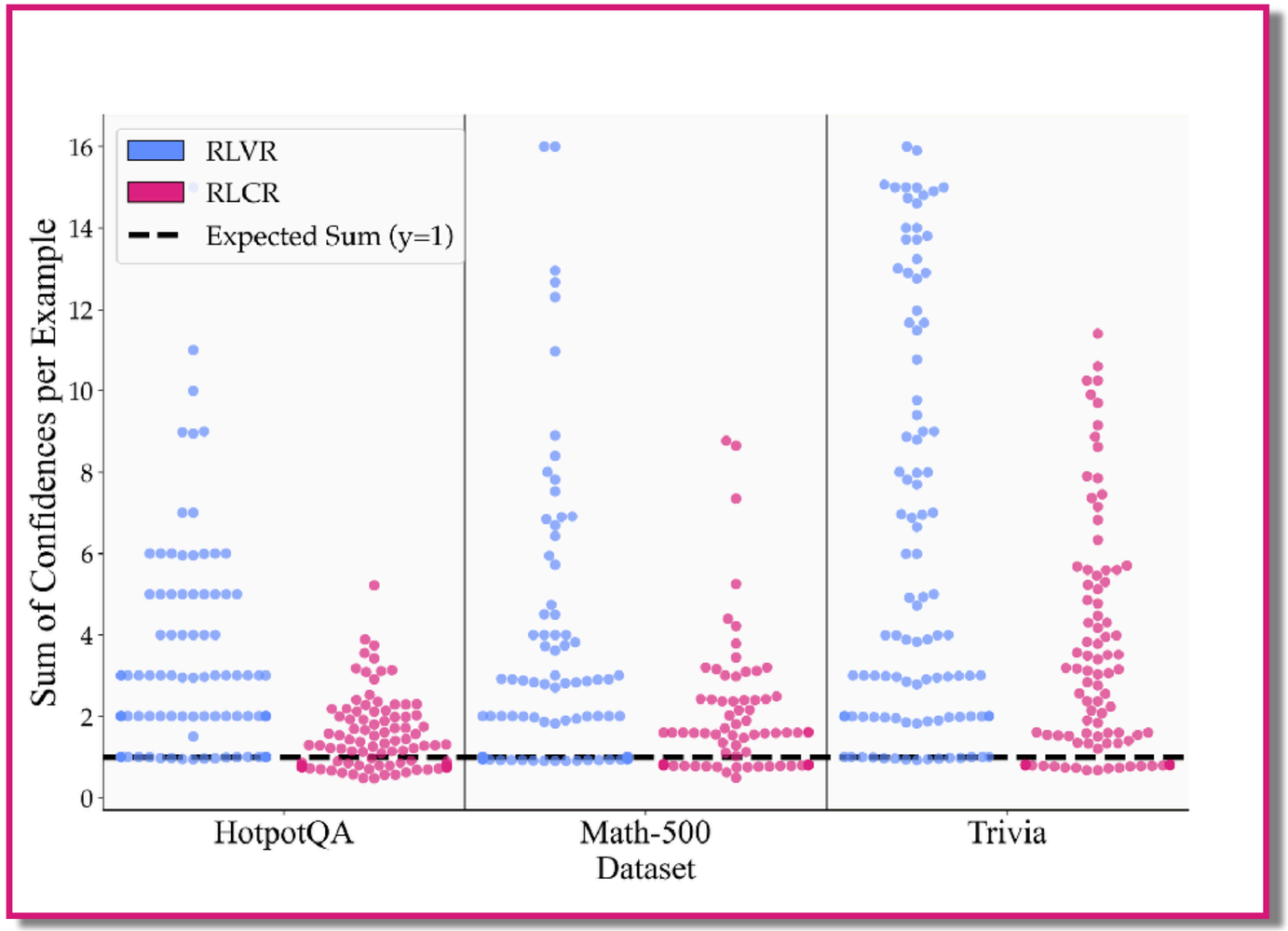
"What is the capital of Canada"?



When given mutually exclusive answers, we want the model to distribute its confidence across distinct answers such that the total confidence is less than or equal to 1 .

Undesirable

Inter Solution Consistency



When given mutually exclusive answers, we want the model to distribute its confidence across distinct answers such that the total confidence is less than or equal to 1 .

RLCR has much better consistency than RLVR, but room for improvement remains.

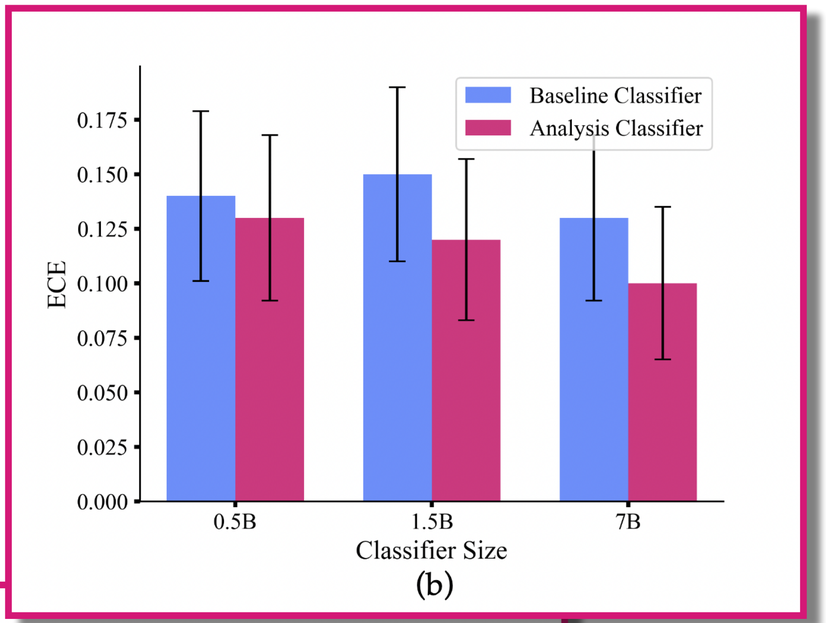
Reasoning About Uncertainty

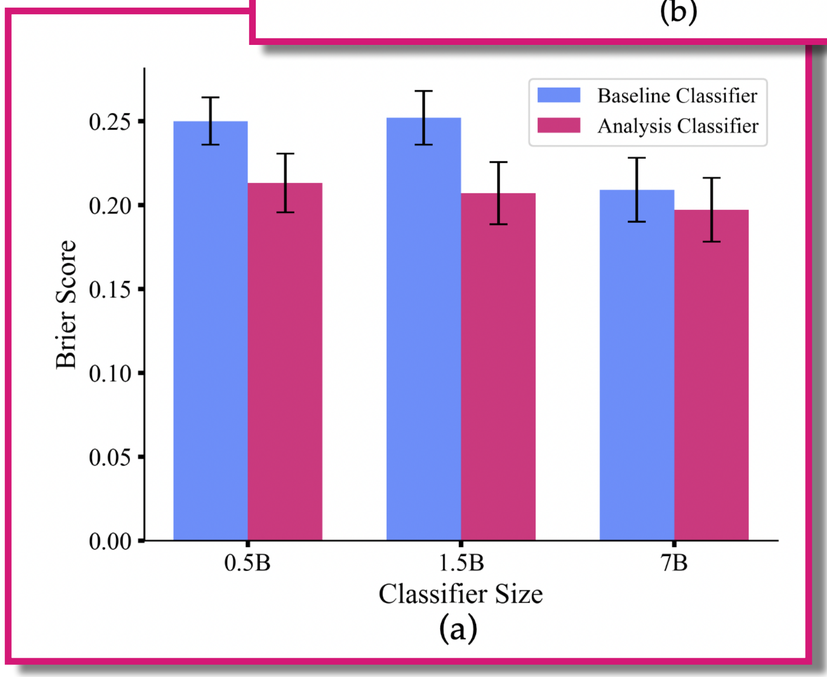
We trained two types of classifiers:

1. ***Baseline: Trained on RLVR CoTs***
   <think> <ans>

2. ***Analysis: Trained on RLCR CoTs, but without <confidence> values.***
   <think> <ans> <analysis>

Result: Analysis classifier outperformed the baseline, particularly for smaller model sizes.





Results

(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**

Training Process