

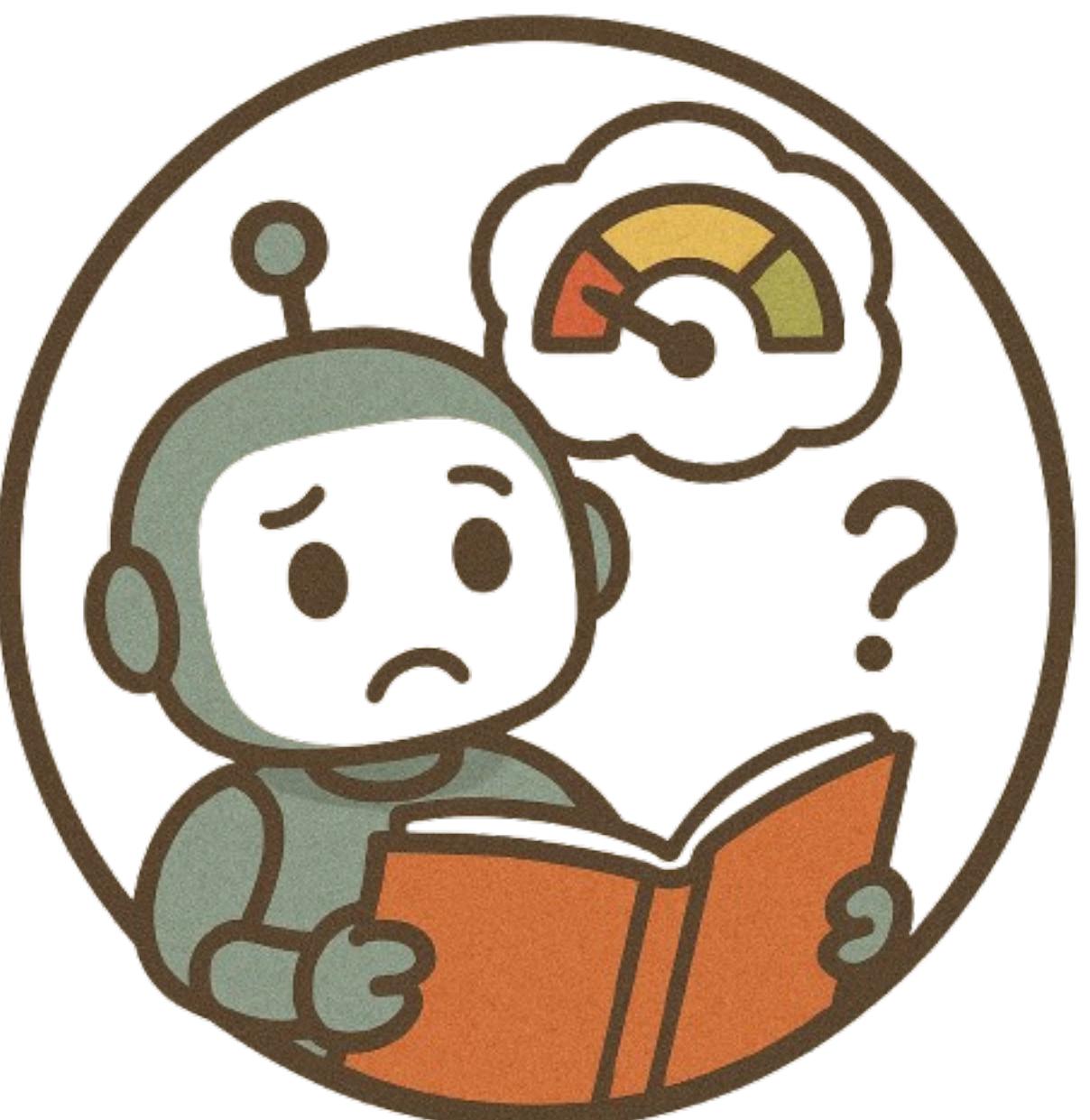


# Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty

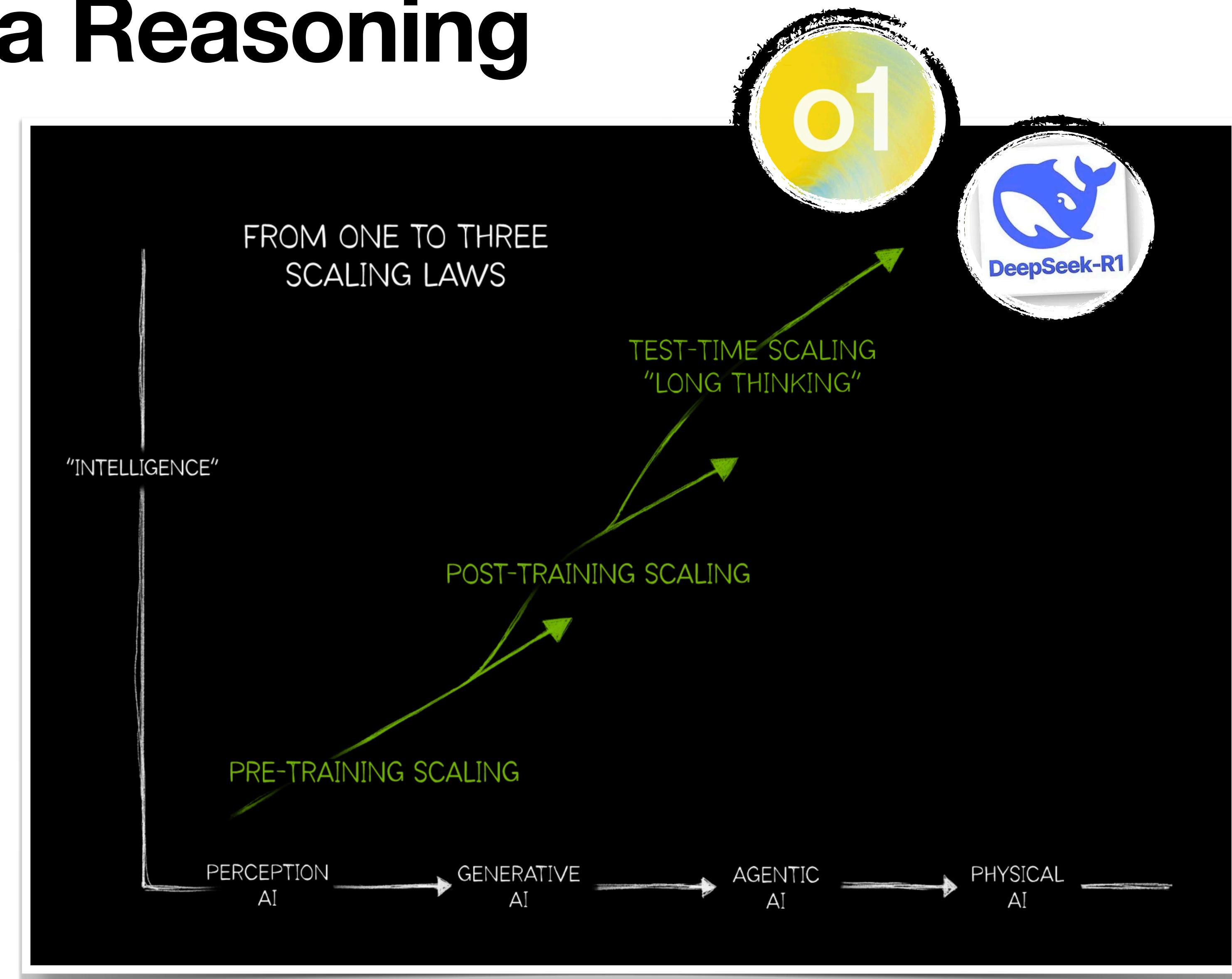
Mehul Damani\*, Isha Puri\*, Stewart Slocum, Idan  
Shenfeld, Leshem Choshen, Yoon Kim, Jacob Andreas  
MIT CSAIL

NVIDIA – October 31, 2025

\*Equal Contribution



# Viva La Reasoning



# Standard Approaches to Reasoning Training

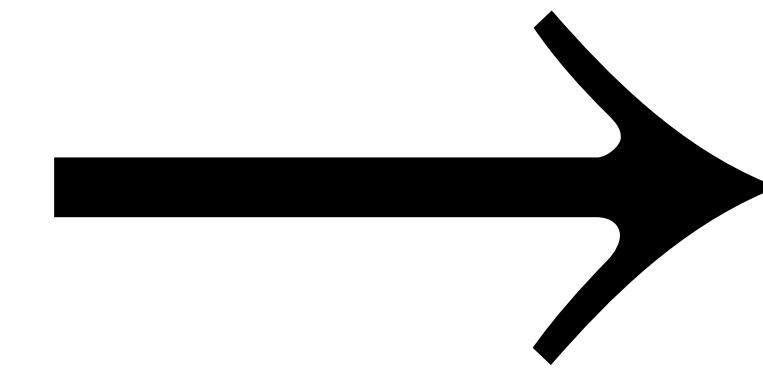
- “Reasoning models” are trained via RL to “think out loud” before answering questions – these models do very well on math and programming
- Standard approach to reasoning training:

**R**einforcement  
**L**earning w/  
**V**erifiable  
**R**ewards



# Standard Approaches to Reasoning Training

Reinforcement  
Learning  
Verifiable  
Rewards



**Binary Correctness Reward:**

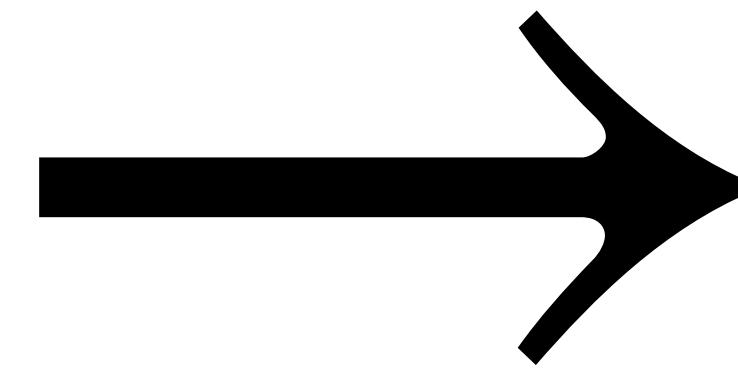
$$R_{correctness} = \begin{cases} 1 & \text{if answer is correct} \\ 0 & \text{if answer is wrong} \end{cases}$$



(The consequences of)

# Standard Approaches to Reasoning Training

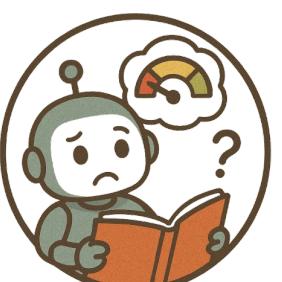
Reinforcement  
Learning  
Verifiable  
Rewards



## 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



(The consequences of)

# Standard Approaches to Reasoning Training

The New York Times

## A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse

A new wave of “reasoning” systems from companies like OpenAI is producing incorrect information more often. Even the companies don’t know why.

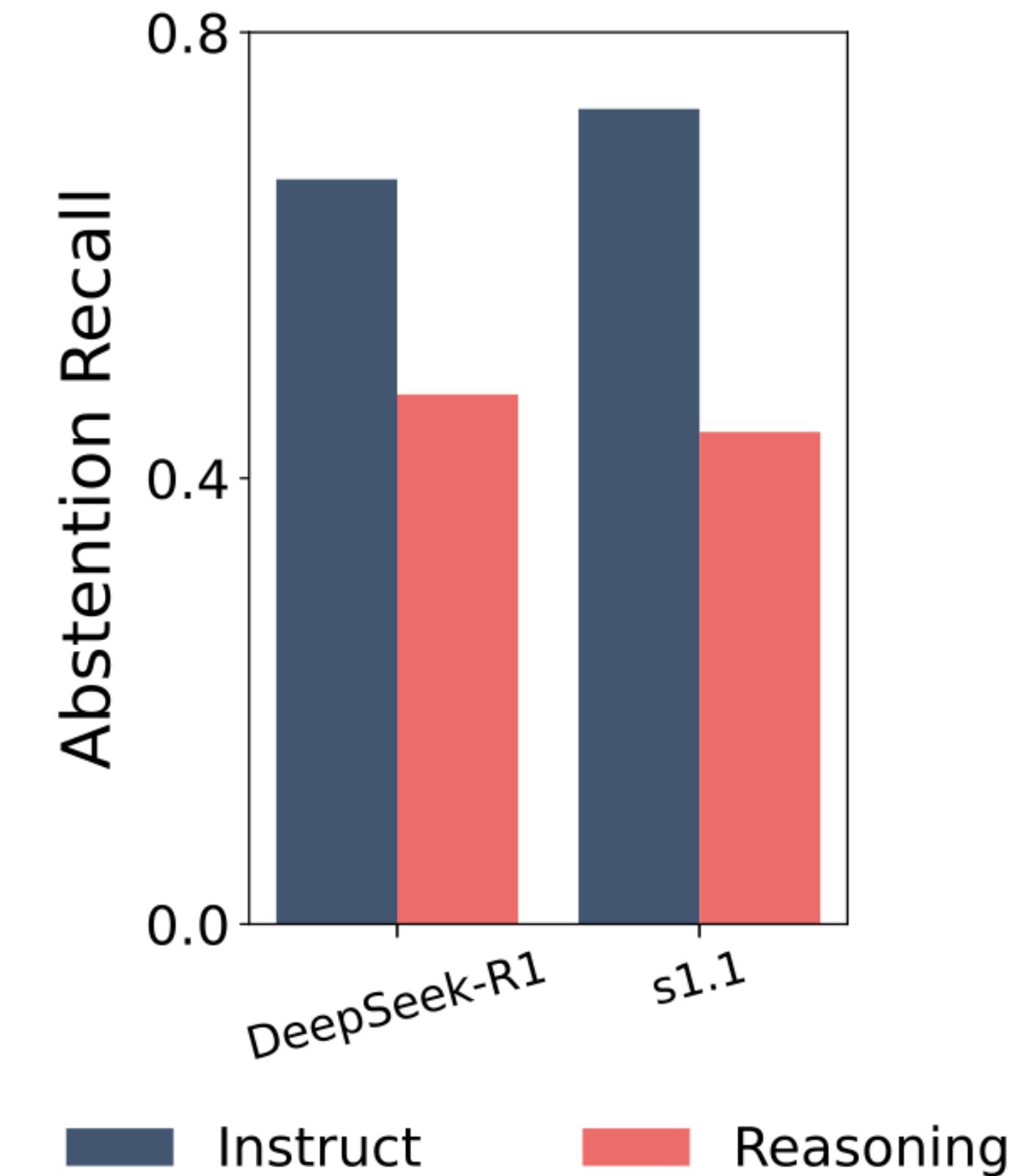
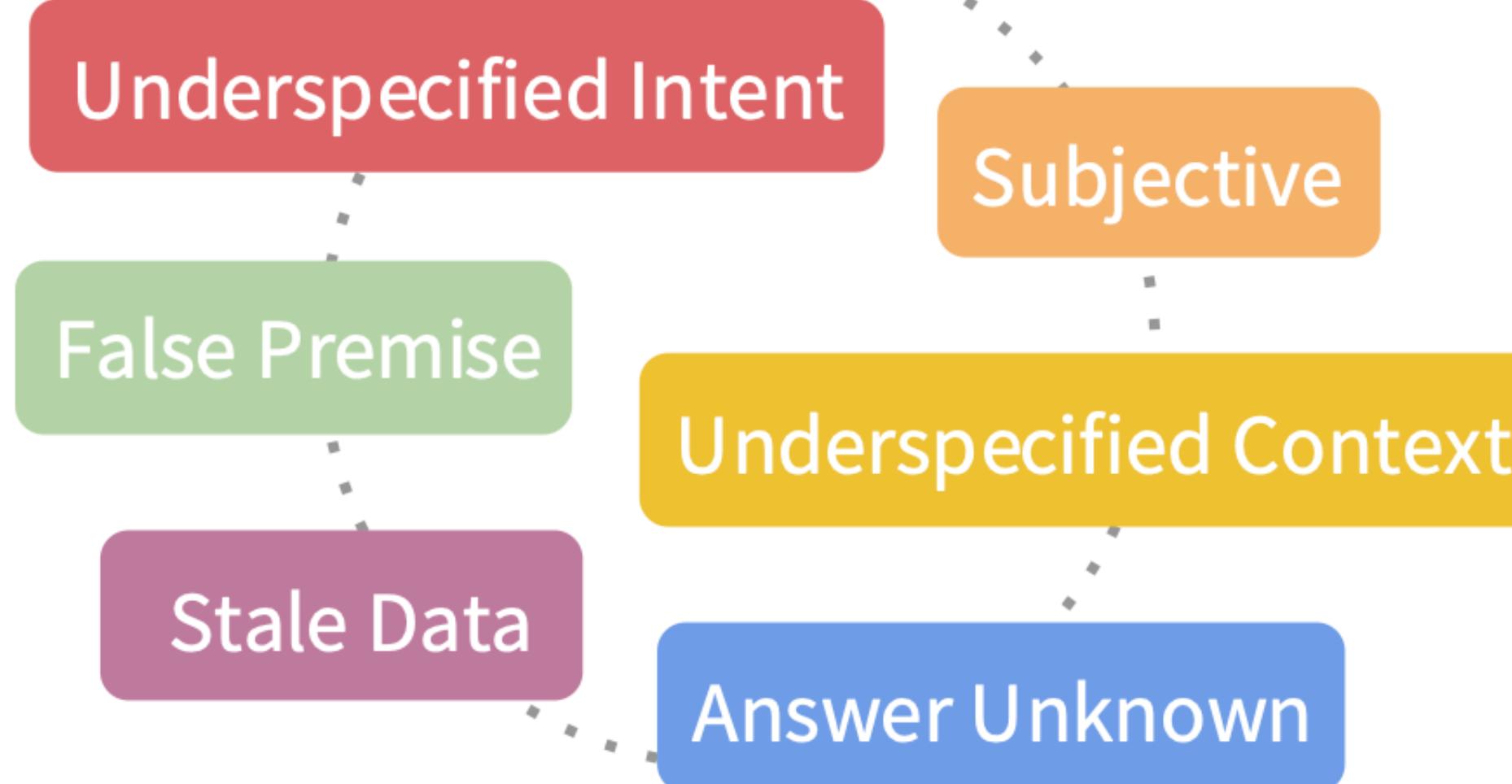
TC TechCrunch

## OpenAI’s new reasoning AI models hallucinate more

Table 4: Hallucination evaluations

Dataset	Metric	o3	o1
SimpleQA	hallucination rate (lower is better)	0.51	0.44
PersonQA	hallucination rate (lower is better)	0.33	0.16

# Models try to answer even ambiguous or unanswerable questions

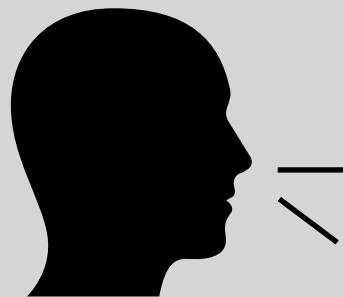


(The consequences of)

# Standard Approaches to Reasoning Training



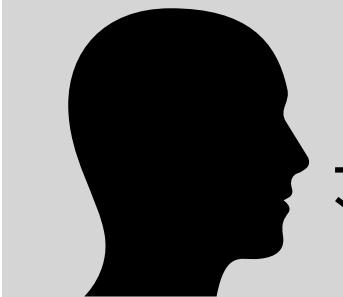
Models don't have a good sense of their own confidence.



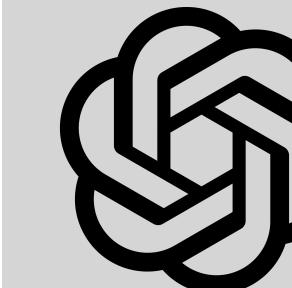
Was Belva Lockwood the first woman to argue before the Supreme Court in November 1880? Give an answer and a numerical estimate of your confidence.



“Yes, 92%”



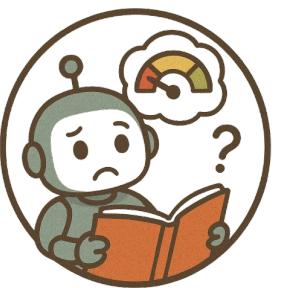
Was Belva Lockwood the first woman to argue before the Supreme Court in November 1877? Give an answer and a numerical estimate of your confidence.



“Yes, 96%”

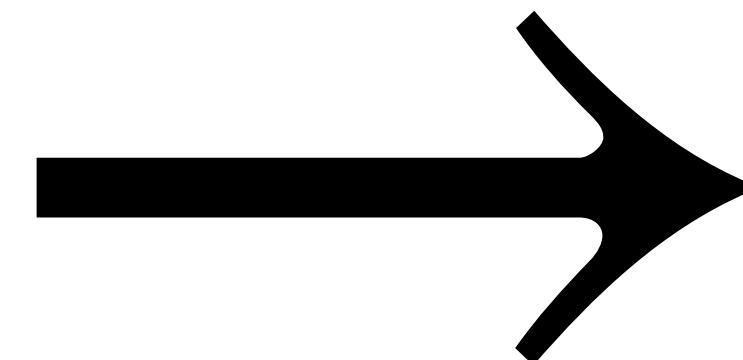


**How can we better calibrate our models  
without compromising on the strengths  
of RLVR?**



# RLCR

**R**einforcement  
L<sub>earning w/</sub>  
**V**erifiable  
**R**ewards



**R**einforcement  
L<sub>earning w/</sub>  
**C**alibrated  
**R**ewards

**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>

## Reason about uncertainty

### 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>

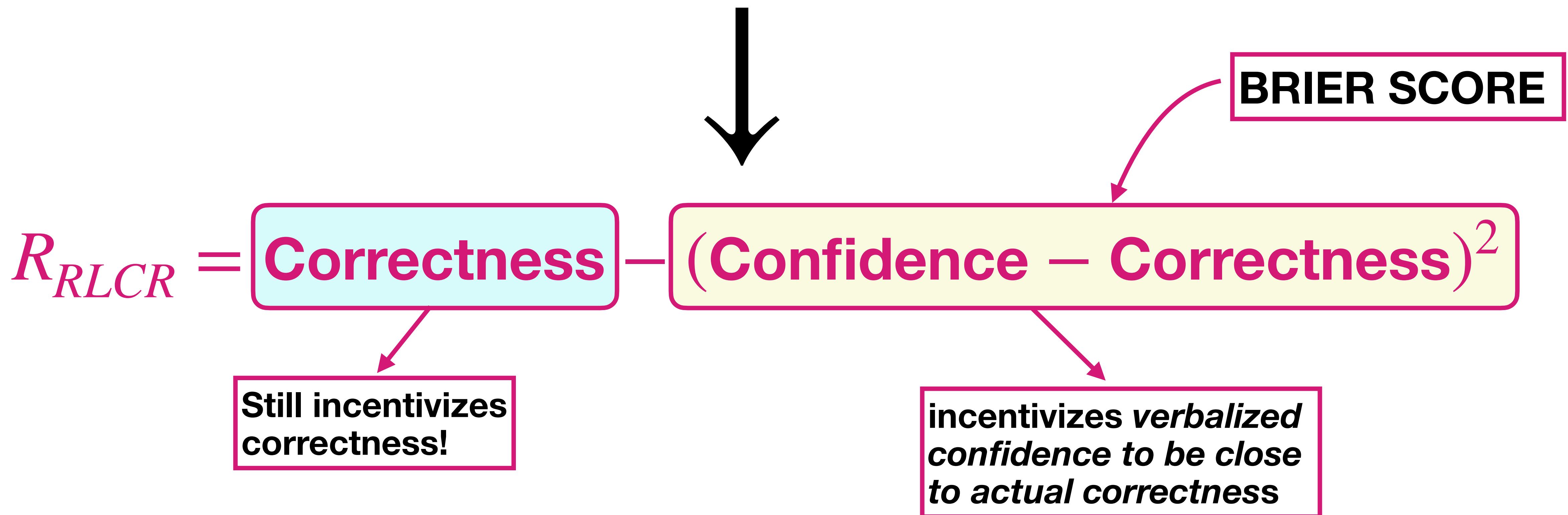


## Verbalize confidence

# RLCR Reward

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

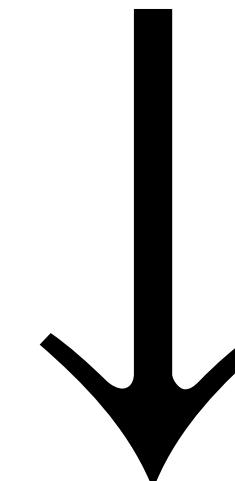
$$R_{RLVR} = \text{Correctness}$$



# RLCR Reward

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

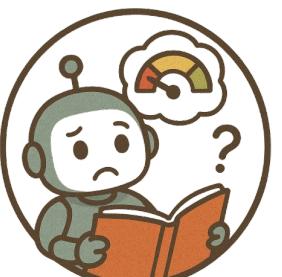
$$R_{RLVR} = 1_{y=y^*}$$



$$R_{RLCR} = 1_{y=y^*} - (q - 1_{y=y^*})^2$$

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

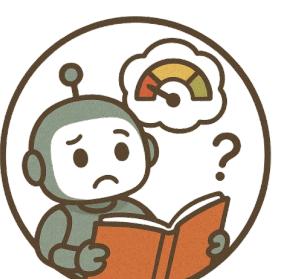
**<think>** The question asks for the song with which Lulu represented the UK in the 1969 Eurovision Song Contest. Lulu is a well-known British singer, and the Eurovision Song Contest is an annual competition where countries submit songs to be performed and judged. I need to recall the specific song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

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

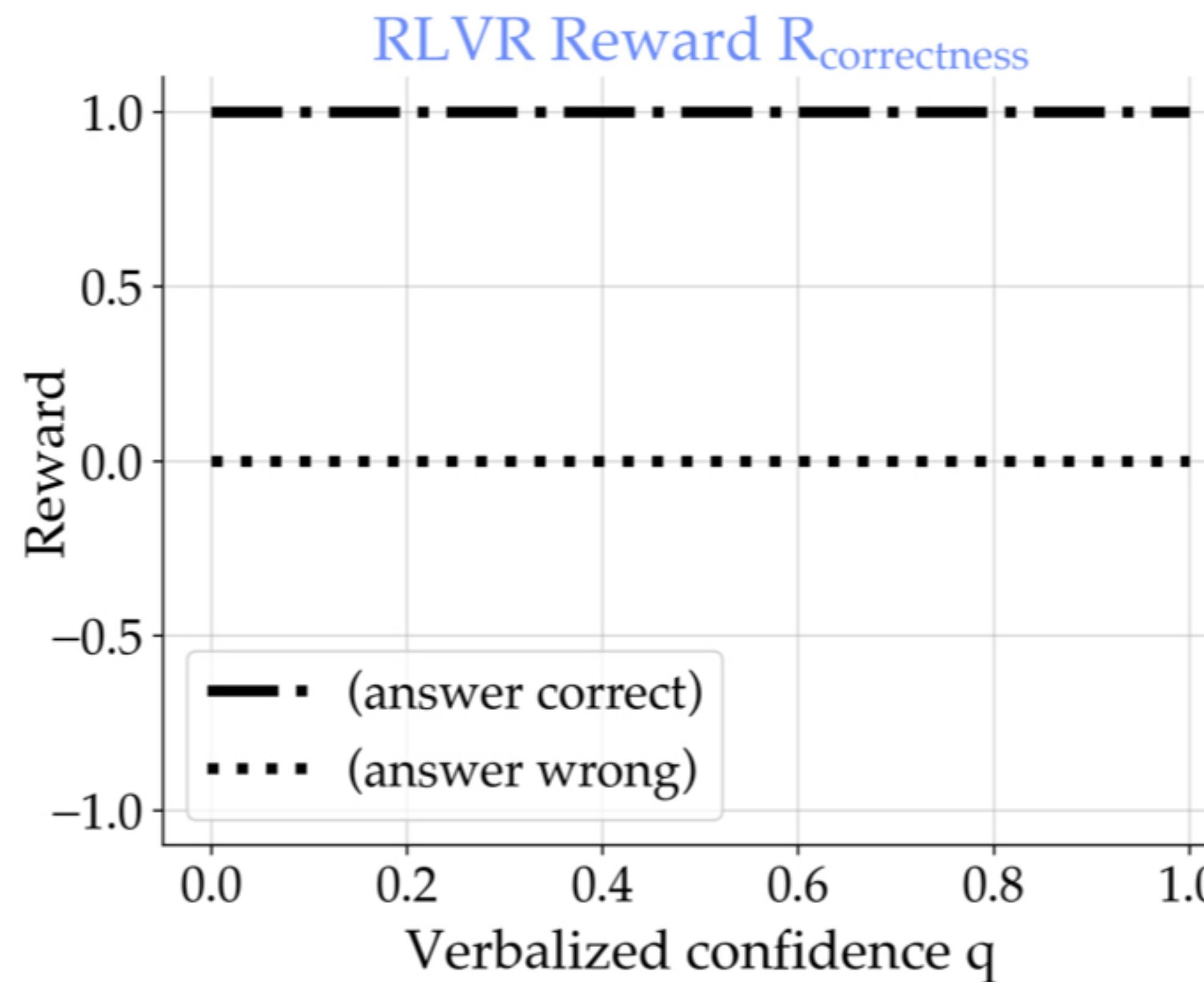
**<analysis>** There is a high level of uncertainty in this answer because while Lulu did represent the UK in the 1969 Eurovision Song Contest, the specific song she performed is not widely known or easily recalled. The song "To Sir With Love" is a popular song by Lulu, but it is not commonly associated 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>**

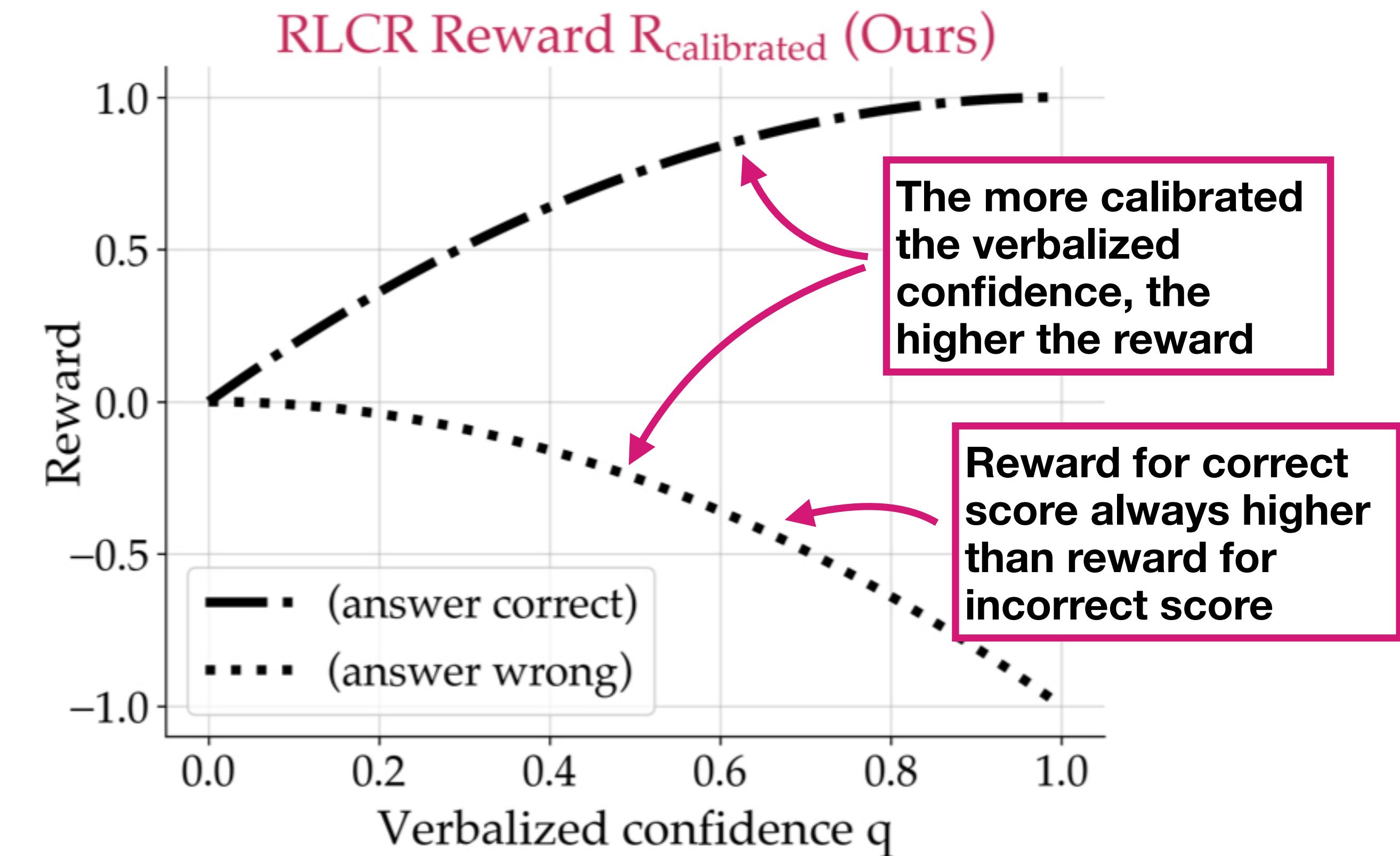
## RLCR CoT



# Reward Comparison



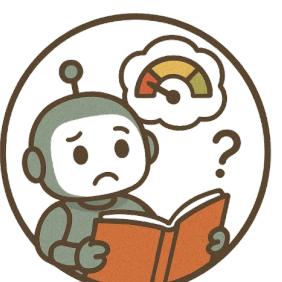
$$R_{\text{RLVR}} = \text{correctness}$$

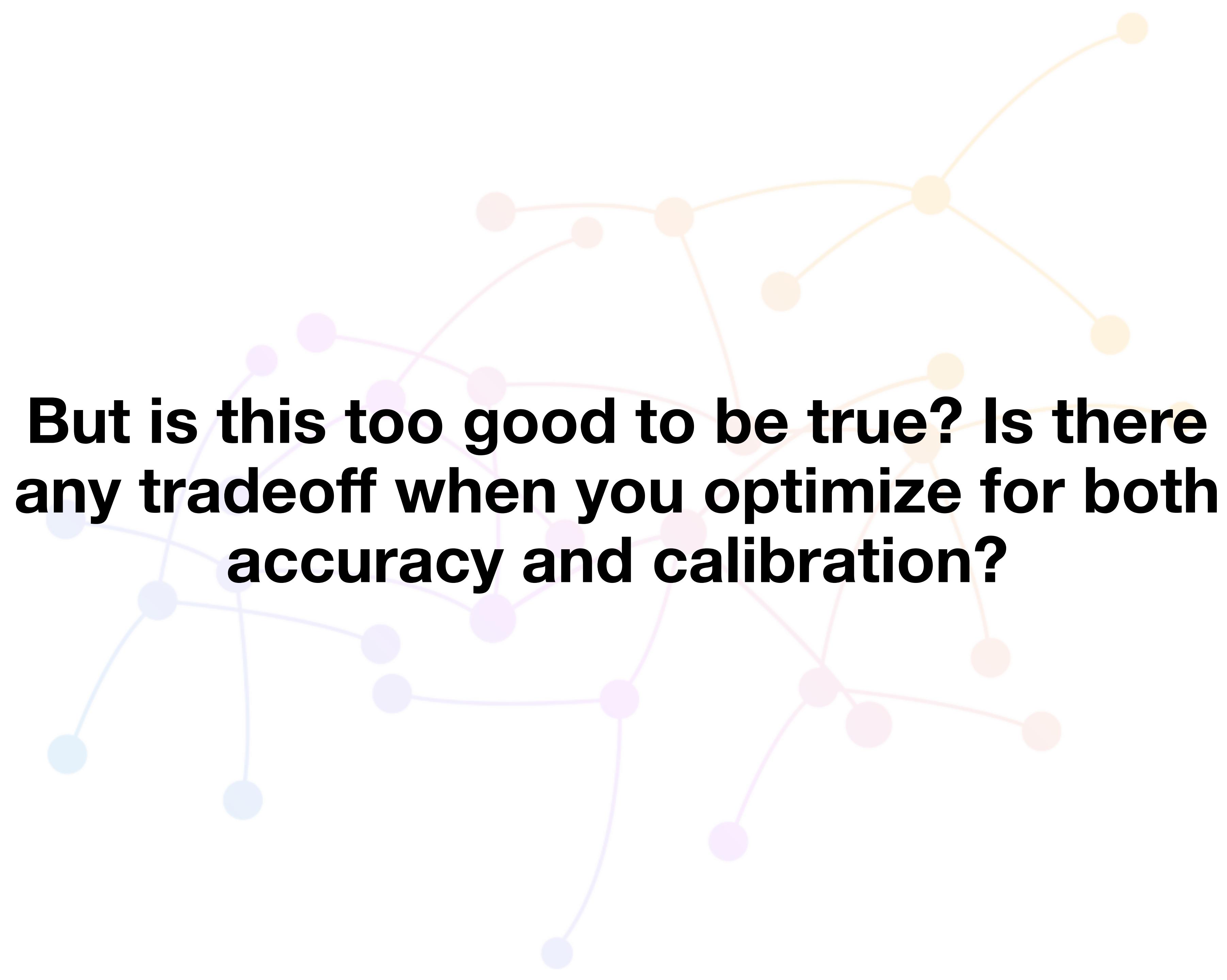


The more calibrated the verbalized confidence, the higher the reward

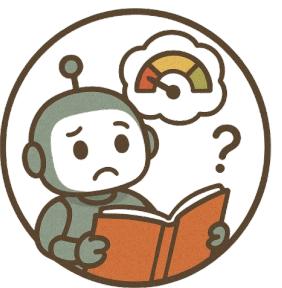
Reward for correct score always higher than reward for incorrect score

$$R_{\text{RLCR}} (\text{Ours}) = \text{correctness} - (\text{confidence} - \text{correctness})^2$$





**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.

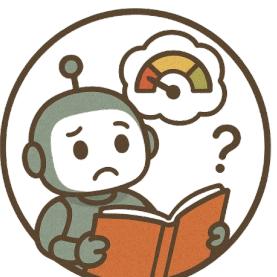


# Proper Scoring Rules

**Scoring Rule:** A scoring rule is a function  $S$  that measures the quality of a confidence estimate  $q$ .

**Proper Scoring Rule:** A scoring rule is **proper** if its expected value is minimized by confidence score that matches true outcome probability

$$\mathbb{E}_{a \sim p(a)} S(p(a), a) \leq \mathbb{E}_{a \sim p(a)} S(q, a) \quad \forall q$$



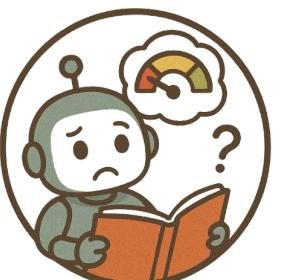
# Proper Scoring Rules - Examples

**Log Loss**

$$S(q, a) = a \log q + (1 - a) \log(1 - q)$$

**Brier Score**

$$S(q, a) = (q - a)^2$$

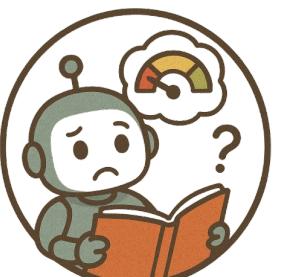


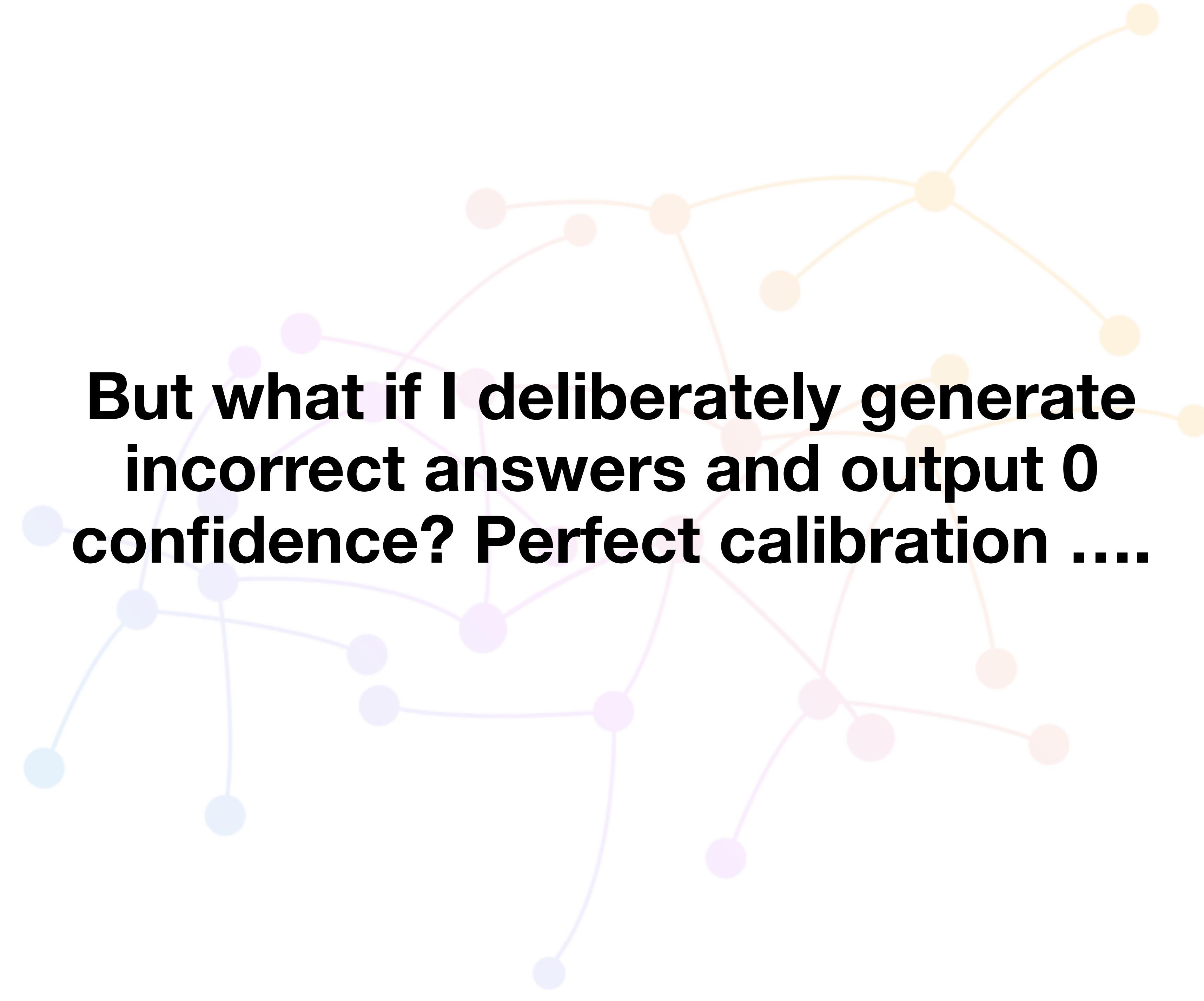
# Calibration Incentive

$$R_{RLCR} = 1_{y=y^*} - (q - 1_{y=y^*})^2$$

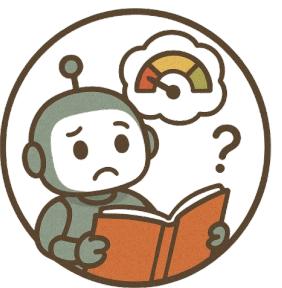
- The first term does not depend on the confidence estimate.
- The second term is exactly the Brier Score!

**Proper scoring rule in reward function incentivizes honest confidences for a given answer!**





**But what if I deliberately generate  
incorrect answers and output 0  
confidence? Perfect calibration ....**



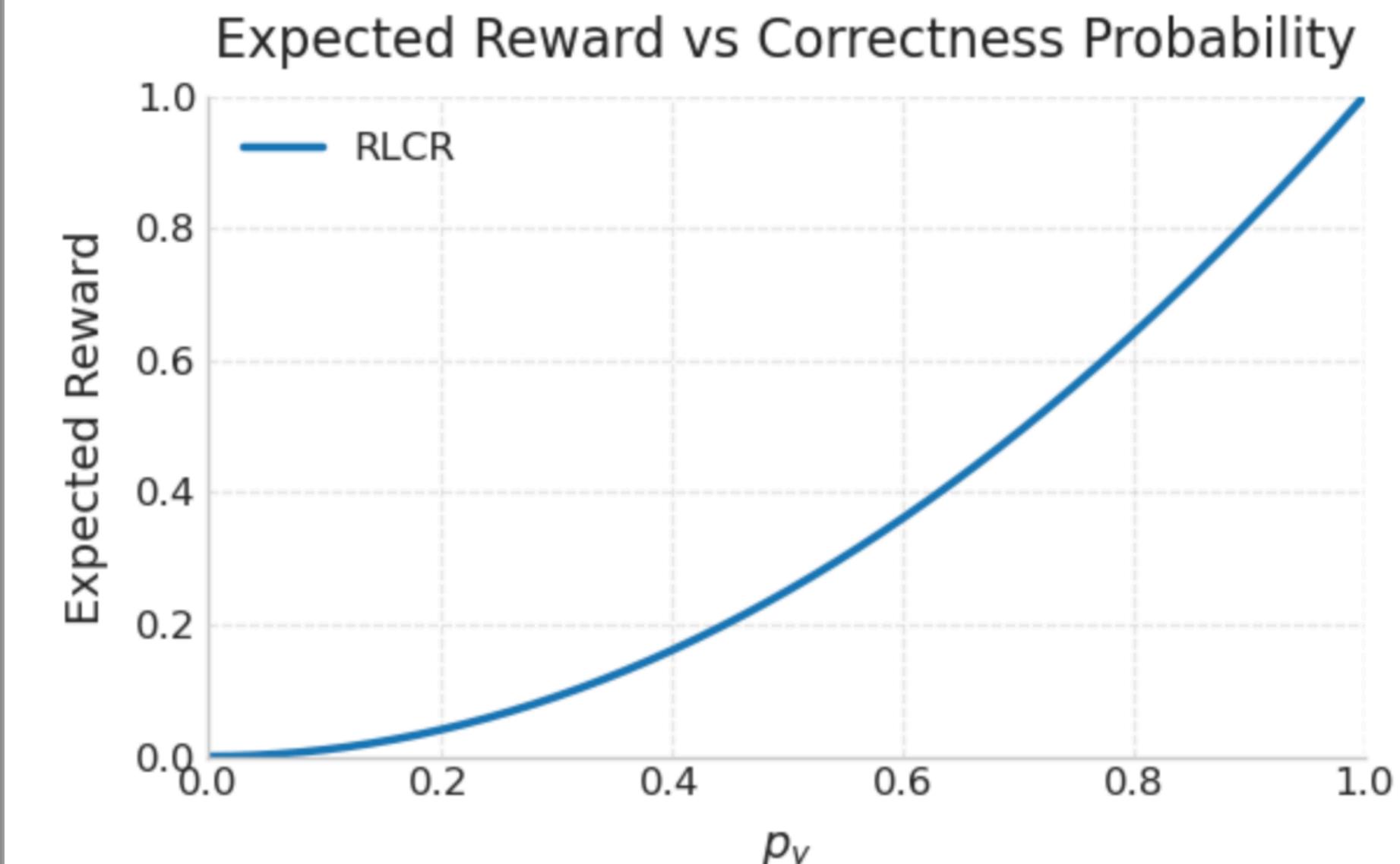
# Correctness Incentive

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

- If true correctness probability for answer  $y$  is  $p_y$ , then this is the confidence model should report.
- **Expected reward** when honestly reporting confidence:

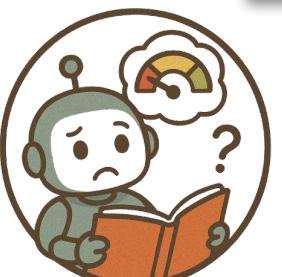
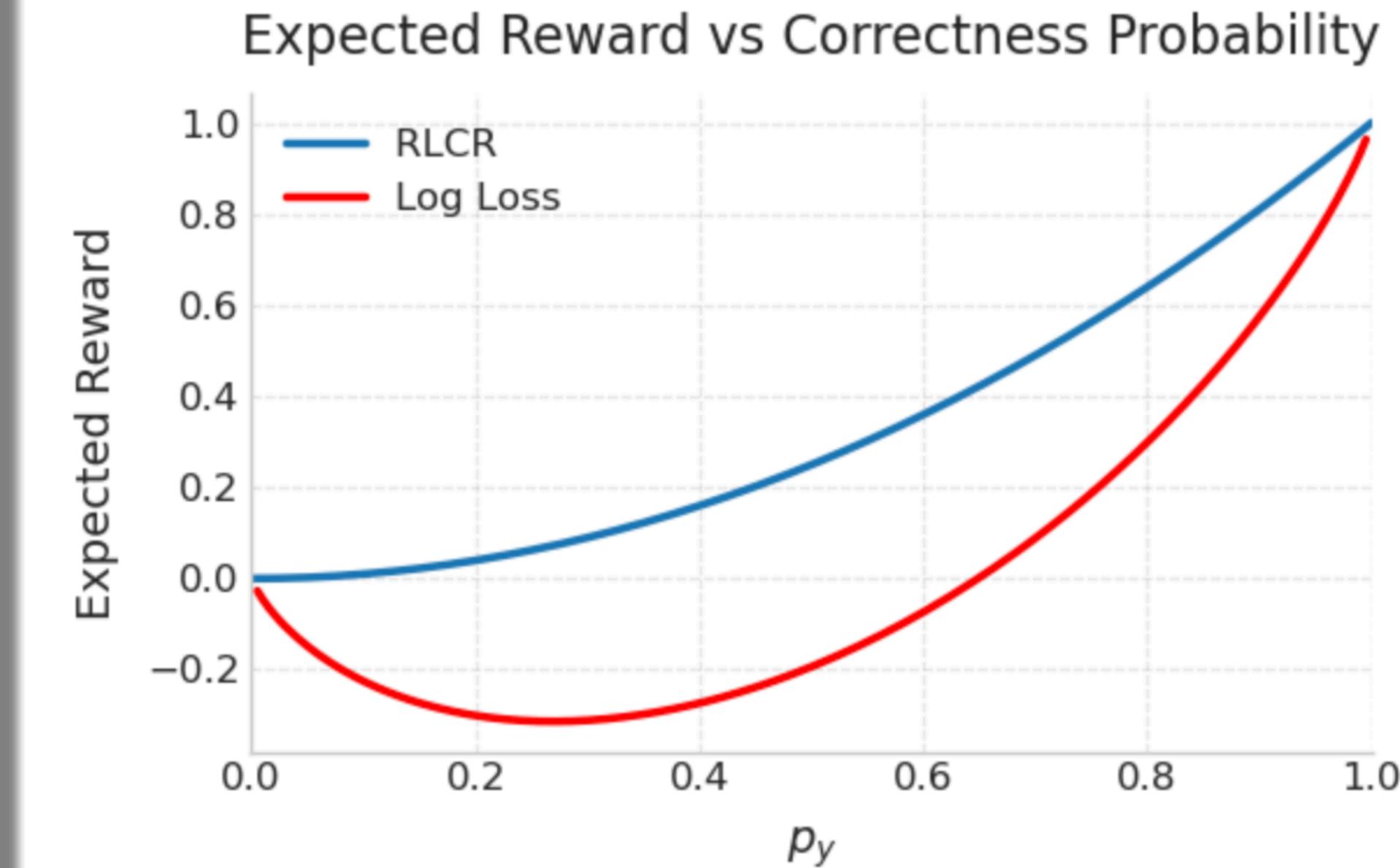
$$\begin{aligned} R &= p_y * (1 - (1 - p_y)^2) + (1 - p_y) * (-p_y^2) \\ &= p_y^2 \end{aligned}$$

- Answers with higher success probability get higher reward!



# Log Loss doesn't have Correctness Incentive

- **Expected reward** when honestly reporting confidence using log loss:  
$$R = p_y * (1 + \log(p_y)) + (1 - p_y) * (\log(1 - p))$$
- **Not strictly increasing** in  $p_y$ .
- **Optimal answer to output is not always the accuracy-maximizing one!**



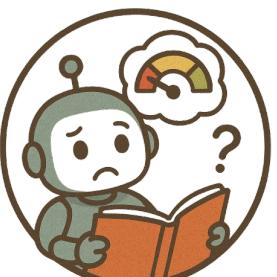
# Theorem Summary

RLCR reward has:

- **Calibration Incentive:** Report honest confidences because reward function uses a proper scoring rule
- **Correctness Incentive:** Output answers that are most likely to be correct as reward increases monotonically with success probability (like RLVR).

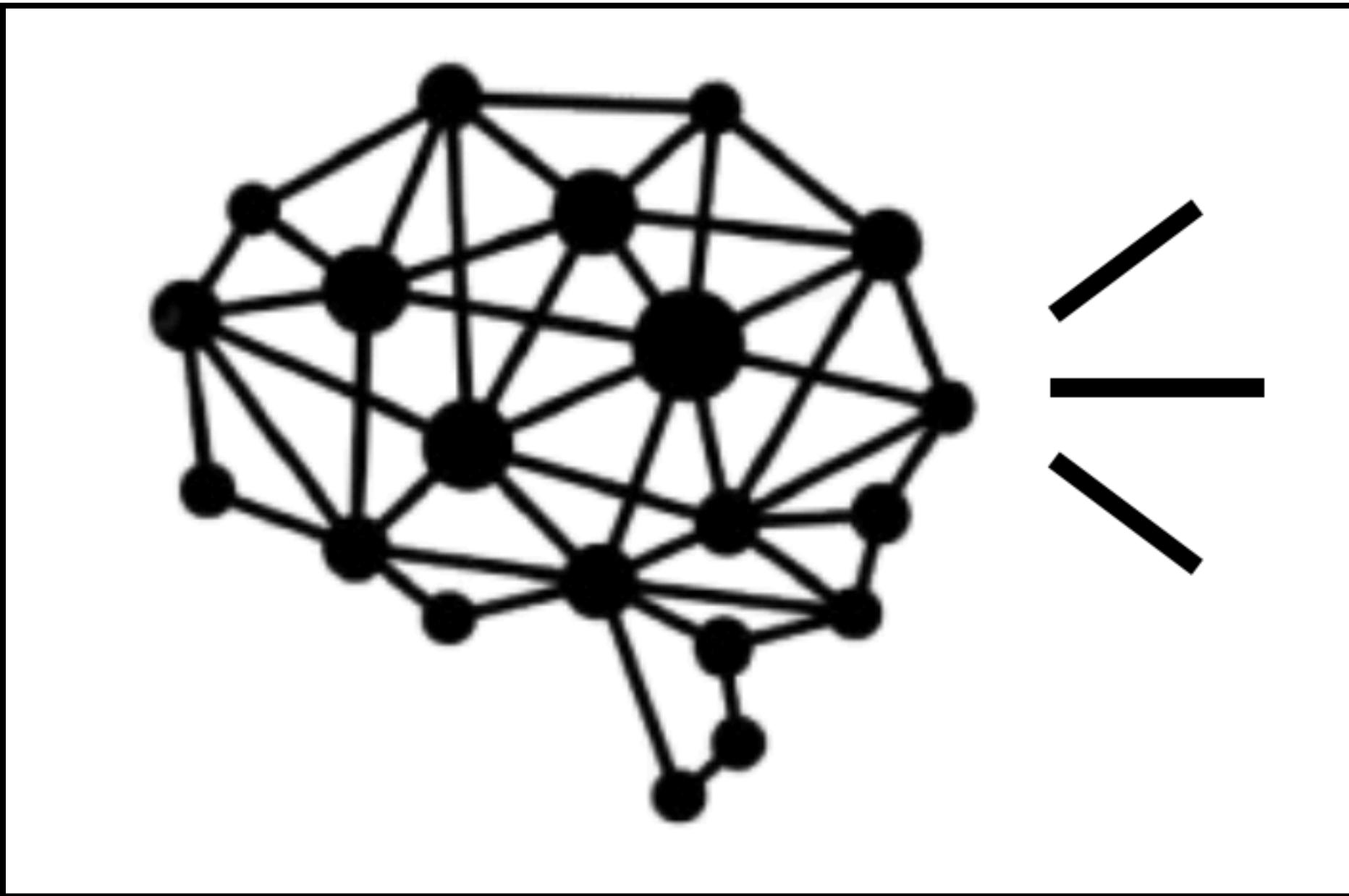
**Theorem 1.** Suppose, for any prediction  $y$  and verbalized confidence  $q$ , that the success indicator  $\mathbb{1}_{y \equiv y^*}$  is distributed as  $Bernoulli(p_y)$ . Then RLCR reward satisfies two properties:

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.

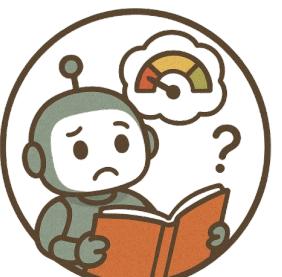


# Why Verbalized Uncertainty?

Why should we care about verbalized uncertainty, instead of, say, a separate uncertainty classifier?



- ***Efficiency (one model)***
- ***Allows model to think and analyze itself before outputting an uncertainty***
- ***(Speculative) maybe uncertainty calibration is a general capability that can enhance other capabilities***

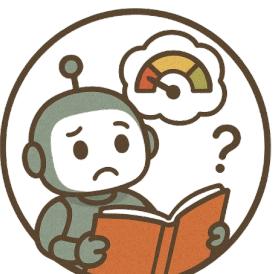


# Experimental Setup

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

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 confidence classifier model (7B params) given **<think> <ans>** solutions from RLVR
4. ⭐ **RLCR: ours!** Train model to output answers and verbalized confidence in a single CoT!



# Datasets - Hotpot

The HotPotQA dataset tests multi-hop reasoning. Given 2 relevant and 8 distractor paragraphs, one must connect pieces of information to answer questions.



To force uncertainty, we remove either 0, 1, or both relevant paragraphs.

# Datasets - Math

The BigMath dataset contains 250k math problems curated for RL

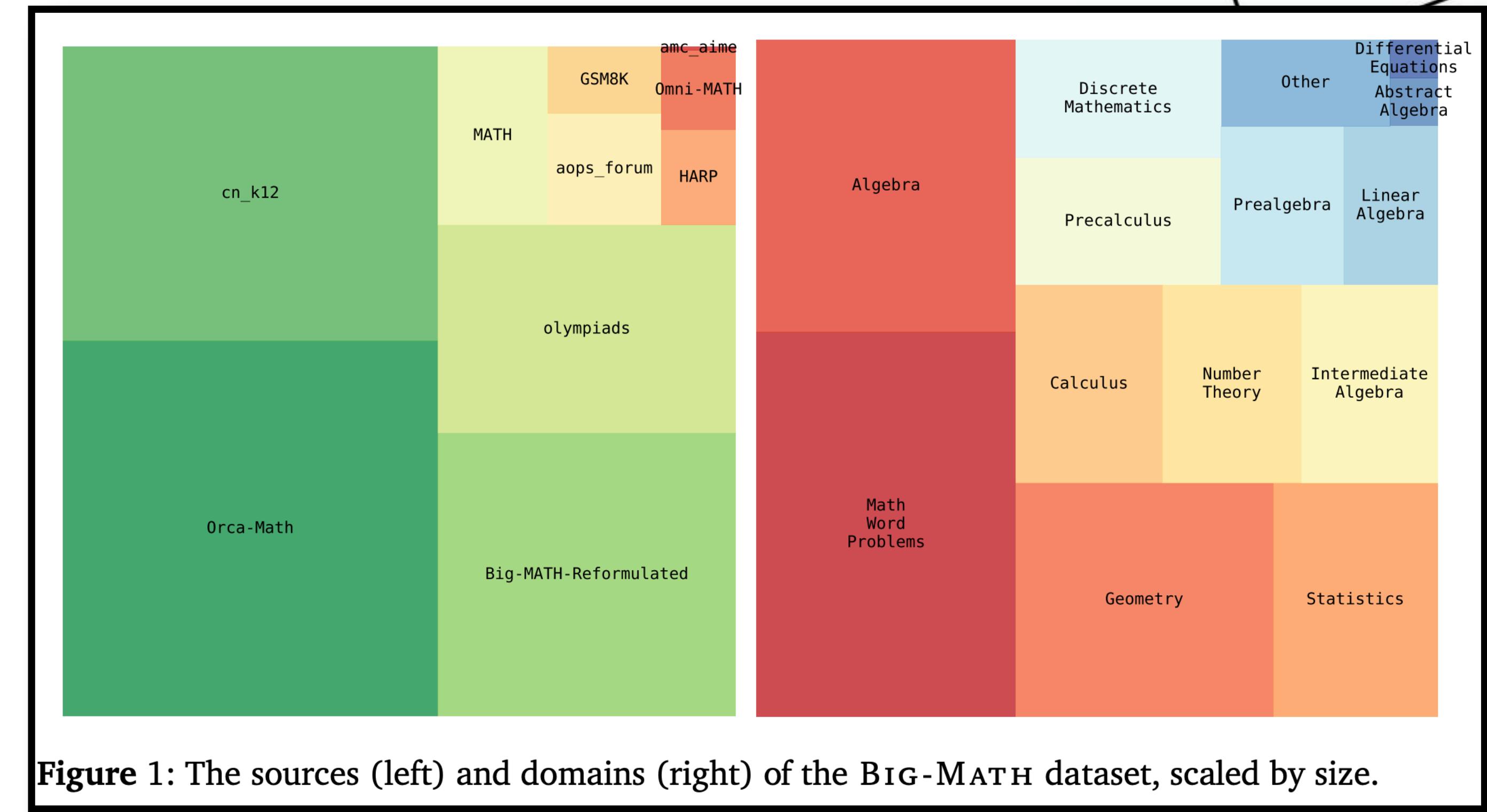
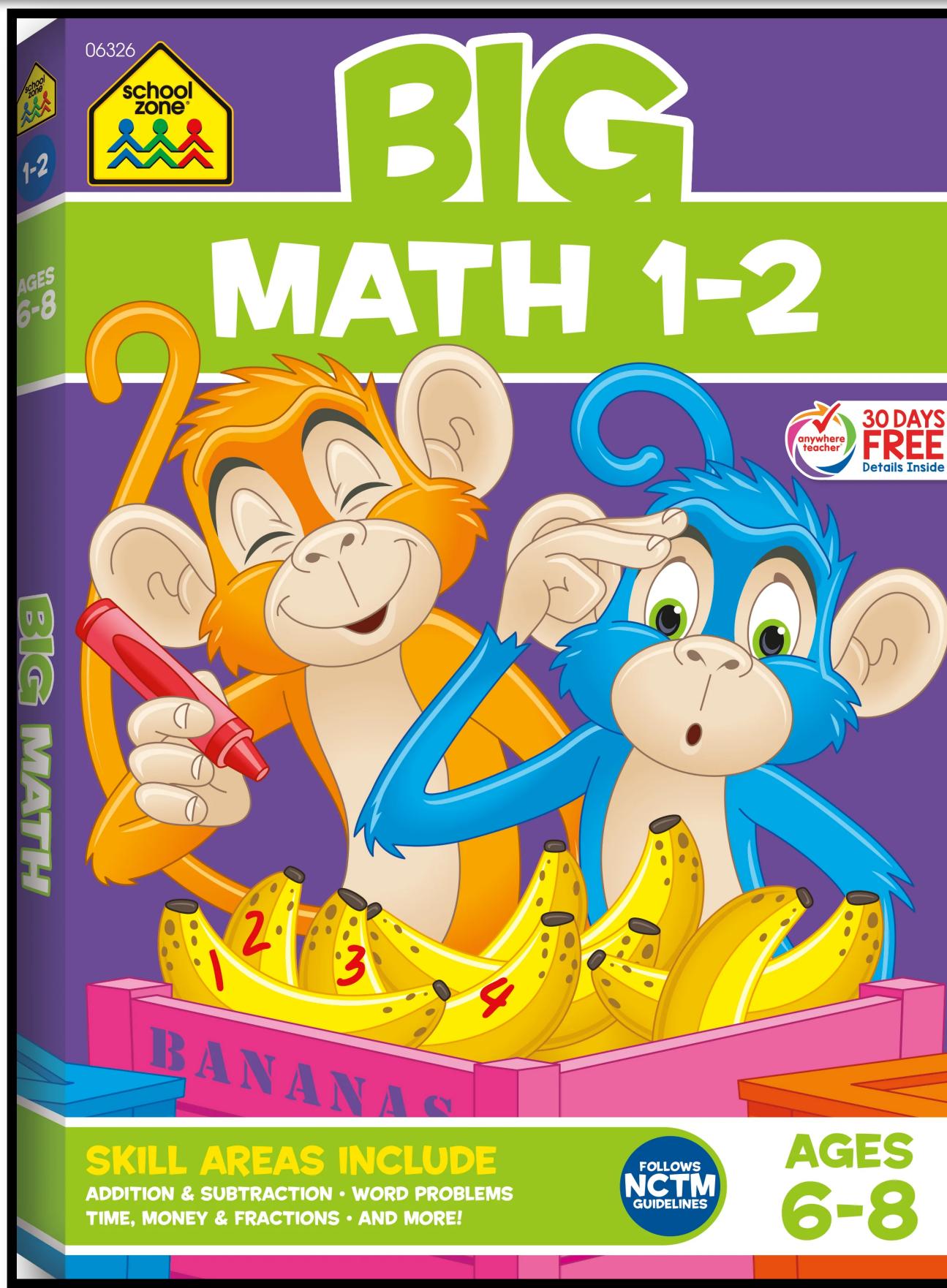
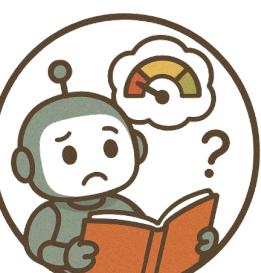


Figure 1: The sources (left) and domains (right) of the BIG-MATH dataset, scaled by size.

# Example Outputs



## 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

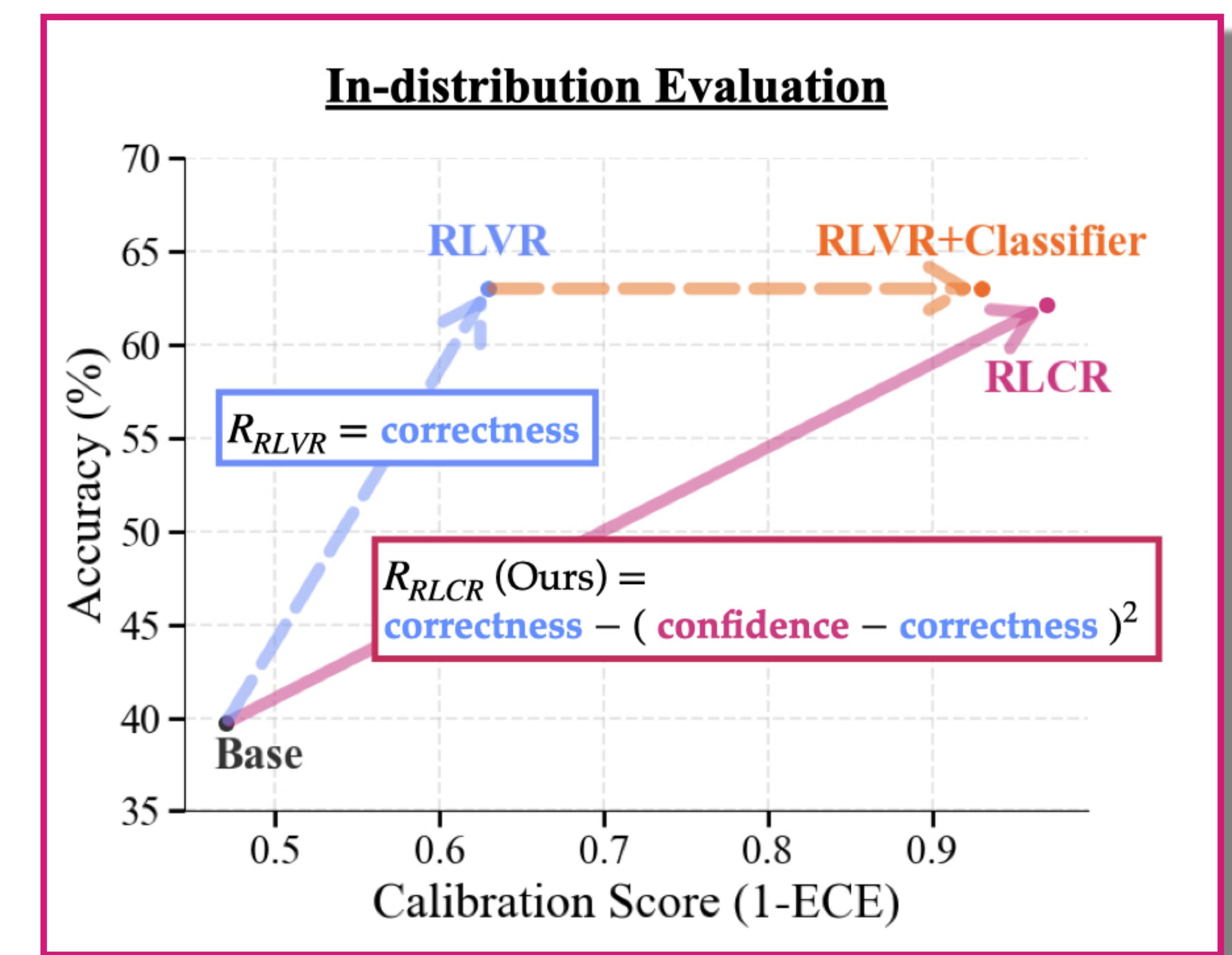
<think> To find out who was awarded the Oceanography Society's Jerlov Award in 2018, I will search for the most recent information on the award winners from the Oceanography Society's official website or a reliable source that covers scientific awards. </think>

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

<confidence> 0.95 </confidence>

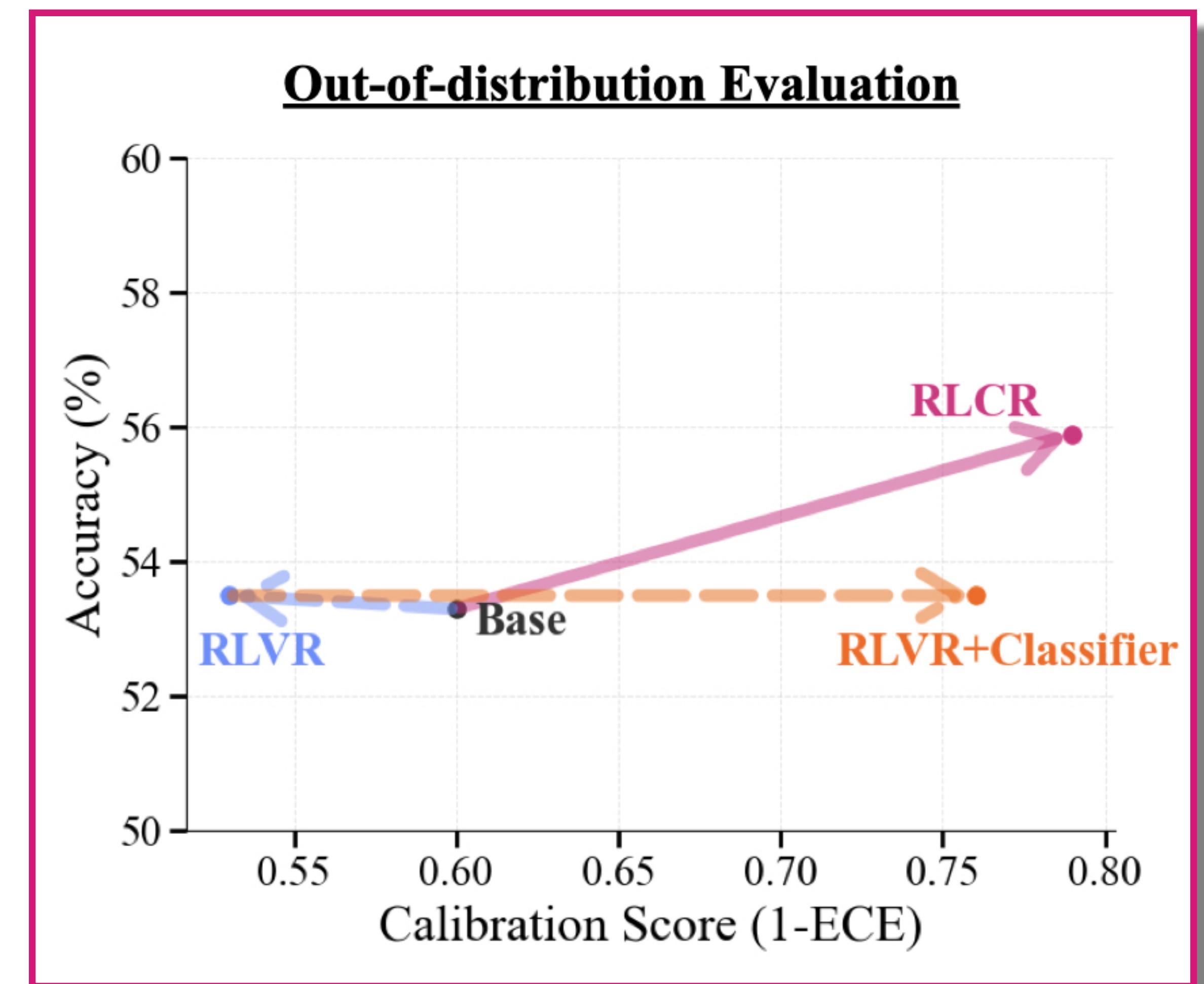
# Results

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



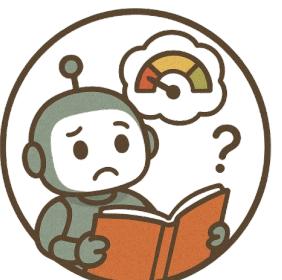
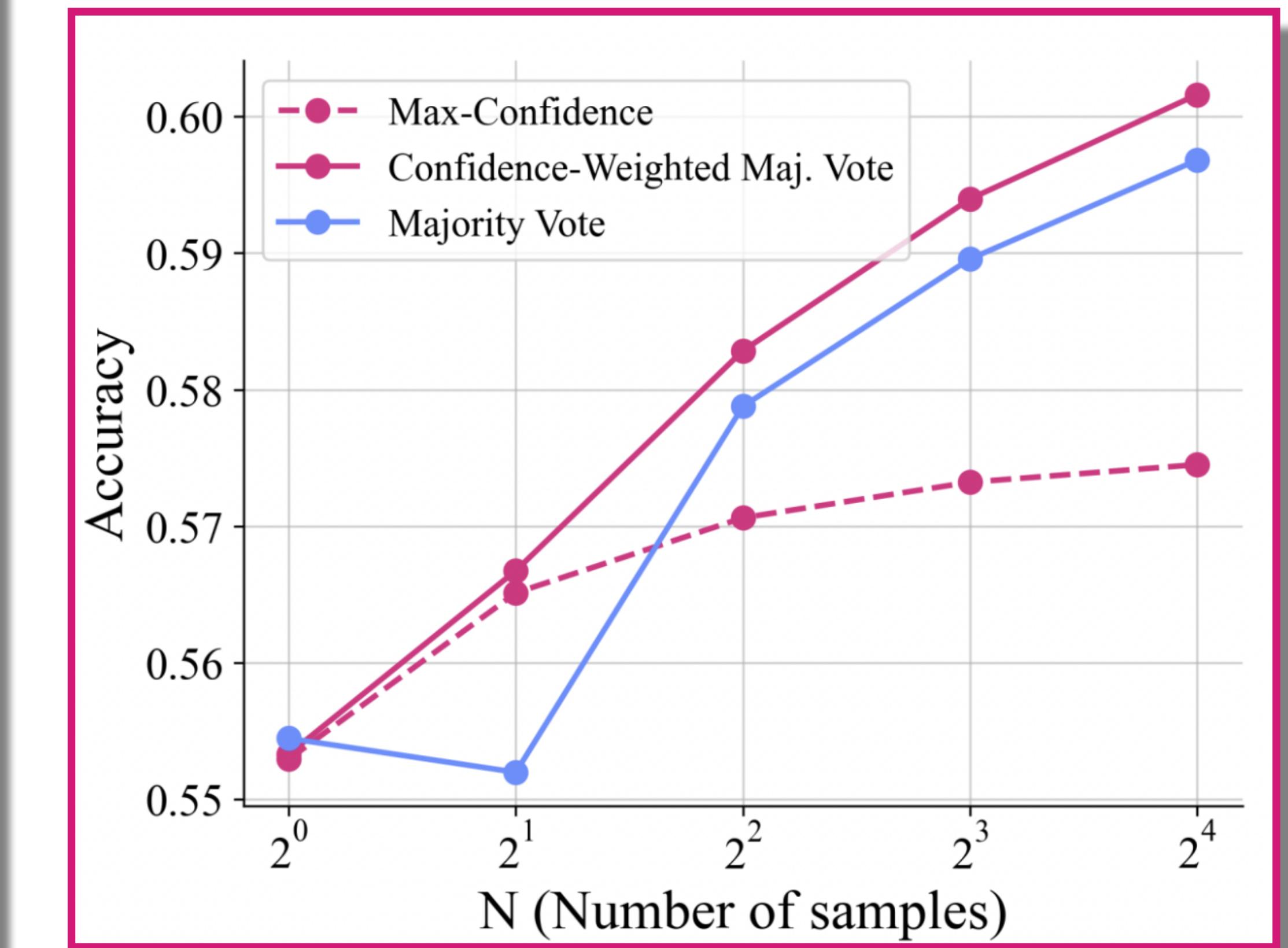
# Results

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



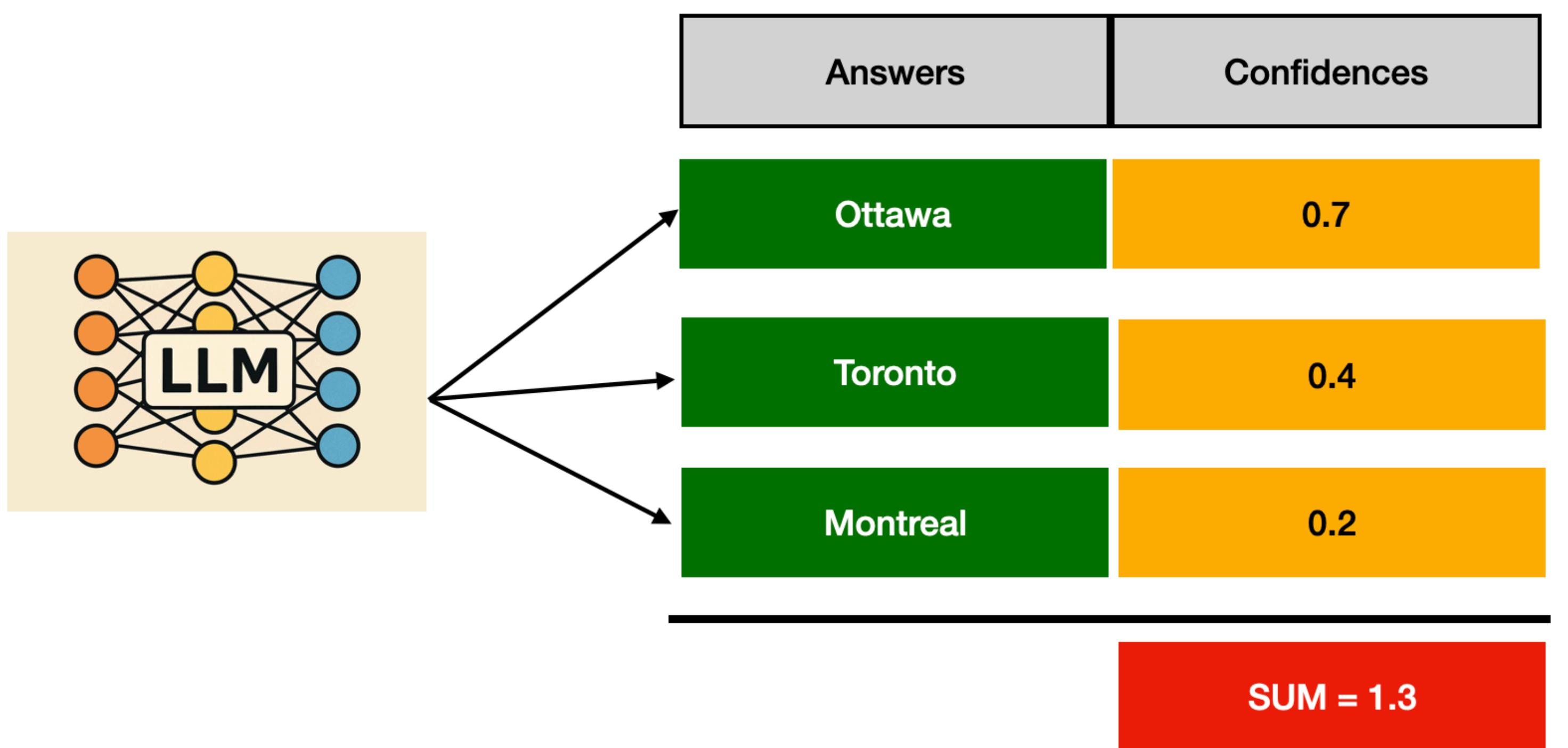
# Using Confidence for Test Time Scaling

- 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!



# Inter Solution Consistency

**“What is the capital of Canada”?**

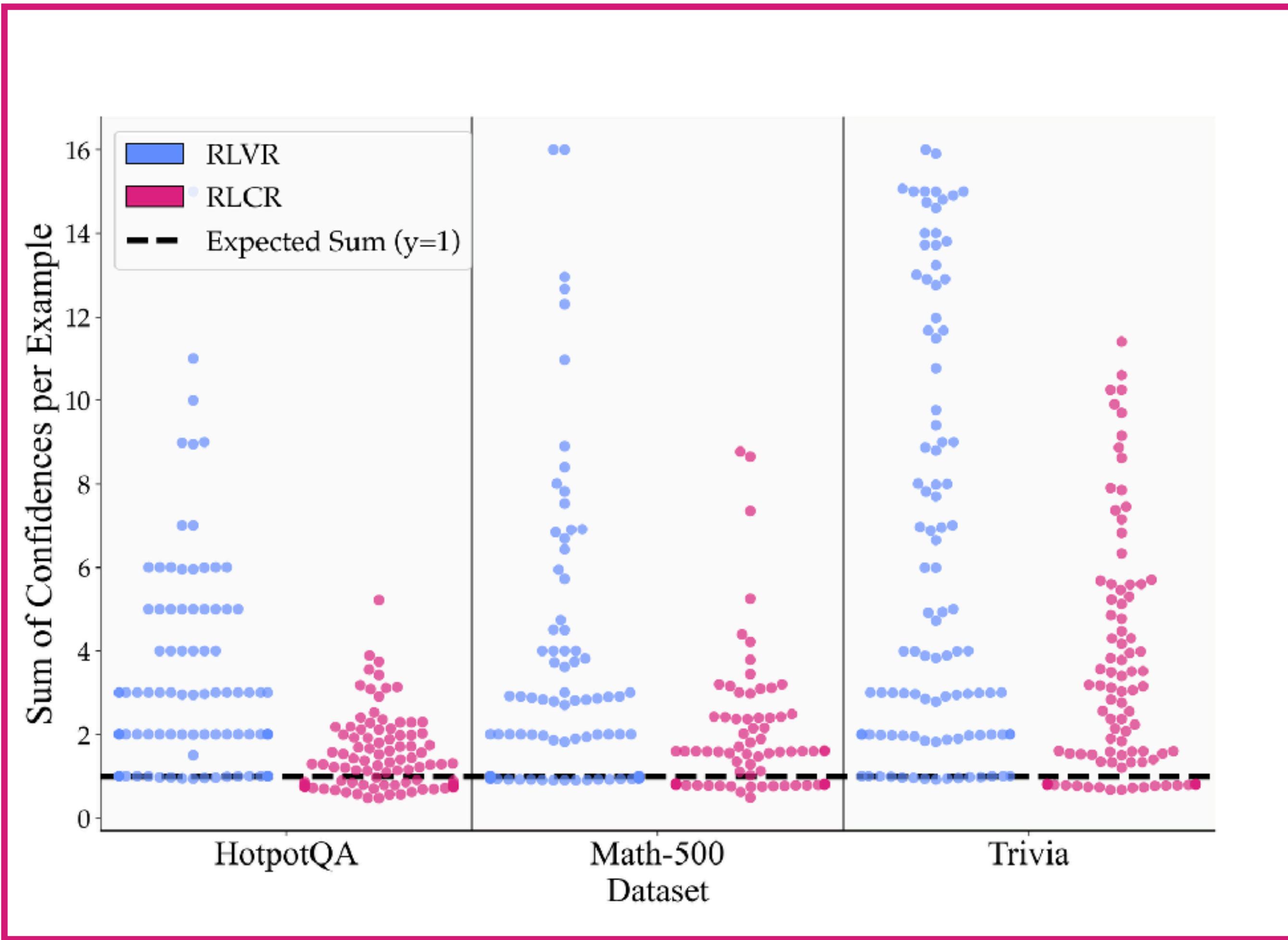


*When given mutually exclusive answers, we want the model to distribute its confidence across distinct answers such that the total confidence is less than or equal to 1 .*

Undesirable

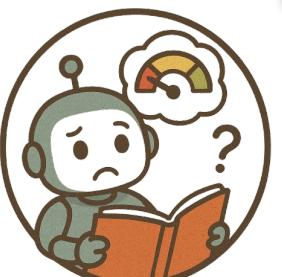


# Inter Solution Consistency



***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.***



# Is Reasoning about Uncertainty useful?

## 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>

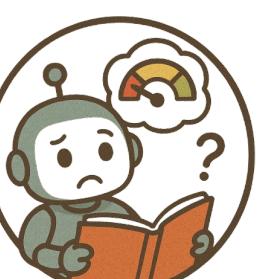
<answer> Dr. David W. Schiedek </answer>

<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>

**Reason about uncertainty**

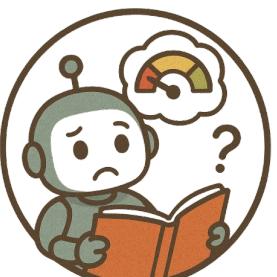
**Verbalize confidence**



# Is Reasoning about Uncertainty useful?

We trained two types of classifiers to output confidence:

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

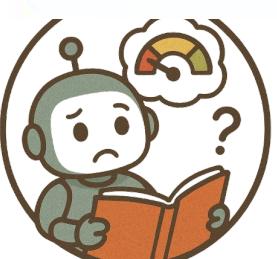


# Is Reasoning about Uncertainty useful?

**<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>**

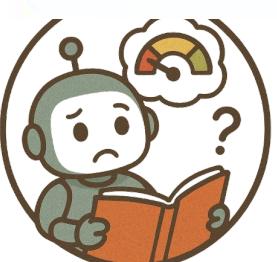
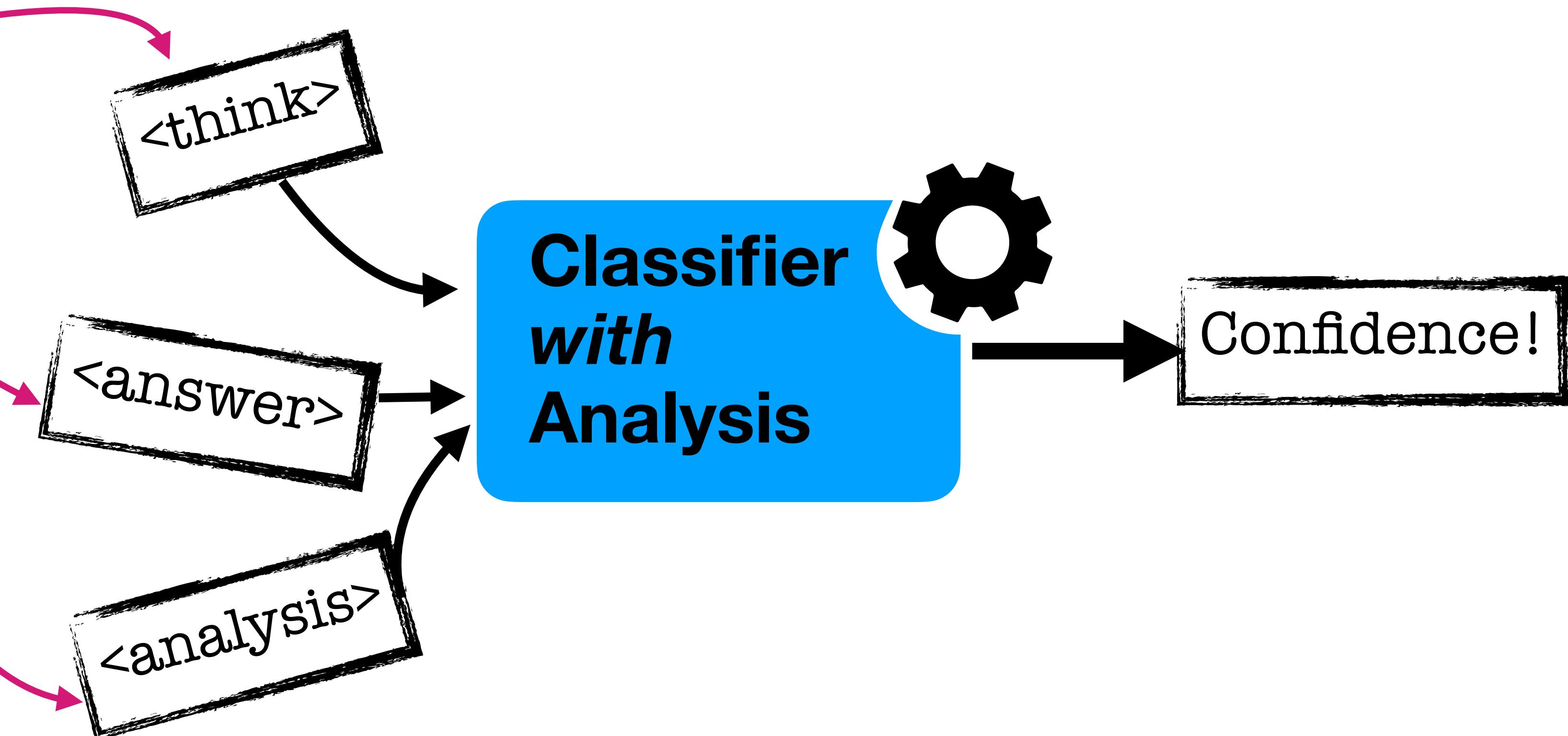


# Is Reasoning about Uncertainty useful?

**<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>**

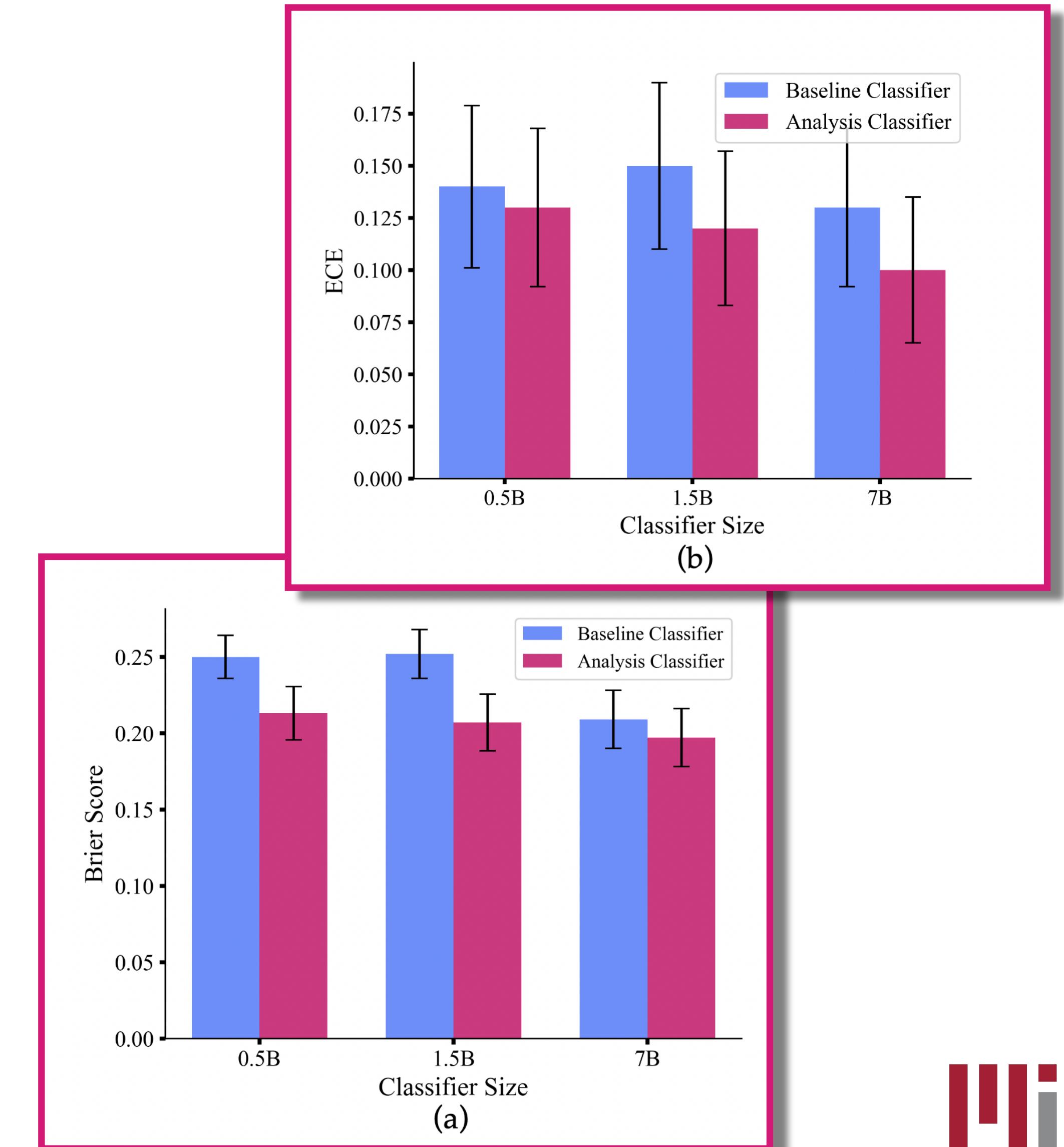


# Is Reasoning about Uncertainty useful?

We trained two types of classifiers:

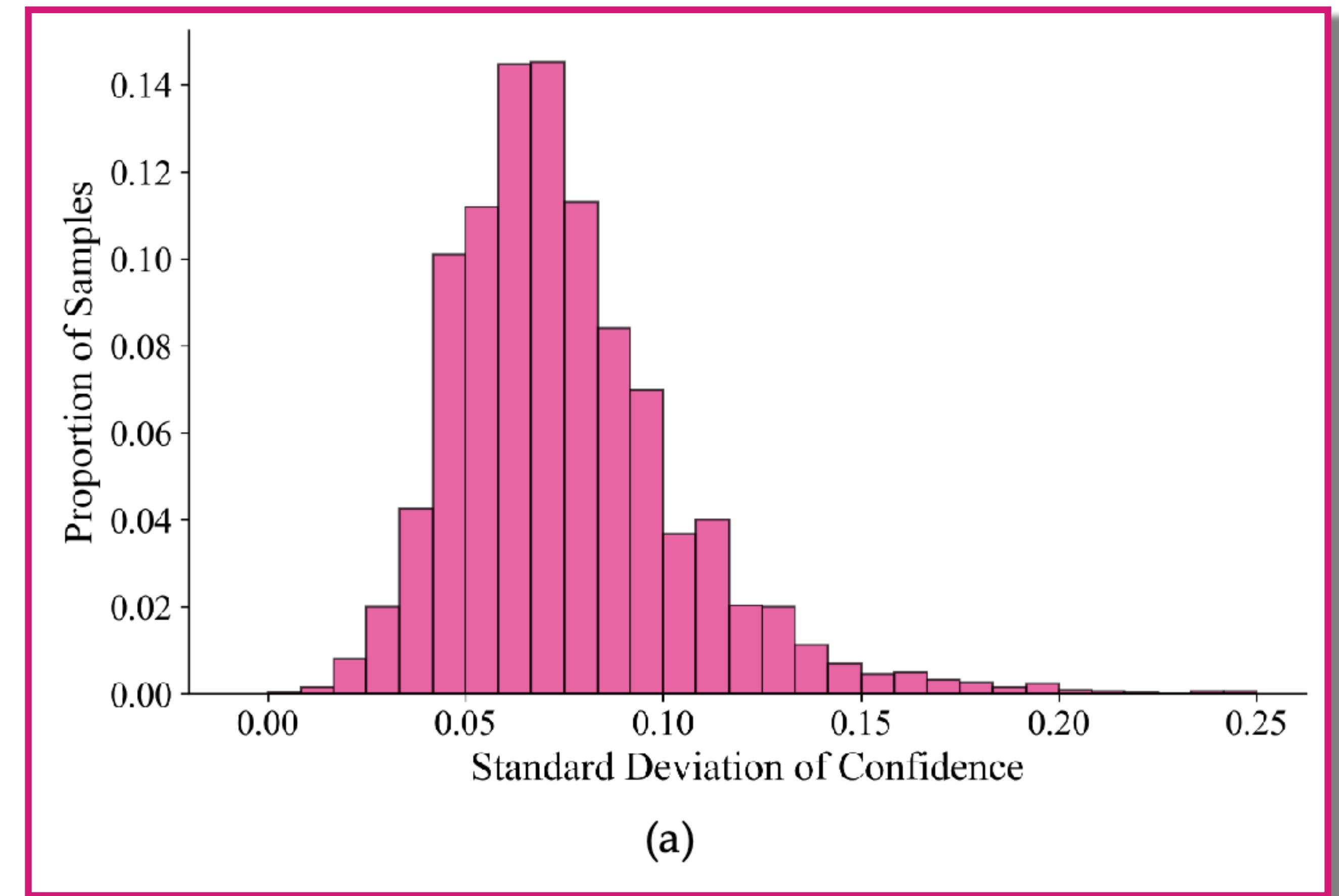
1. **Baseline: Trained on RLVR CoTs**  
<think> <ans>
2. **Analysis: Trained on RLCR CoTs, but without <confidence> values.**  
<think> <ans> <analysis>

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



# 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



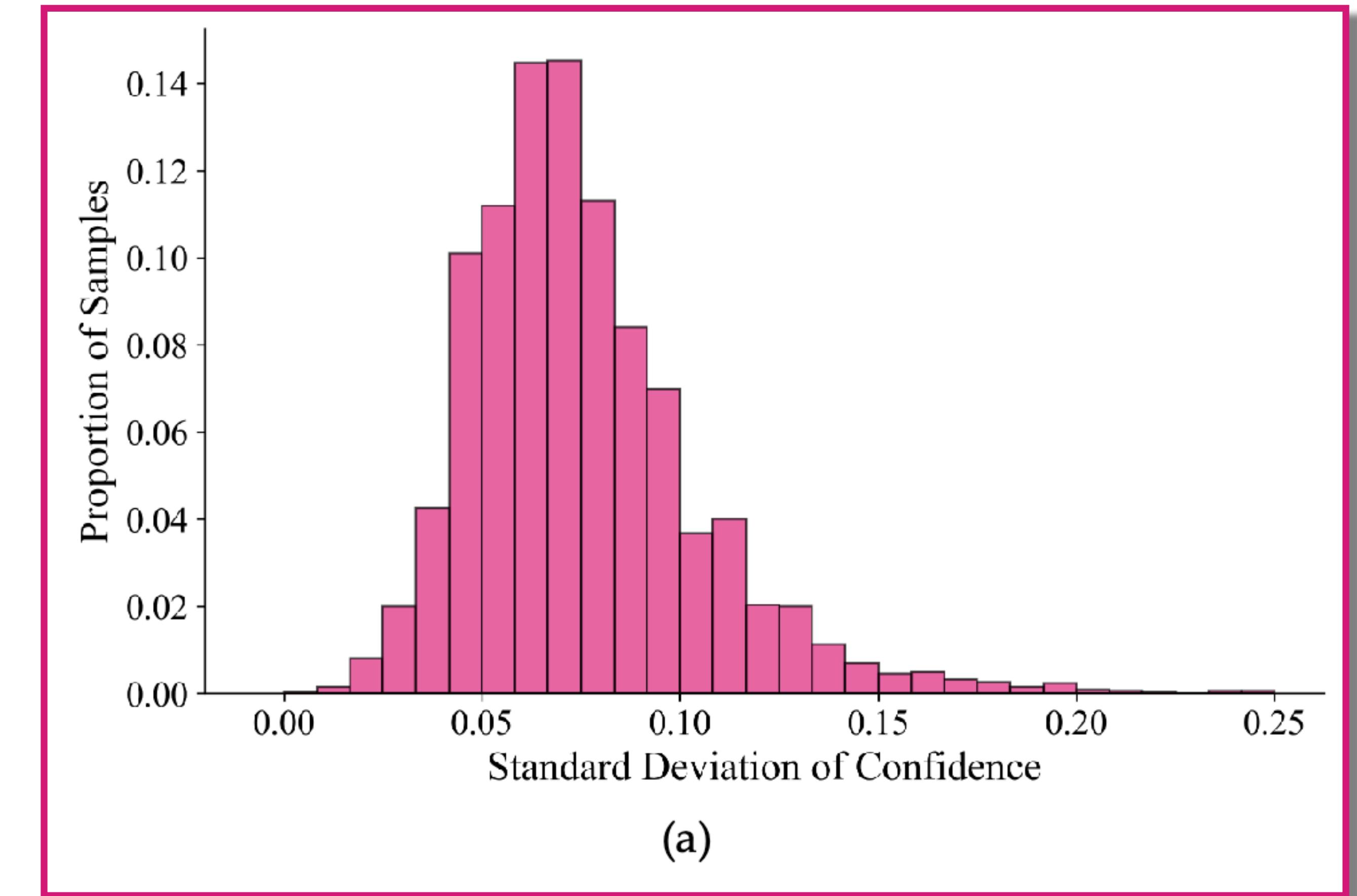
# Uncertainty about Uncertainty

## **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.



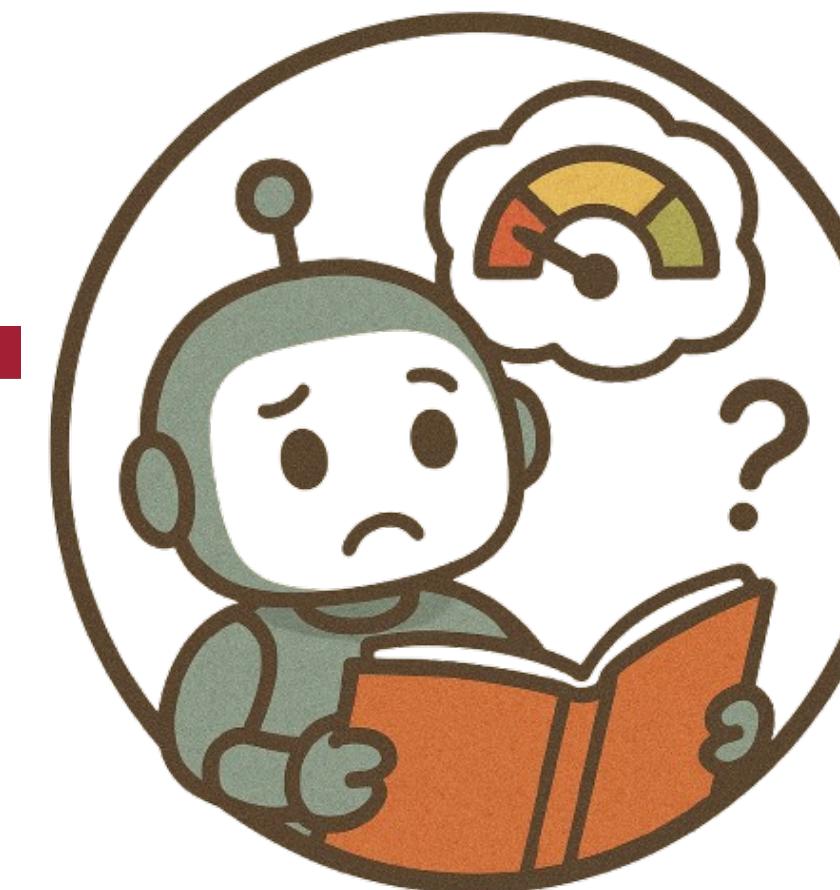
# RLCR:

- 💡 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](https://rl-calibration.github.io)



Mehul Damani\*



Isha Puri\*



Stewart Slocum



Idan Shenfeld



Leshem Chosen



Yoon Kim



Jacob Andreas

\*Equal Contribution