Reasoning

[most slides from Haw-Shiuan Chang]

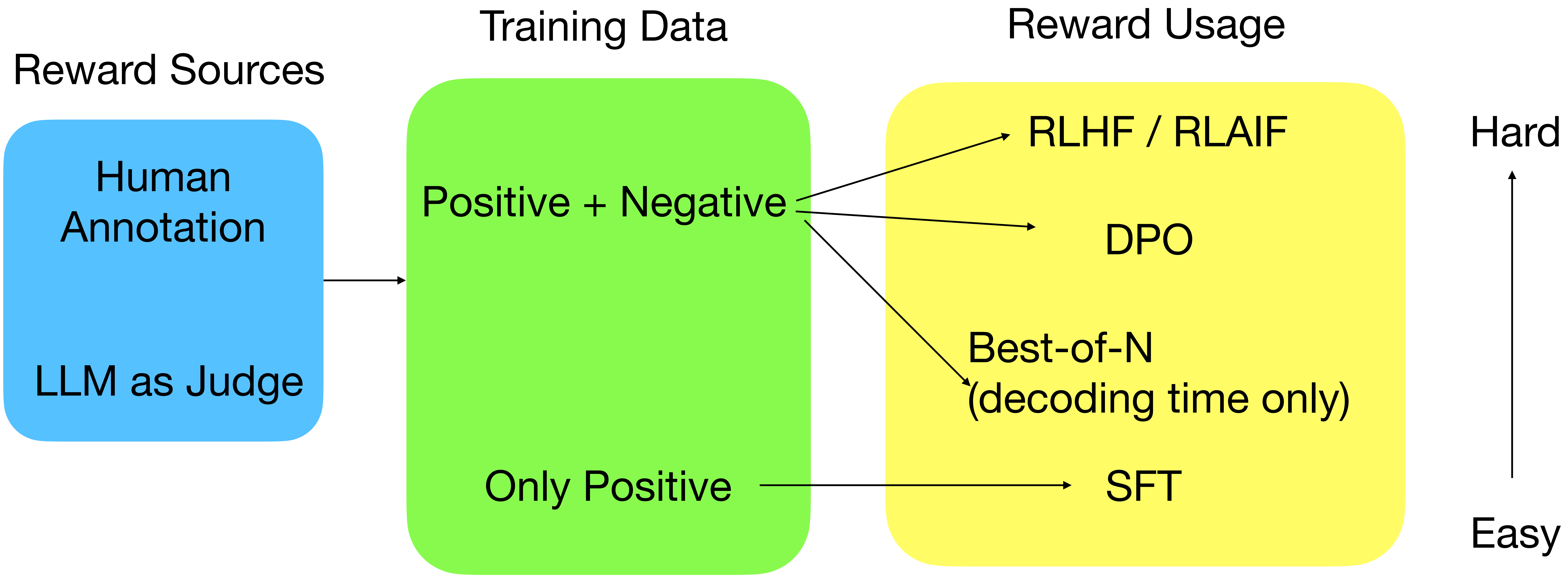
CS 685, Fall 2025
Advanced Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

- Marisa: Tuesdays 10–11 AM, in room CS207, Cube 1
- Rohan: Wednesdays 2:30–3:30 PM, in room CS207, Cube 1
- Brendan: Thursdays 9:30-10:30 AM, in office CS238.
- **Nguyen: Fridays 1-2 PM, in room CS207, Cube 4**

Review

Alignment Methods



Reward Sources

Training Data

Reward Usage

Human Annotation

LLM as Judge

Positive + Negative

Only Positive

RLHF / RLAIF

DPO

Best-of-N (decoding time only)

SFT

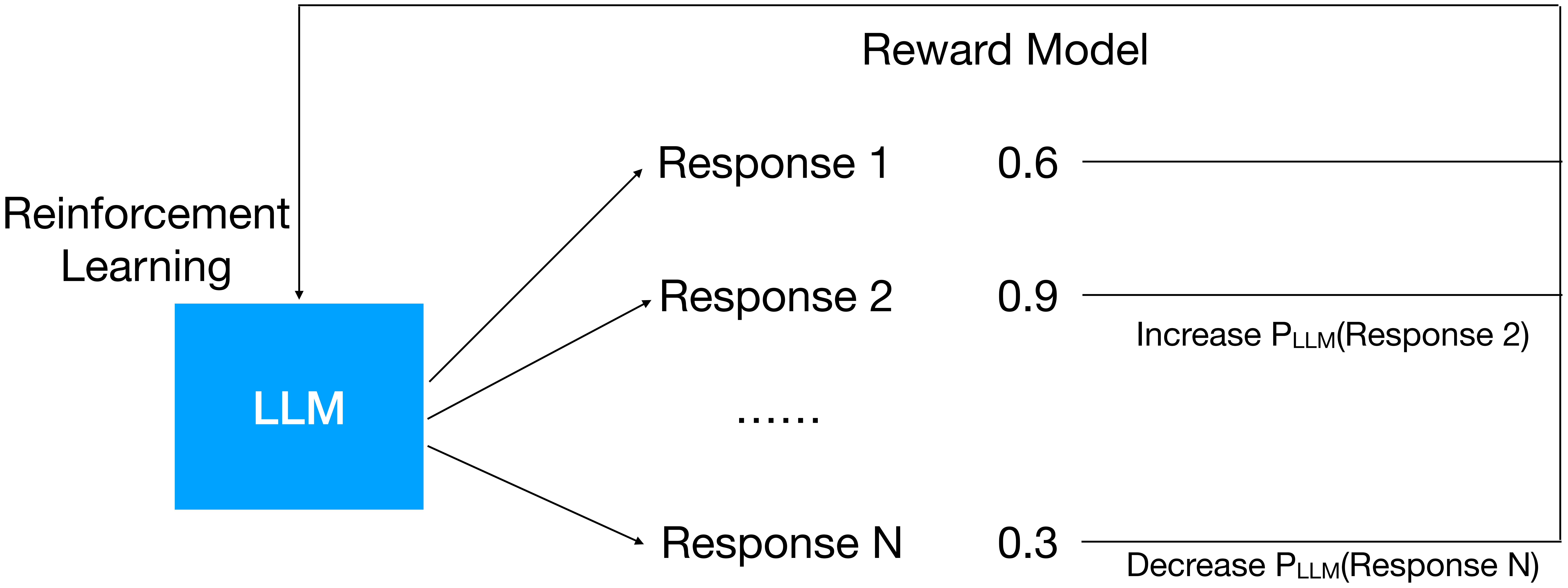
Hard

Easy

Best of N

- The simplest alignment baseline
- Uses separate, pretrained LM and reward functions
- At decoding time: take N samples from the base model. Choose highest-reward output.

RLHF



Reward Model

Reinforcement Learning

LLM

Response 1 0.6

Response 2 0.9

Increase P_{LLM}(Response 2)

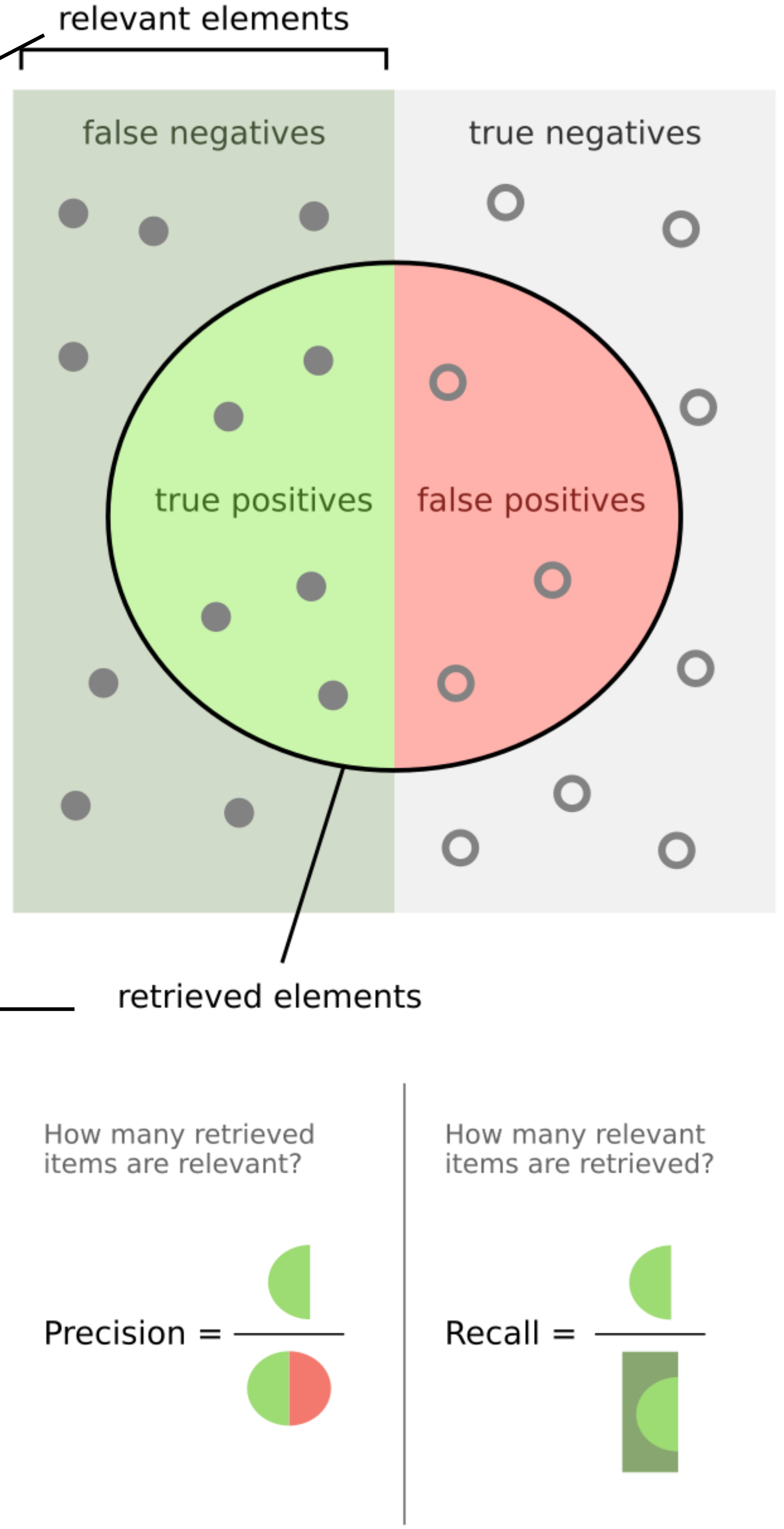
……

Response N 0.3

Decrease P_{LLM}(Response N)

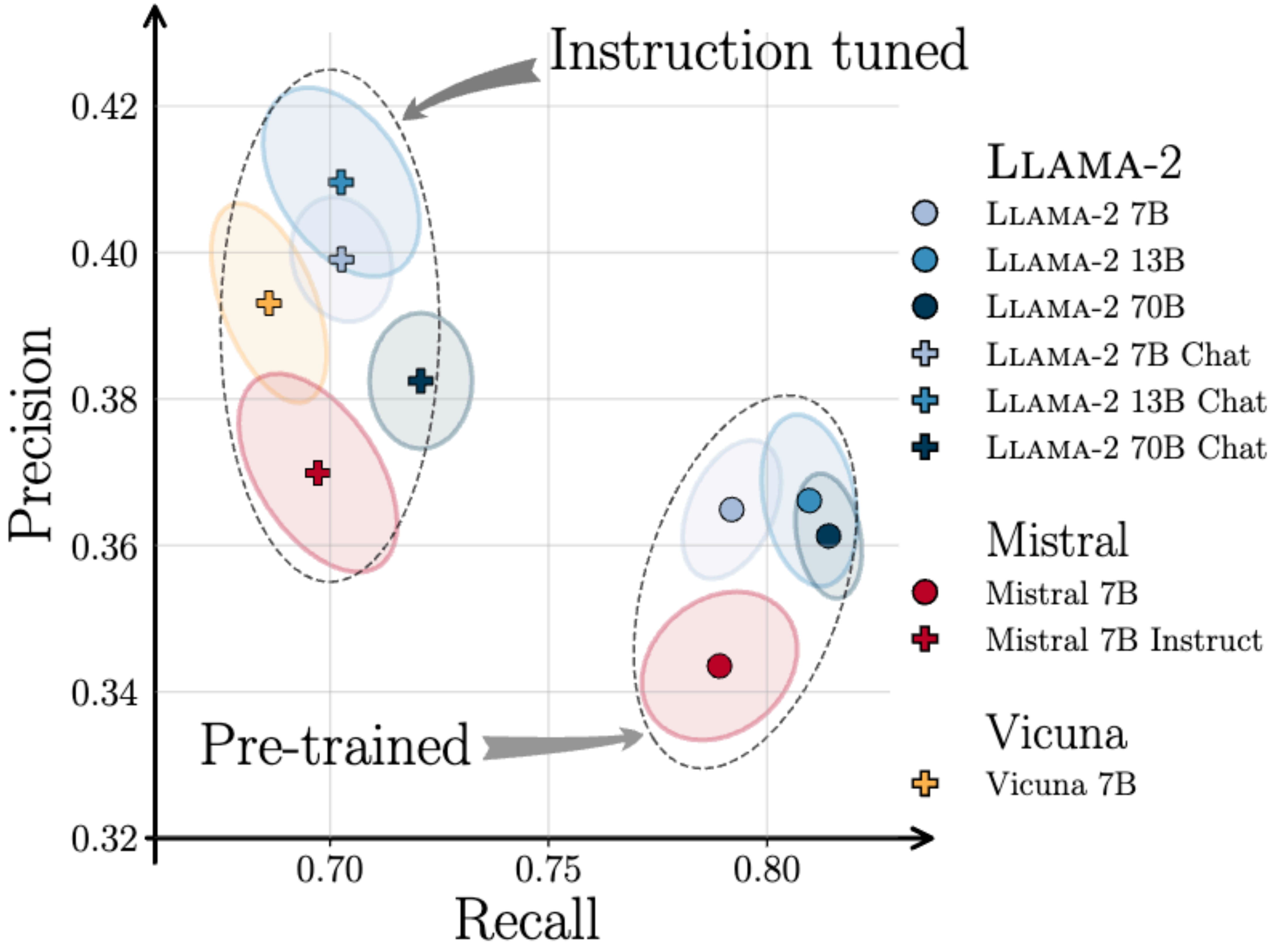
RLHF Decreases Output Diversity

All the correct answers



Predicted answers

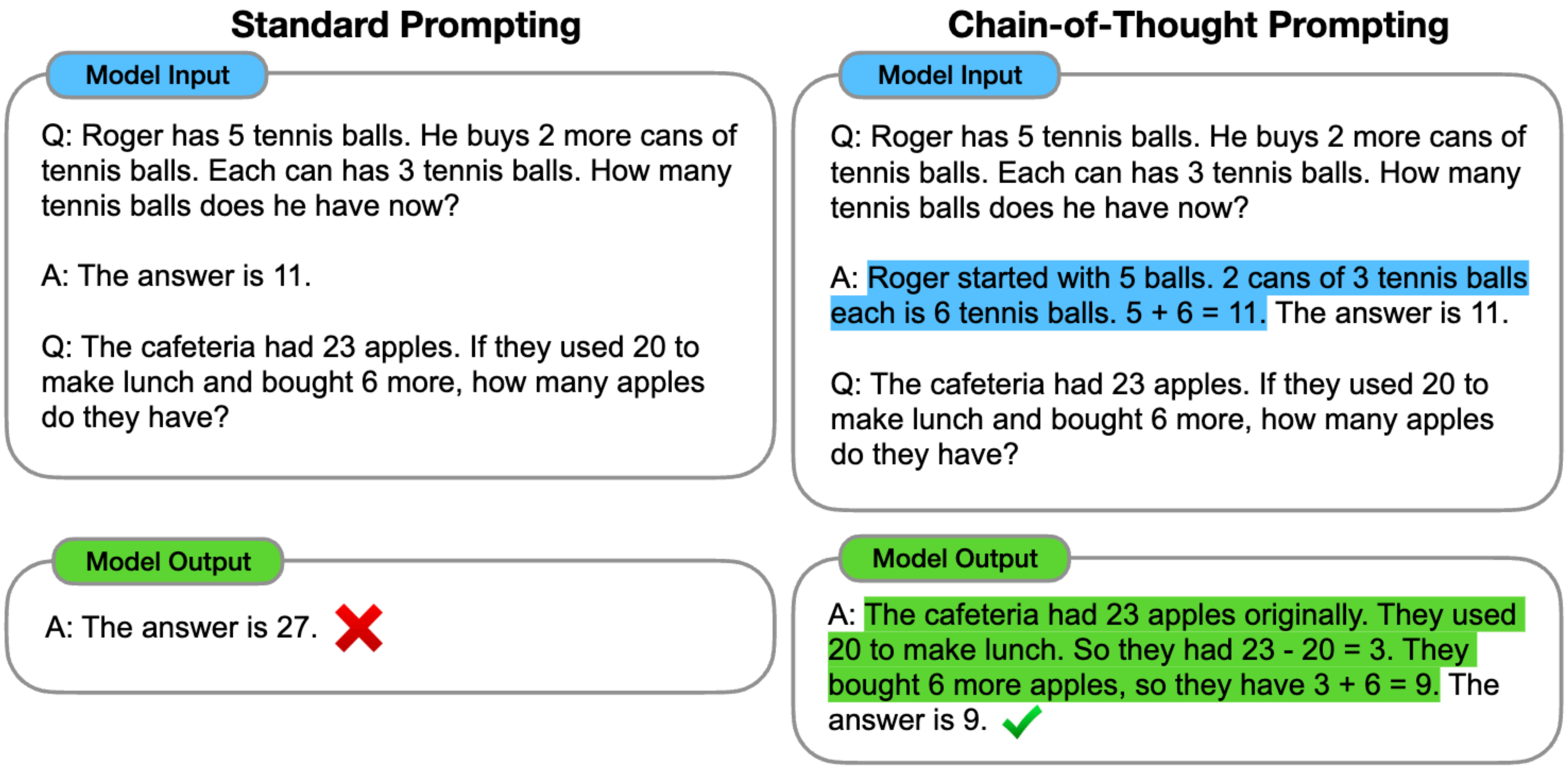
https://en.wikipedia.org/wiki/Precision_and_recall



Exploring Precision and Recall to assess the quality and diversity of LLMs (Le Bronnec et al., ACL 2024; https://arxiv.org/pdf/2402.10693)

Large Reasoning Model (LRM)

Chain of Thought



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

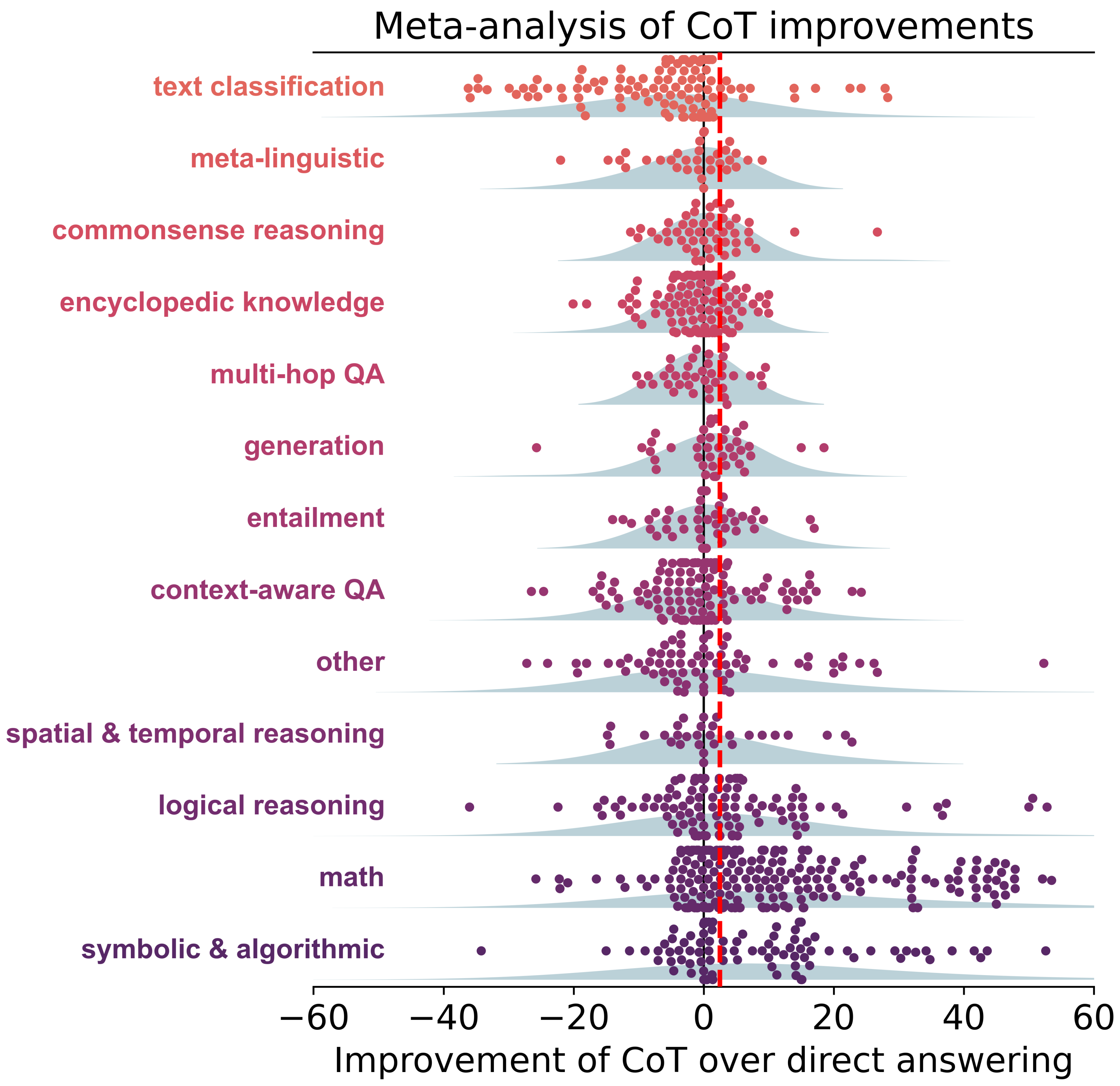
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

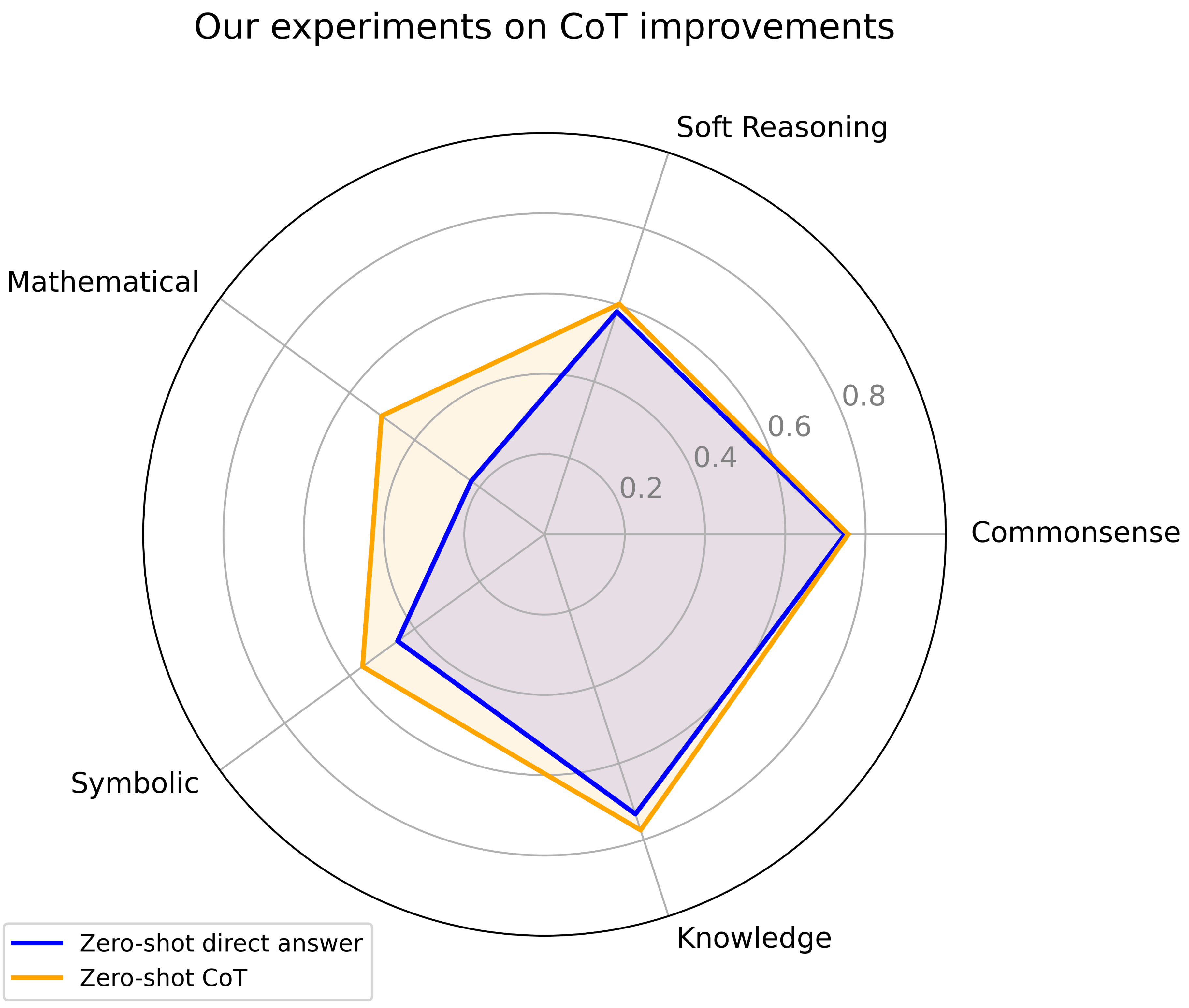
Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., NeurIPS 2022; https://arxiv.org/pdf/2201.11903)

Chain of Thought

- Meta-analysis of 100 papers using CoT
  [Left: diff. in performance for one (LLM, task) pair]


Meta-analysis of CoT improvements


Our experiments on CoT improvements

Sprague et al., ICLR 2025, "To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning."

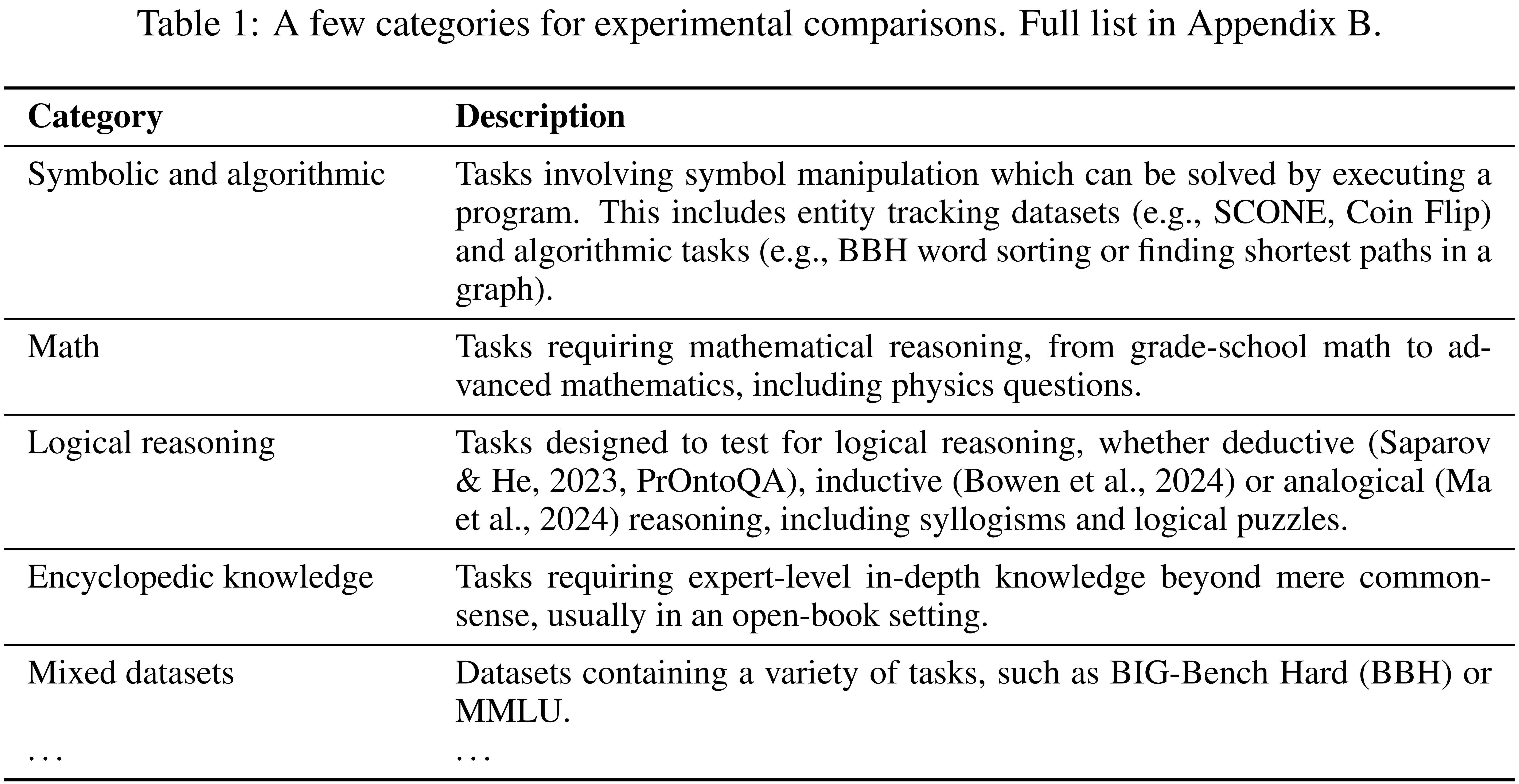


Table 1: A few categories for experimental comparisons. Full list in Appendix B.

Category	**Description**
Symbolic and algorithmic	Tasks involving symbol manipulation which can be solved by executing a program. This includes entity tracking datasets (e.g., SCONE, Coin Flip) and algorithmic tasks (e.g., BBH word sorting or finding shortest paths in a graph).
Math	Tasks requiring mathematical reasoning, from grade-school math to advanced mathematics, including physics questions.
Logical reasoning	Tasks designed to test for logical reasoning, whether deductive (Saparov & He, 2023, PrOntoQA), inductive (Bowen et al., 2024) or analogical (Ma et al., 2024) reasoning, including syllogisms and logical puzzles.
Encyclopedic knowledge	Tasks requiring expert-level in-depth knowledge beyond mere commonsense, usually in an open-book setting.
Mixed datasets	Datasets containing a variety of tasks, such as BIG-Bench Hard (BBH) or MMLU.
…	…

Sprague et al., ICLR 2025, "To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning."

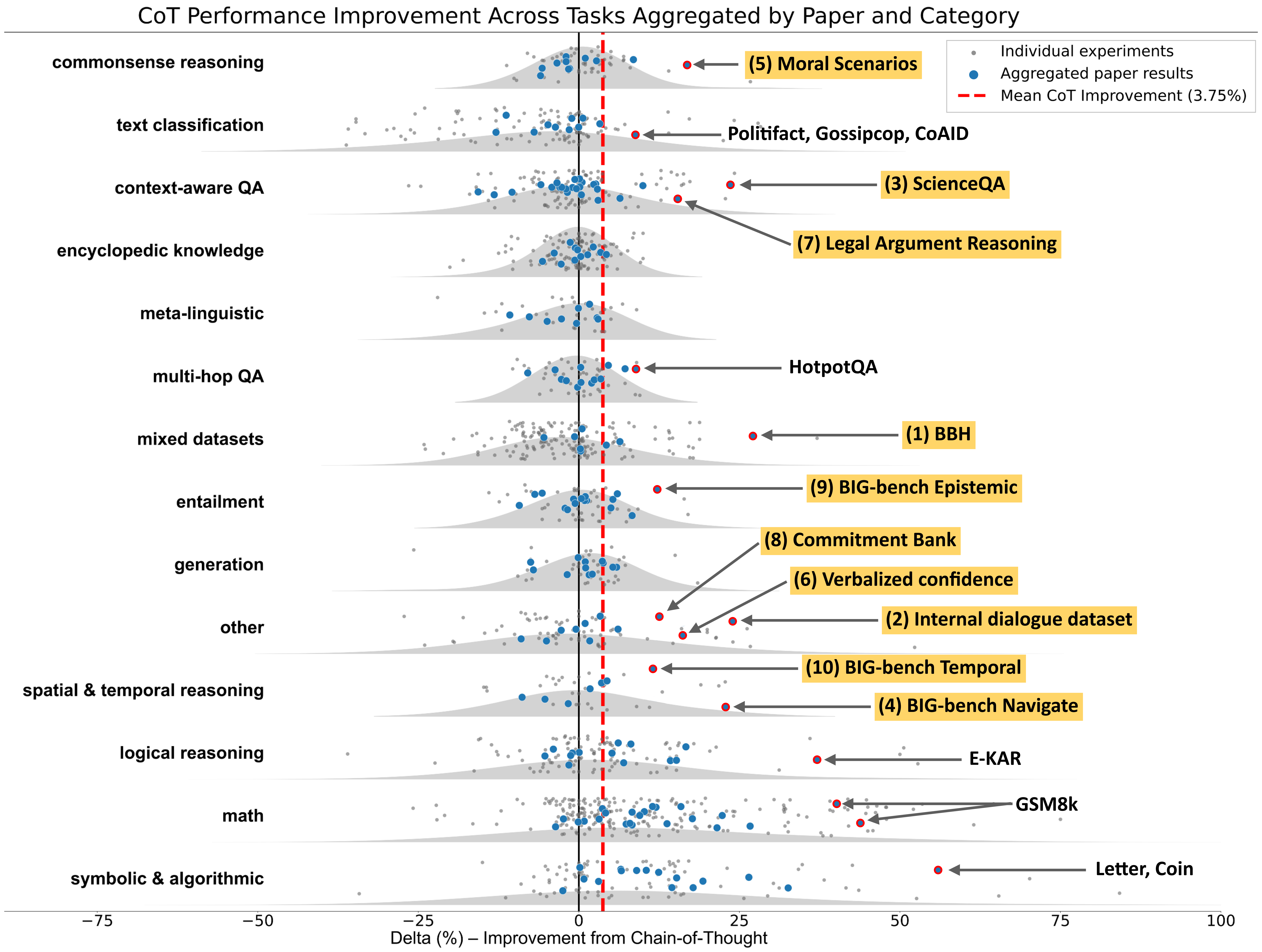
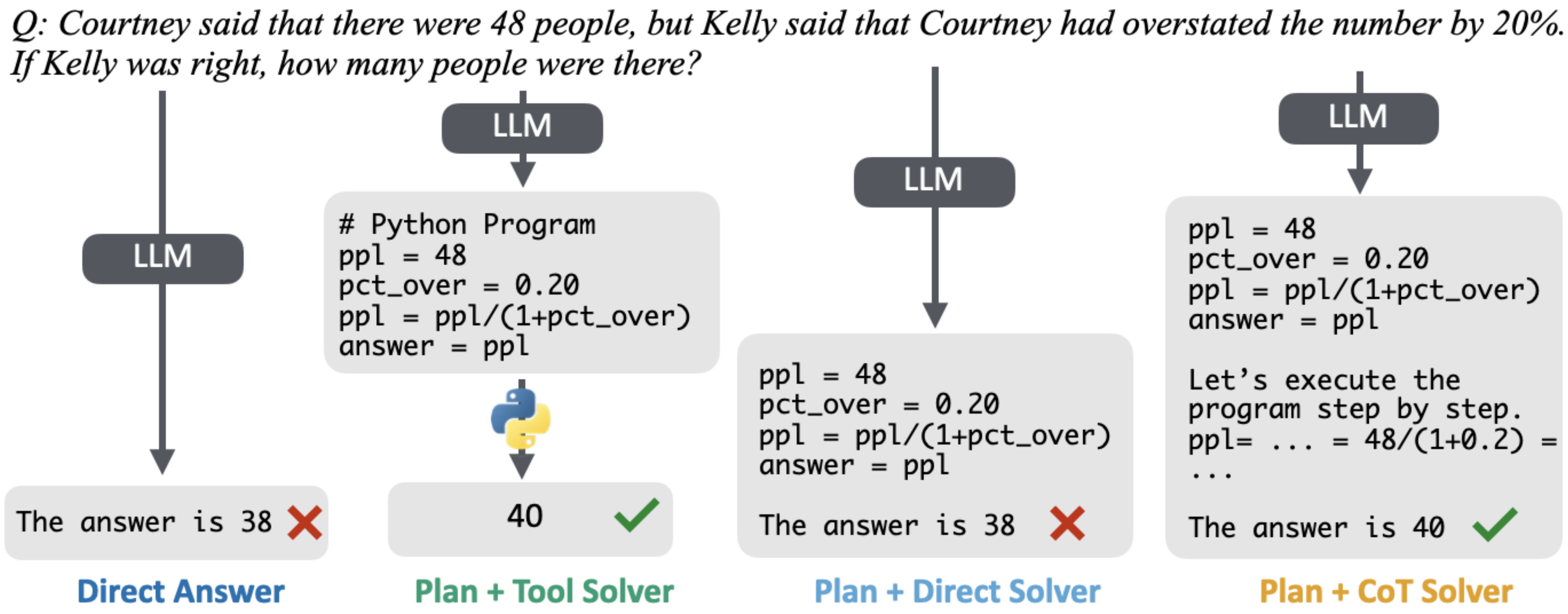


Figure 2: Results from our meta-analysis (grey dots) aggregated by paper and category (blue dots).

Sprague et al., ICLR 2025, "To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning."



Q: Courtney said that there were 48 people, but Kelly said that Courtney had overstated the number by 20%. If Kelly was right, how many people were there?

```
# Python Program
ppl = 48
pct_over = 0.20
ppl = ppl/(1+pct_over)
answer = ppl
```

```
ppl = 48
pct_over = 0.20
ppl = ppl/(1+pct_over)
answer = ppl

The answer is 38
```

```
ppl = 48
pct_over = 0.20
ppl = ppl/(1+pct_over)
answer = ppl

Let's execute the
program step by step.
ppl= ... = 48/(1+0.2) =
...

The answer is 40
```

Figure 5: Prompt variants that separate planning and execution for GSM8K. For all prompt variants besides direct answer and CoT (not shown), we few-shot prompt an LLM to first generate a Python program as a solution plan. For Plan + Direct Solver, the LLM is prompted to directly give an answer from the plan; for Plan + CoT Solver, the LLM is prompted to solve the plan step-by-step with CoT and give an answer; for Plan + Tool Solver, we feed the plan into a Python interpreter.

- Demonstrating planning vs execution in prompts

Sprague et al., ICLR 2025, "To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning."

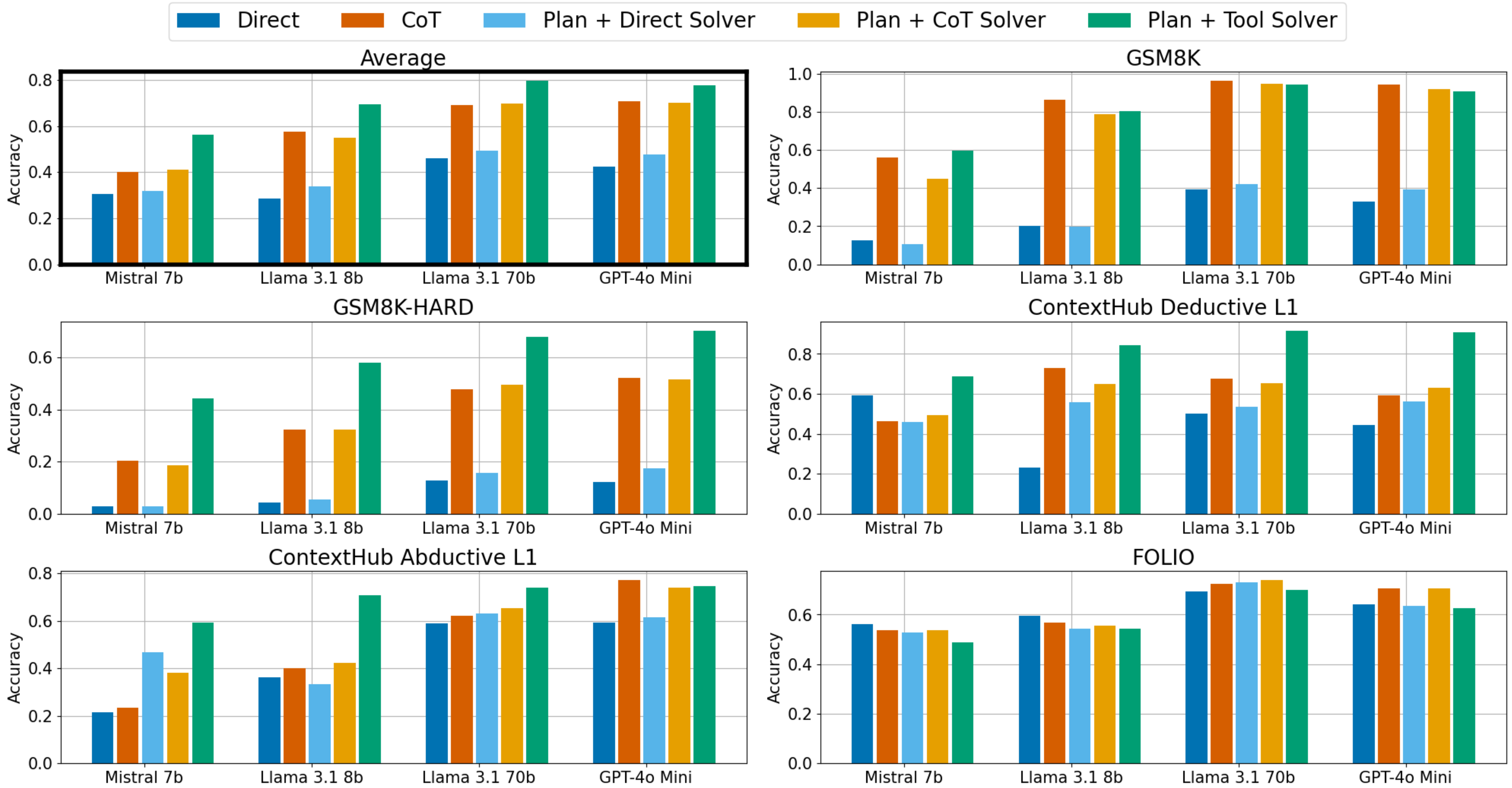


Figure 6: Performance of prompt variants that separate planning and execution for math and logical reasoning datasets. Despite outperforming direct answer for solving a formal plan and deriving the final answer, CoT is still limited in performing symbolic computations: there is a large performance boost from Plan + Tool Solver over CoT and Plan + CoT Solver on average across all models.

Sprague et al., ICLR 2025, "To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning."

DeepSeek R1

How is a state-of-the-art LRM trained?

DeepSeek V3

- A large LLM with 671B
- Using Mixture of Expert Architecture

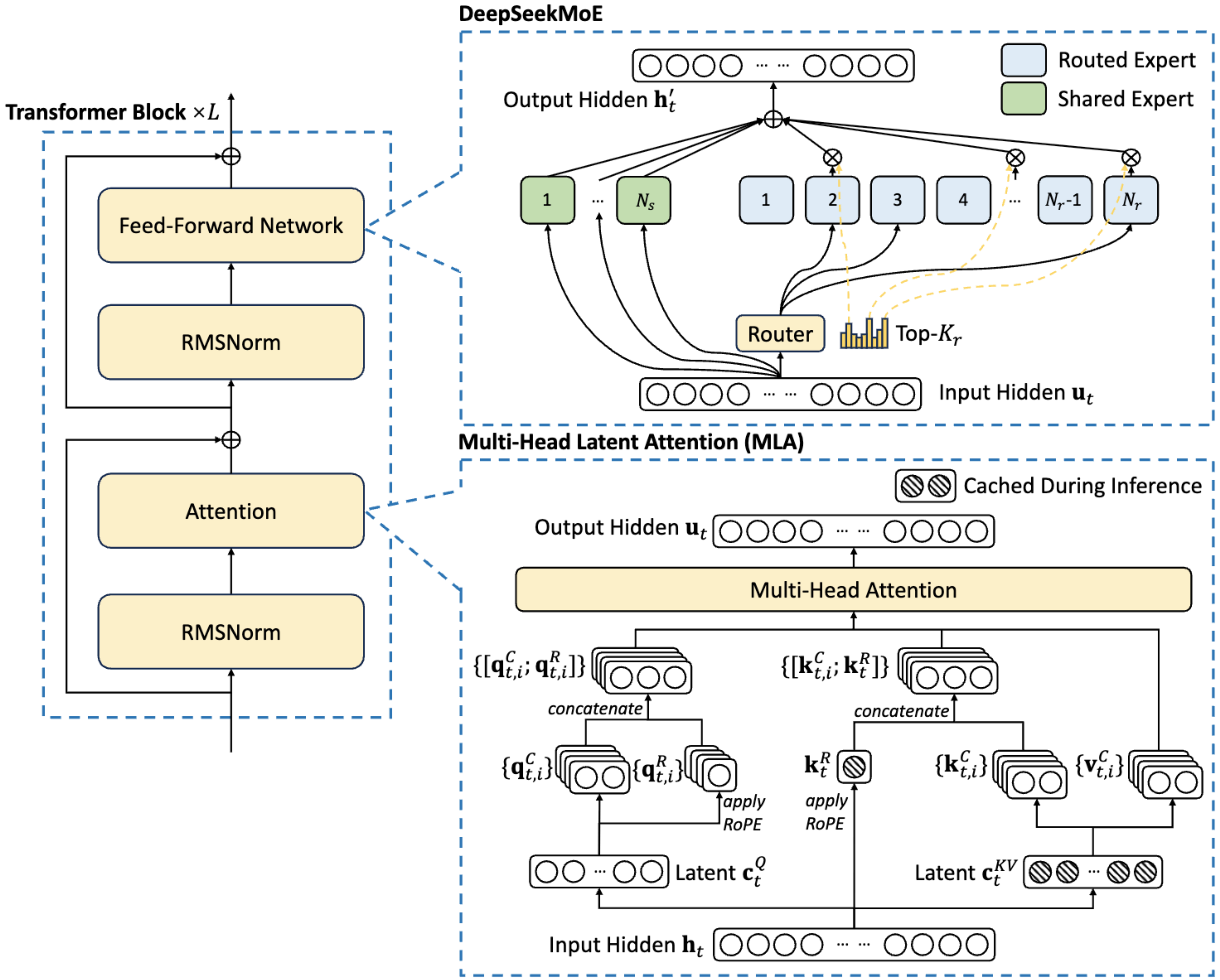
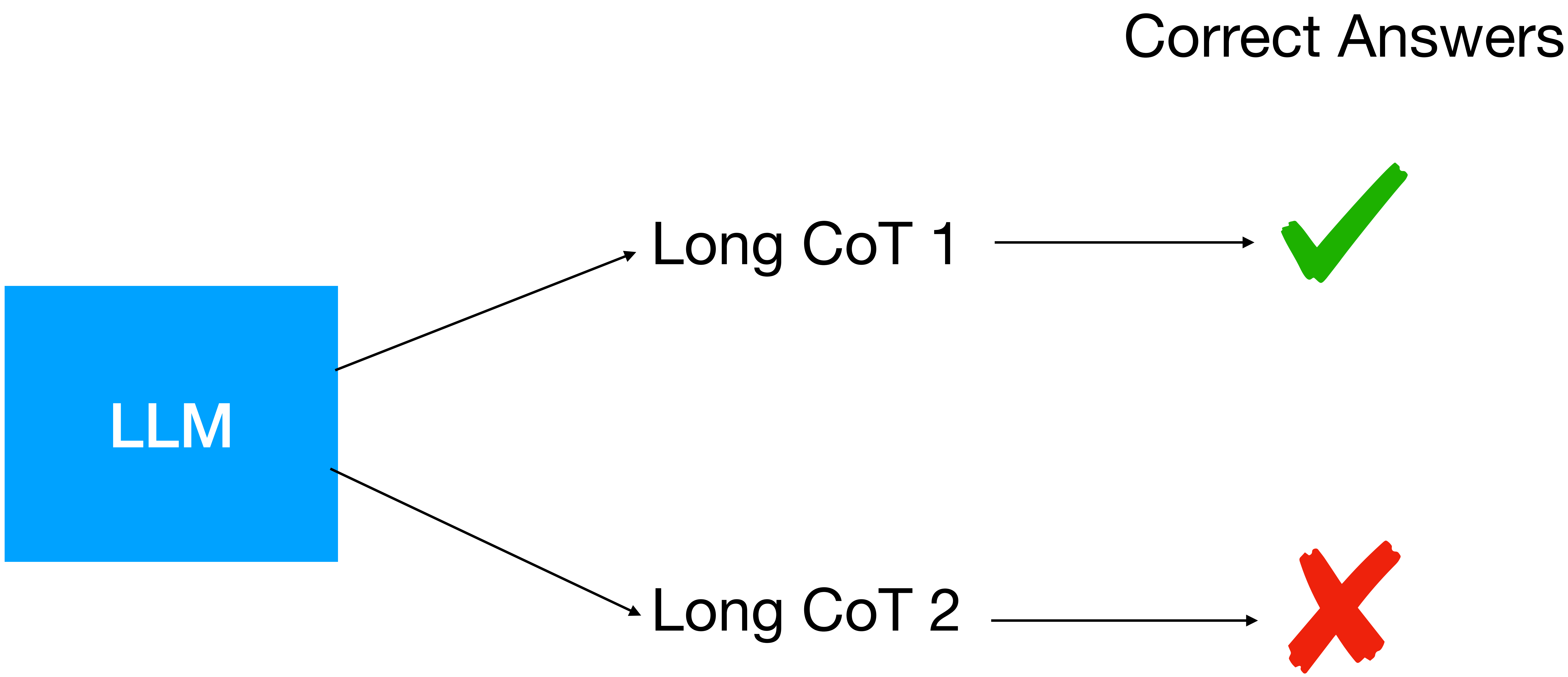


Figure 2 | Illustration of the basic architecture of DeepSeek-V3. Following DeepSeek-V2, we adopt MLA and DeepSeekMoE for efficient inference and economical training.

[Jan 2025, arxiv]

Reasoning



Correct Answers

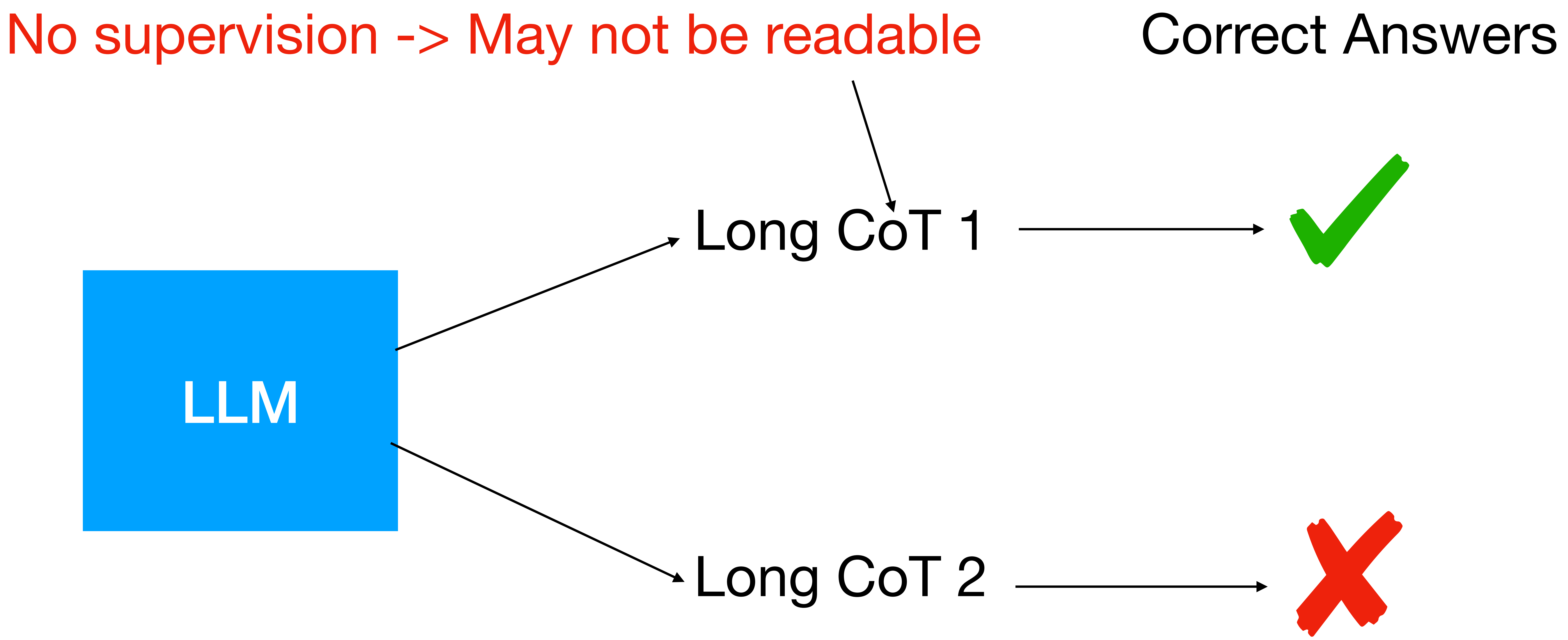
LLM

Long CoT 1 ✓

Long CoT 2 ✗

We Should
Encourage the
LLM to output
more of this

Readability Issue



No supervision -> May not be readable

Correct Answers

LLM

Long CoT 1 ✓

Long CoT 2 ✗

We Should Encourage the LLM to output more of this

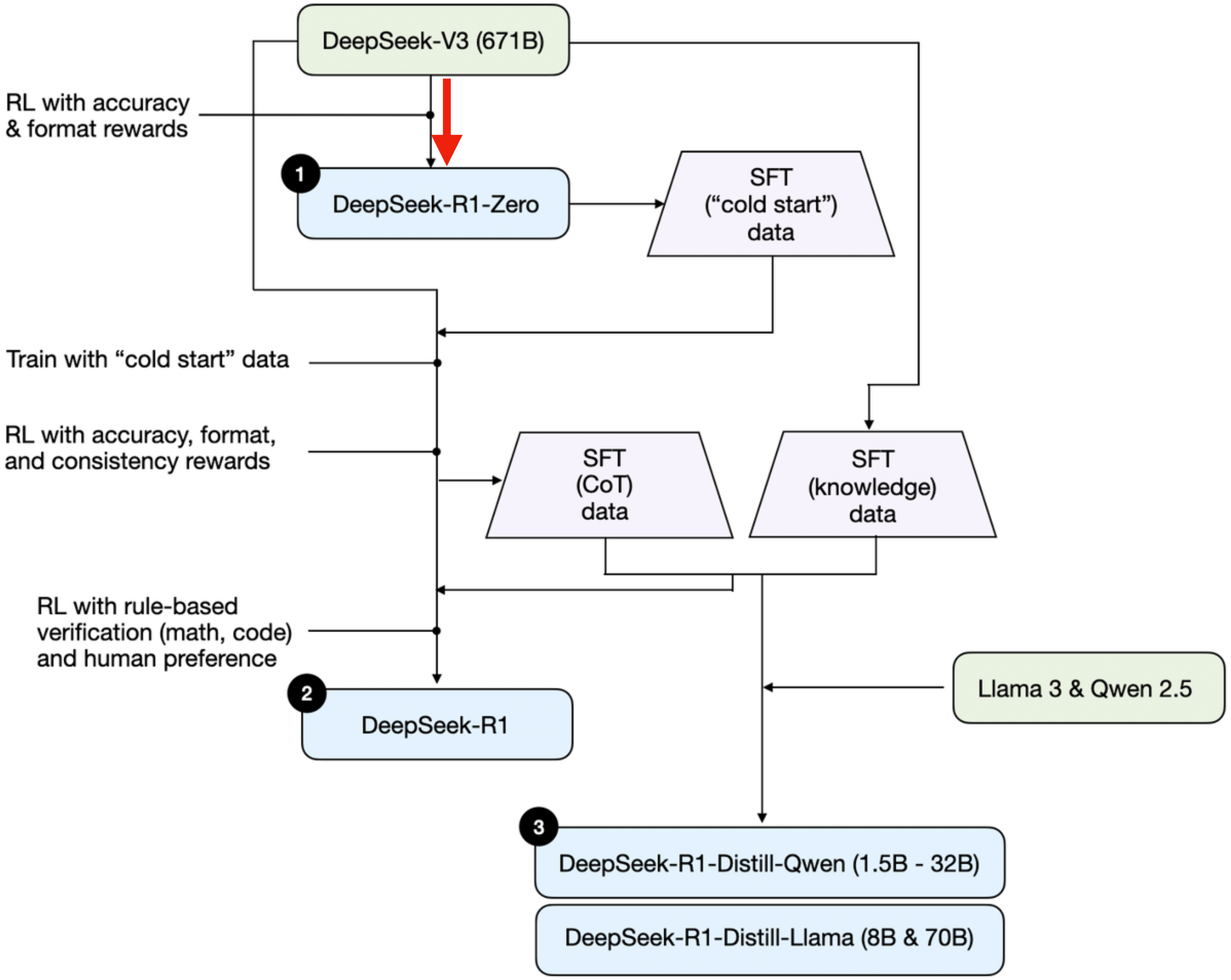
Challenges

- 1. Does not have process supervision/reward
  - CoT after RL is often not readable
  - RL is unstable
  - Expensive because CoT is long and model is large
- 2. LRM also needs to handle the normal queries that do not need reasoning

DeepSeek R1 Training Pipeline

- Accuracy reward:
  - Leetcode compiler
  - Rule-based answer checking
- Format reward:
  - LLM to judge if the reasoning is inside the <think> tag

https://www.linkedin.com/pulse/understanding-reasoning-llms-sebastian-raschka-phd-1tshc/?trackingId=L4cJD57IRs2Pl2nUoLs%2FLw%3D%3D



Pure Reinforcement Learning

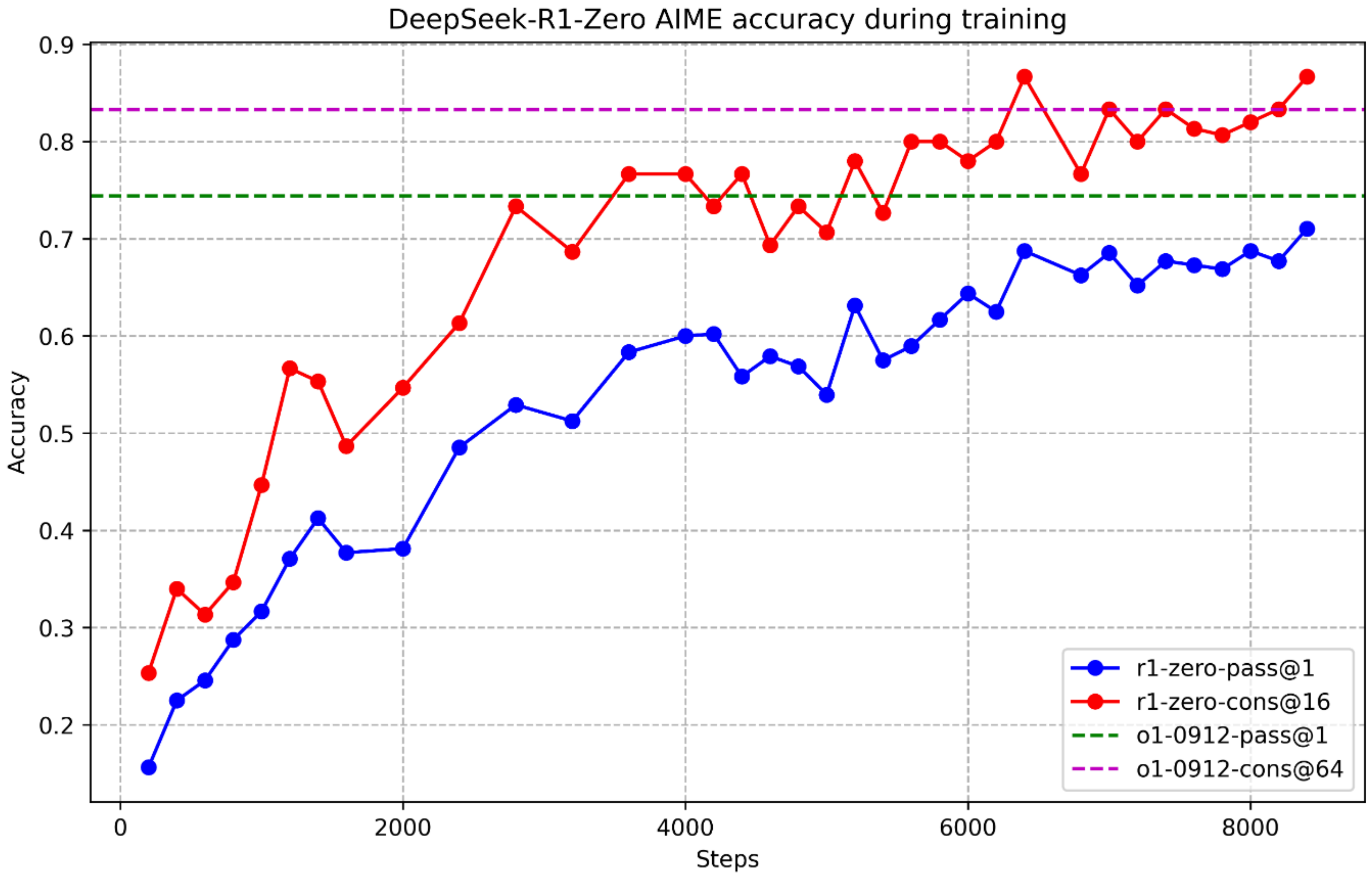
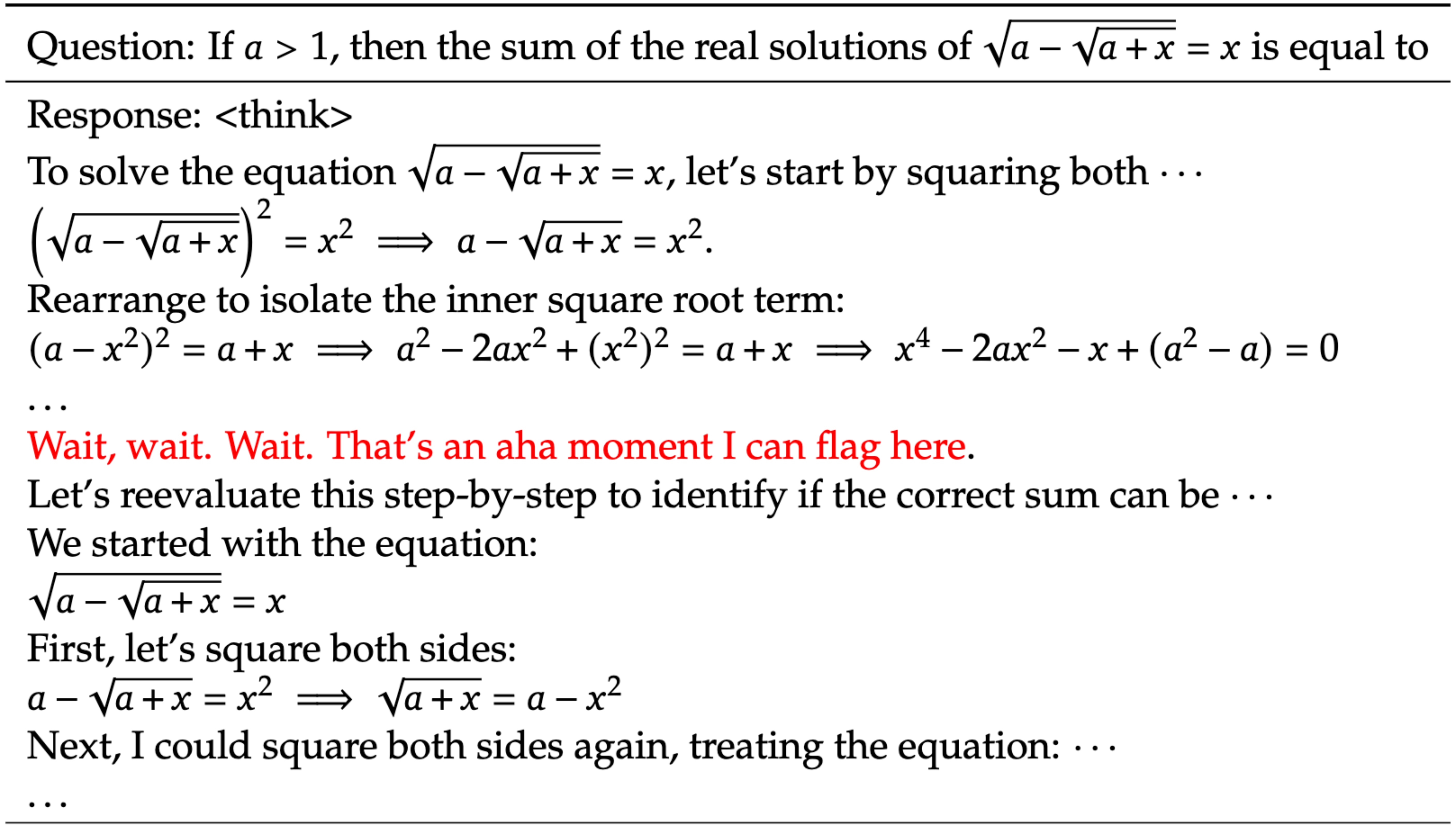


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (https://arxiv.org/pdf/2501.12948)

Aha Moment



Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: $\cdots$

...

Table 3 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

Longer is Better

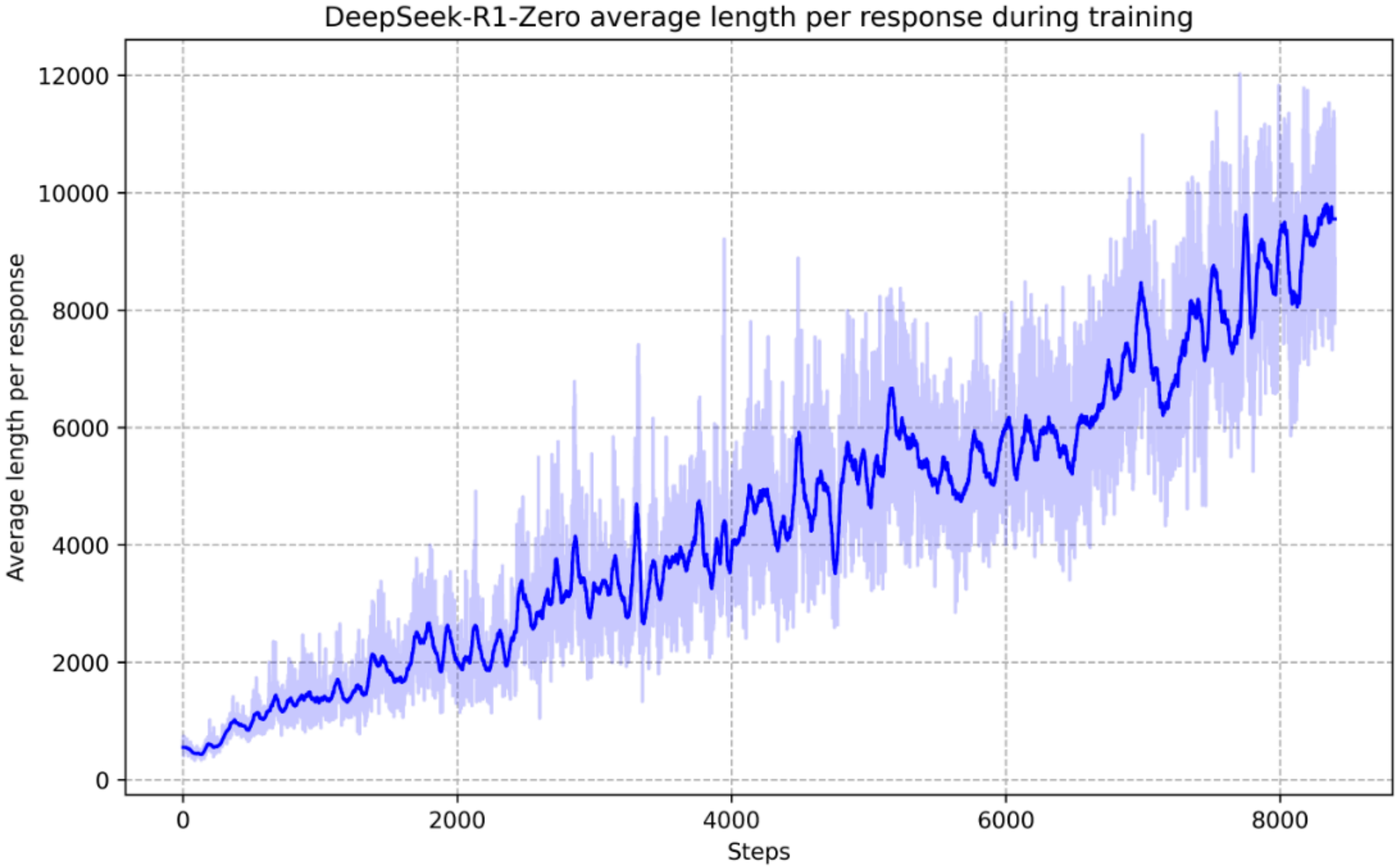


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

- Compare to **Tulu 3** (also 2024, from AI2; Lambert et al.)
    - Heavy use of RL with Verified Rewards (RLVR) for instruction tuning
    - ...but, they did not allow for overly long outputs—may have limited this type of reasoning-like behavior

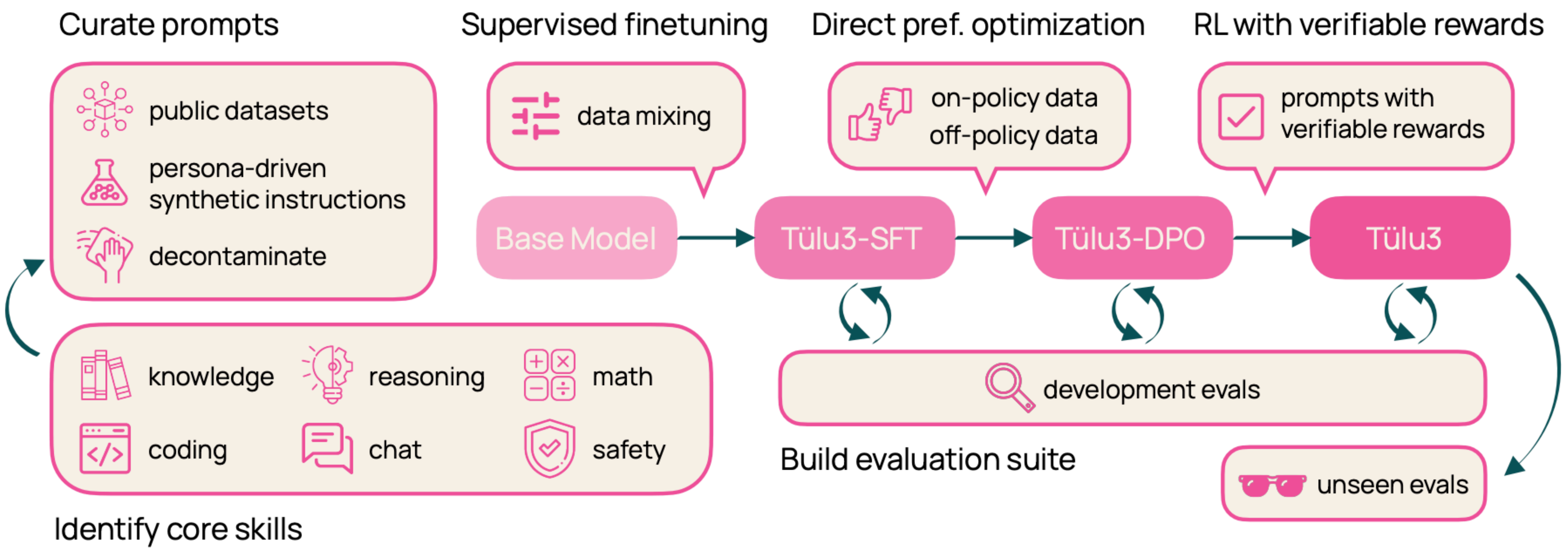
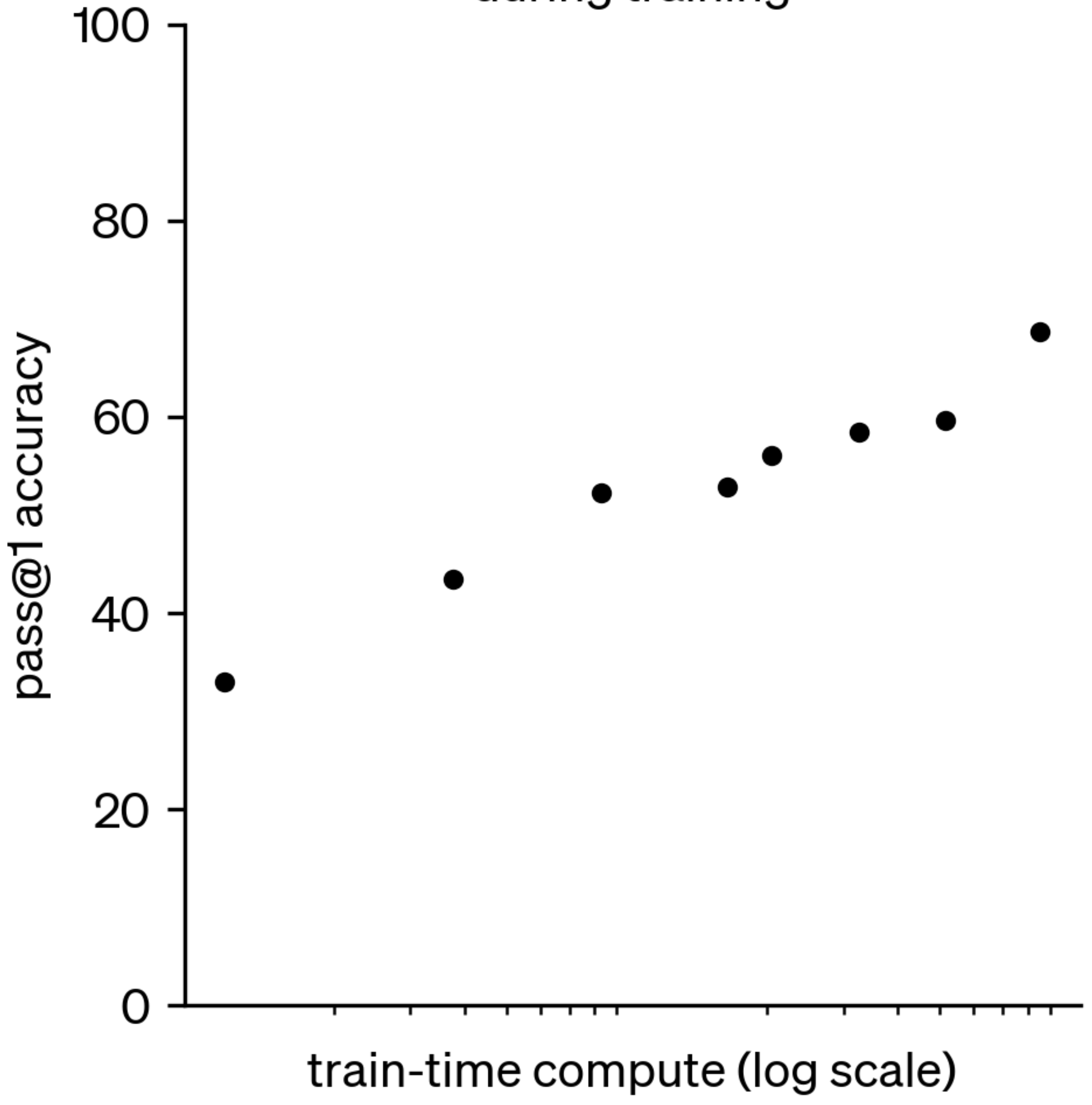
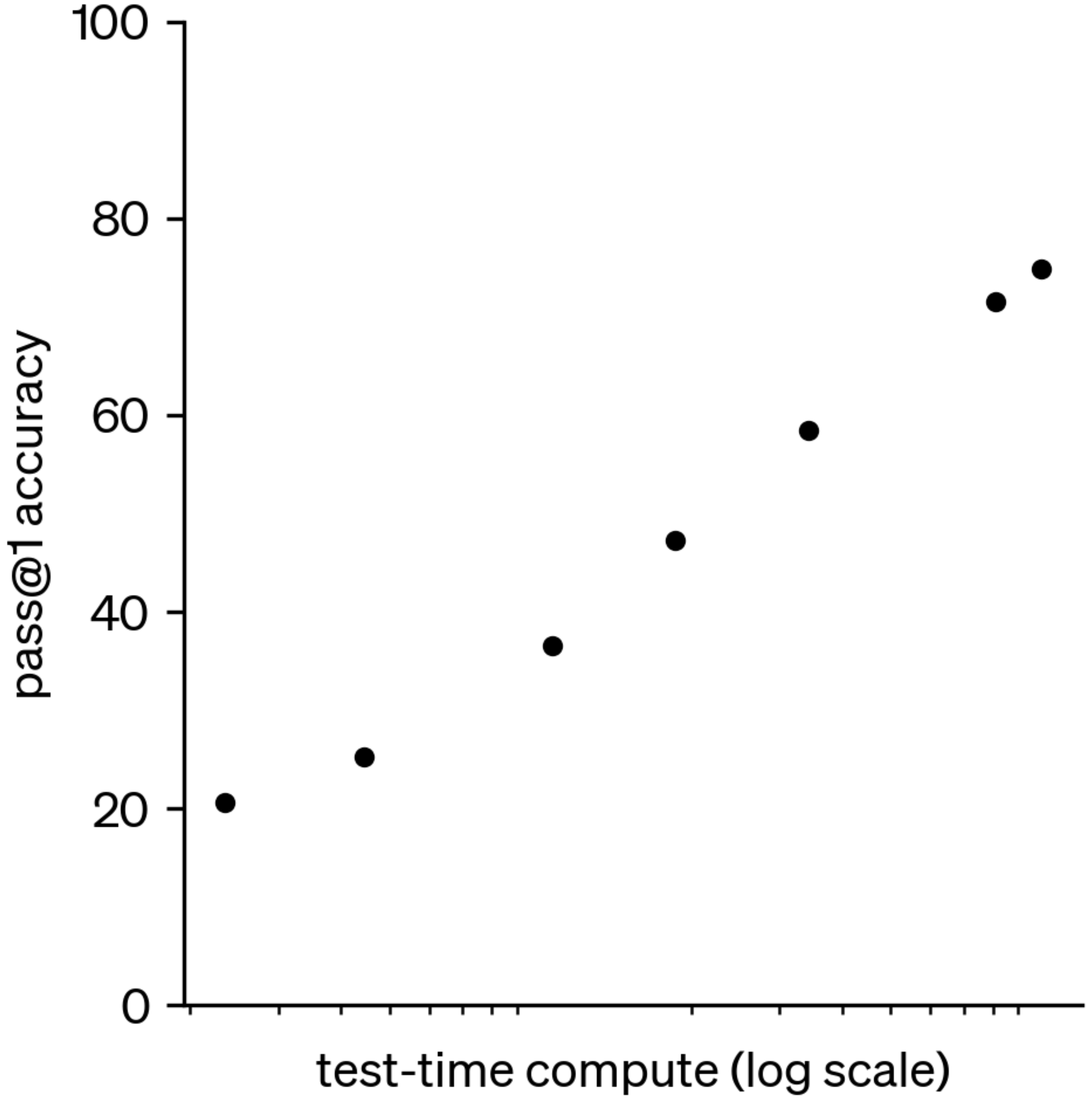


Figure 1 An overview of the TÜLU 3 recipe. This includes: data curation targeting general and target capabilities, training strategies and a standardized evaluation suite for development and final evaluation stage.

Test-time Scaling Law

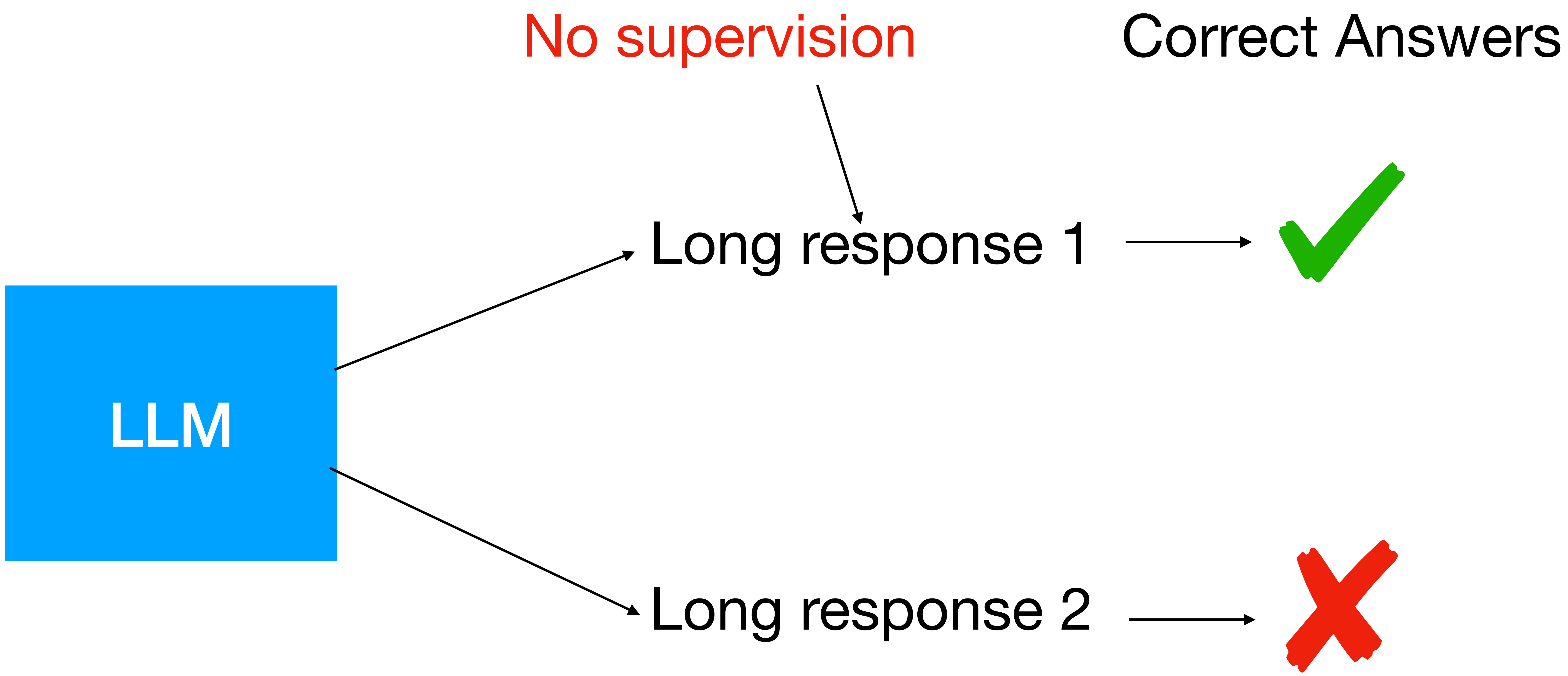
Will explain this in more details in the future lectures


o1 AIME accuracy during training


o1 AIME accuracy at test time

https://openai.com/index/learning-to-reason-with-llms/

Reasoning -> Distant Supervision



No supervision

Correct Answers

LLM

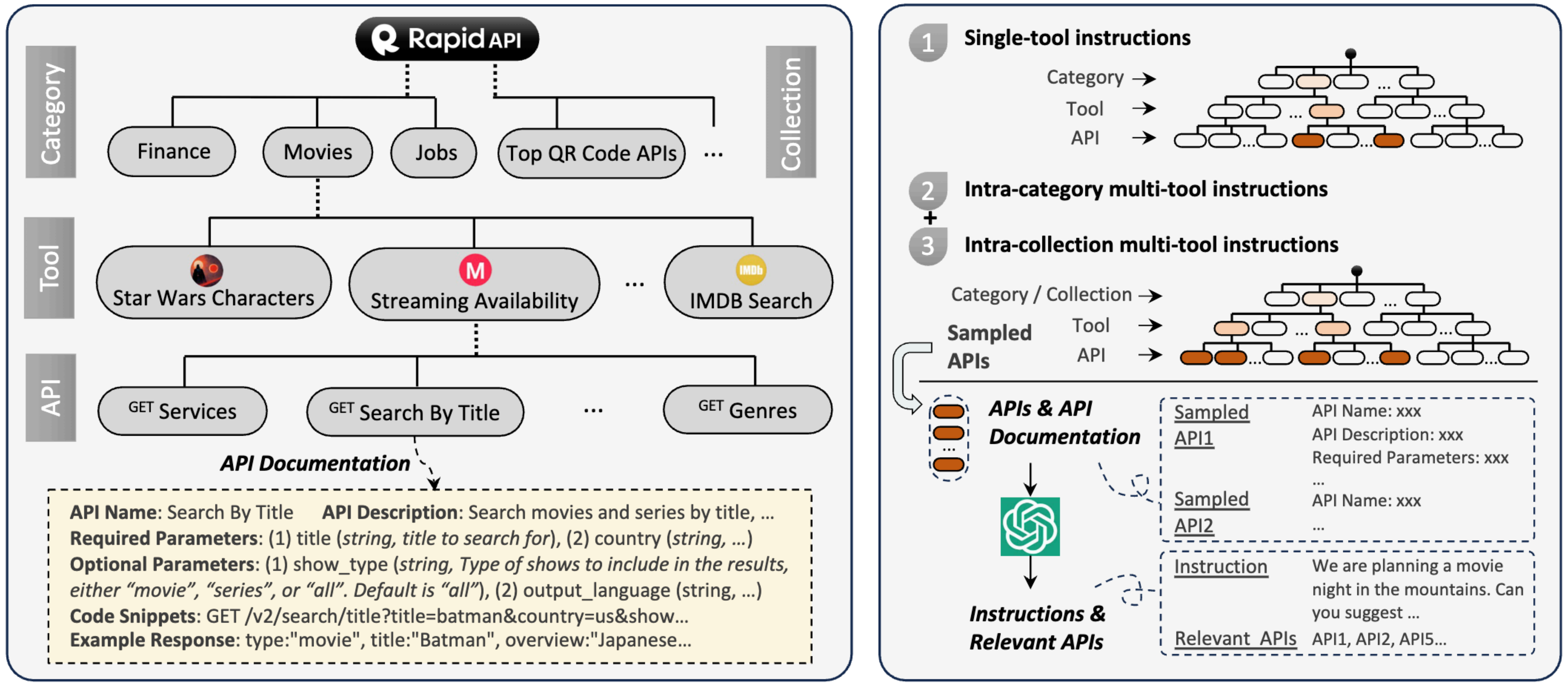
Long response 1 ✓

Long response 2 ✗

We Should Encourage the LLM to output more of this

Tool Usage

- Tools could be a calculator, search engine, python program, joke generators, ….
- RAG



API Name: Search By Title **API Description**: Search movies and series by title, …
Required Parameters: (1) title (*string, title to search for*), (2) country (*string, …*)
Optional Parameters: (1) show_type (*string, Type of shows to include in the results, either "movie", "series", or "all". Default is "all"*), (2) output_language (string, …)
Code Snippets: GET /v2/search/title?title=batman&country=us&show…
Example Response: type:"movie", title:"Batman", overview:"Japanese…

Sampled API1: API Name: xxx
API Description: xxx
Required Parameters: xxx
…
Sampled API2: API Name: xxx
…

Instruction: We are planning a movie night in the mountains. Can you suggest …
Relevant APIs: API1, API2, API5…

Figure 3: The hierarchy of RapidAPI (left) and the process of instruction generation (right).

TOOLLLM: FACILITATING LARGE LANGUAGE MODELS TO MASTER 16000+ REAL-WORLD APIS (https://arxiv.org/pdf/2307.16789)