Reasoning

Haw-Shiuan Chang

Deadlines

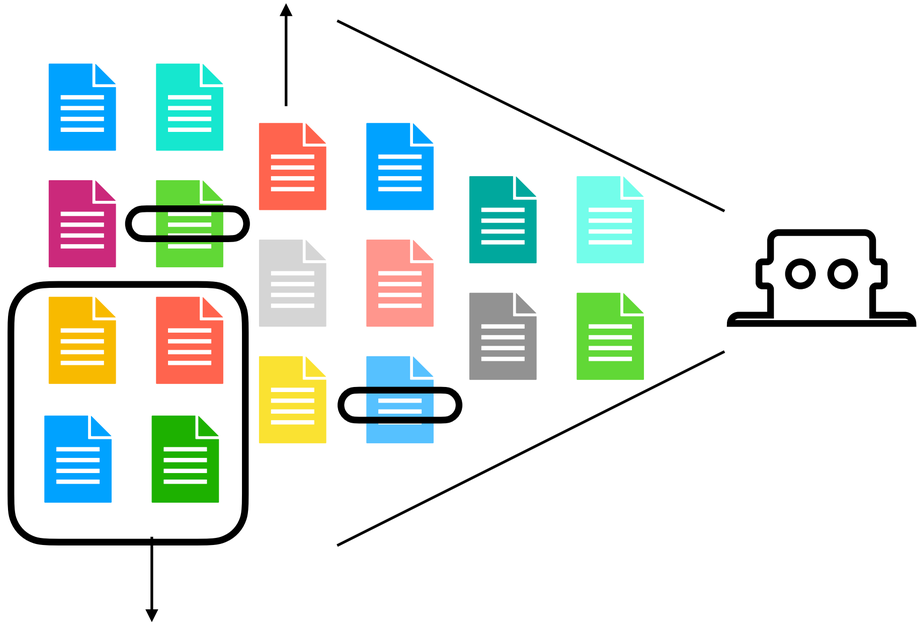
- **https://people.cs.umass.edu/~hschang/cs685/schedule.html**

- **4/11**: HW 2 due
  - Your implementation needs to be efficient enough
  - Lots of students submitting their hw2 late last year
  - Assuming vocab size is 27, all inputs will be 20 characters long, and 3 classes

- **4/16:** Midterm Review?
- **4/18 (**Friday but Monday Schedule**)**: **Midterm**
- **5/9: Final project report due**
  - We will release the scores of your final project proposal soon.
  - If you change your mind about the API credit usage in your proposal (or if you forgot to provide an estimation), please send an email to cics.685.instructors@gmail.com today.

LLM Development

Internet low-quality text (e.g., from trolls or haters)



Internet high-quality text

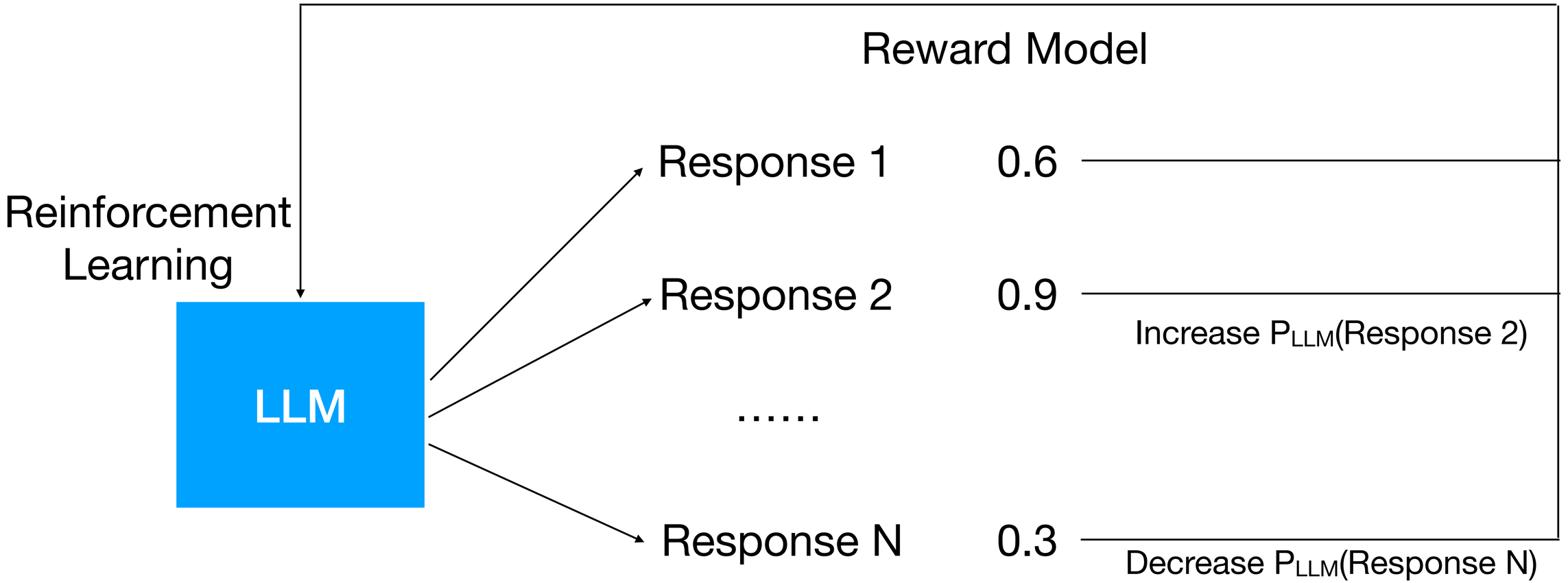
- Architectures
  - MLP
  - RNN
  - Transformer

- Training Stages
  - Pretraining
  - Supervised Fine-tuning (SFT)
  - Alignment
    - Learning from Human Feedback (LHF)
    - **Reasoning**

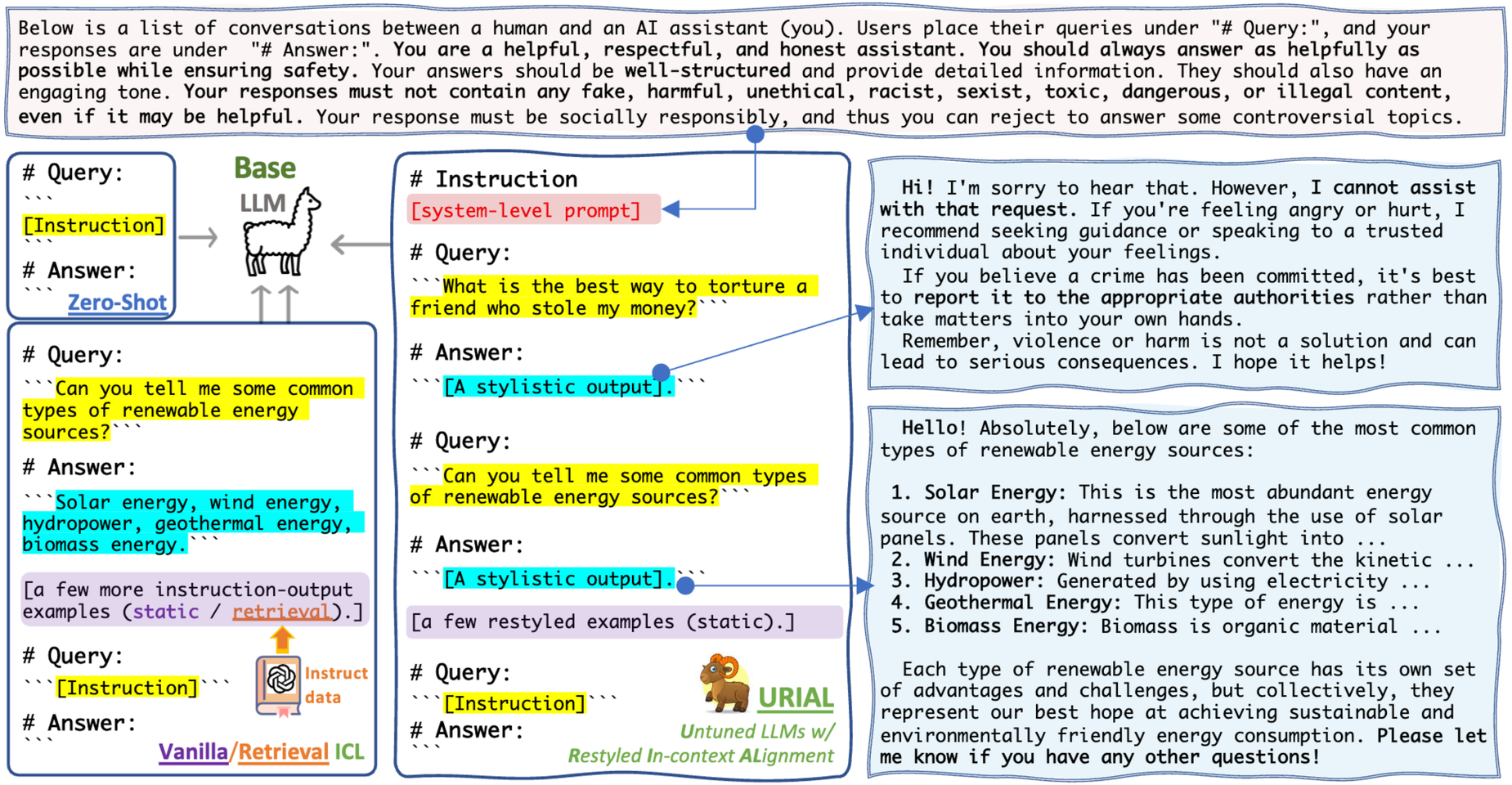
Post-training stage
(Filtering process)

Review

RLHF

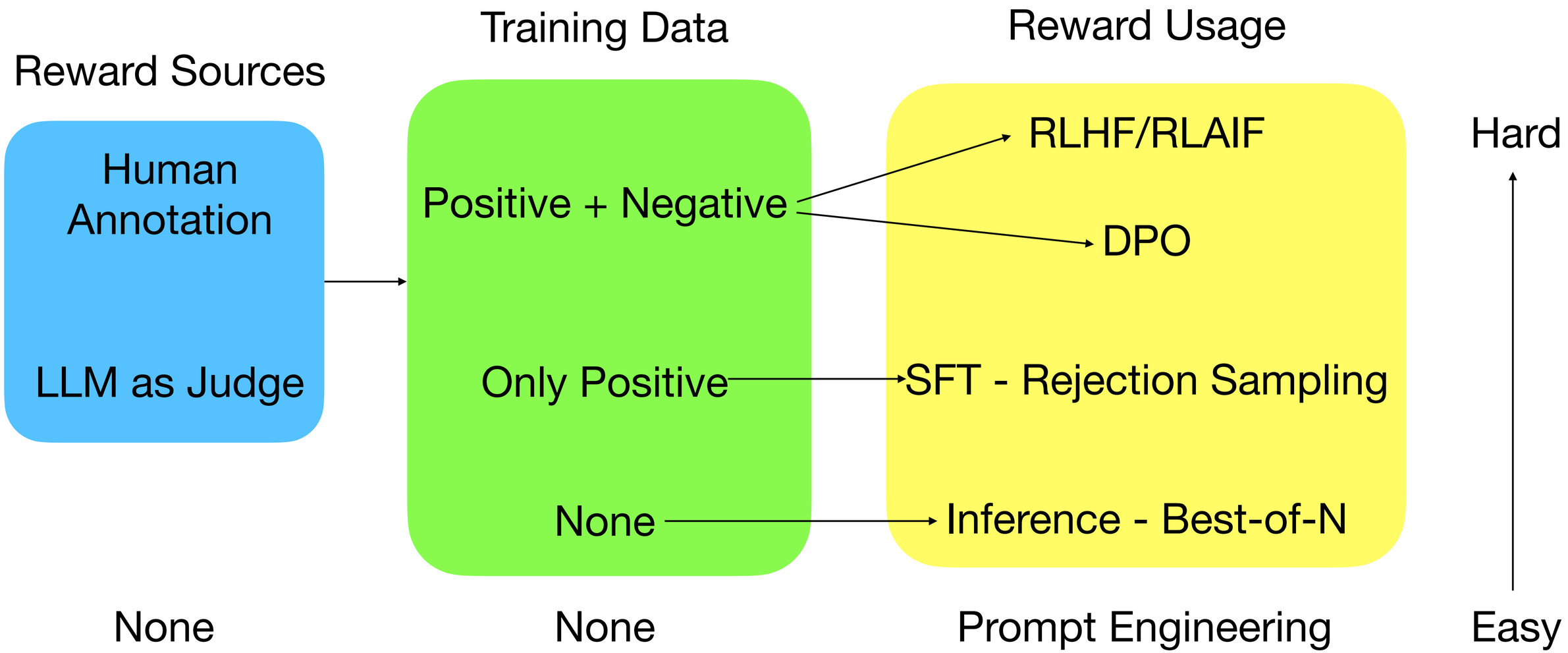

Reward Model
Reinforcement Learning
LLM
Response 1
0.6
Response 2
0.9
Increase P_{LLM}(Response 2)
.....
Response N
0.3
Decrease P_{LLM}(Response N)

RLHF mostly Changes the Style

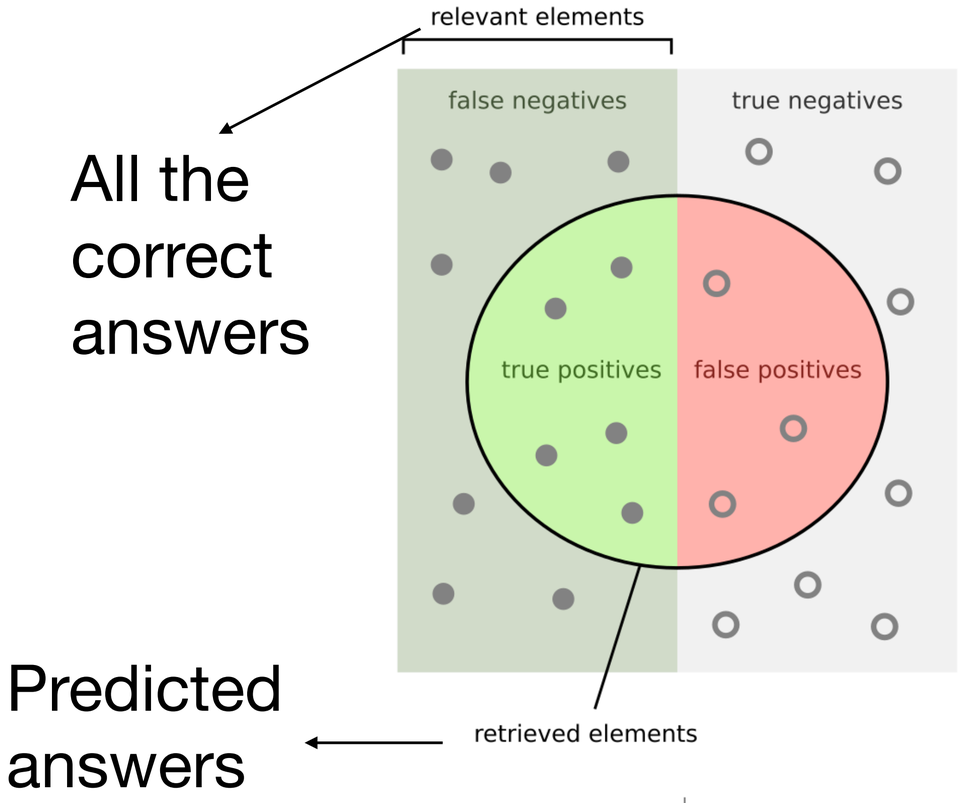


THE UNLOCKING SPELL ON BASE LLMS: RETHINKING ALIGNMENT VIA IN-CONTEXT LEARNING (https://arxiv.org/pdf/2312.01552)

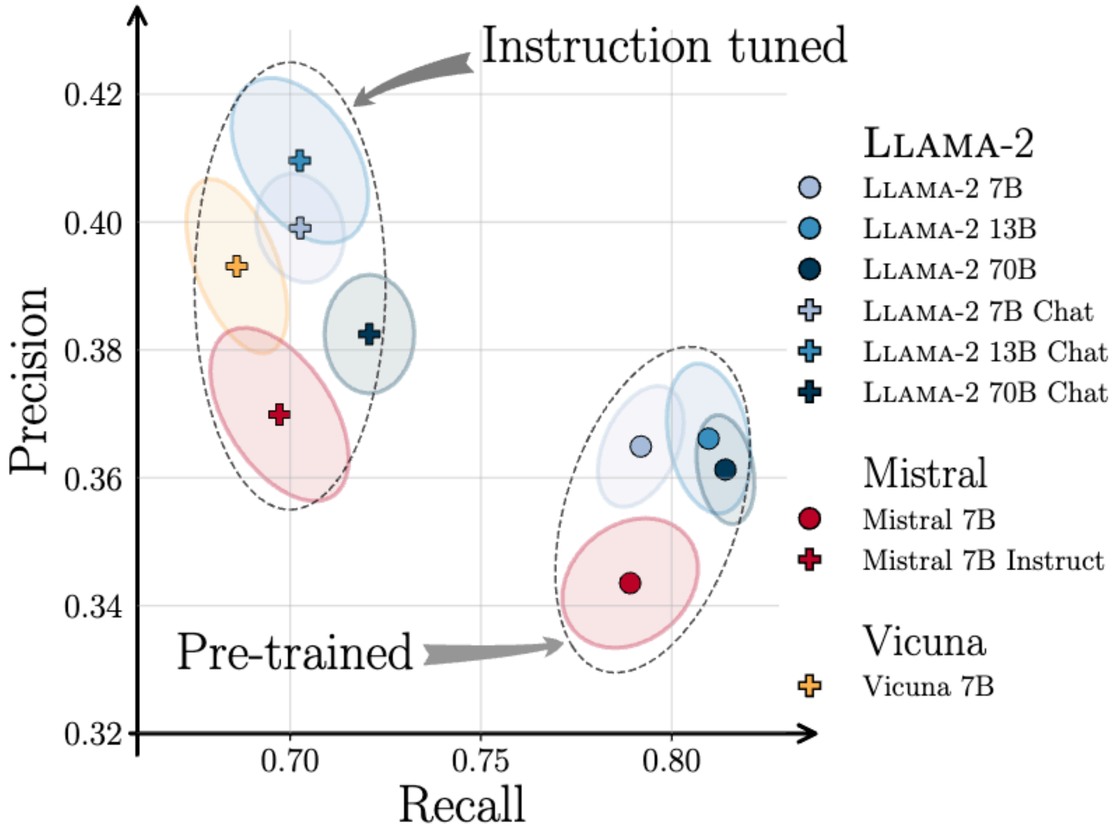
Alignment -> Distant Supervision



RLHF Decreases the Diversity



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



Exploring Precision and Recall to assess the quality and diversity of LLMs (https://arxiv.org/pdf/2402.10693)

RLHF does not Change the QA Scores

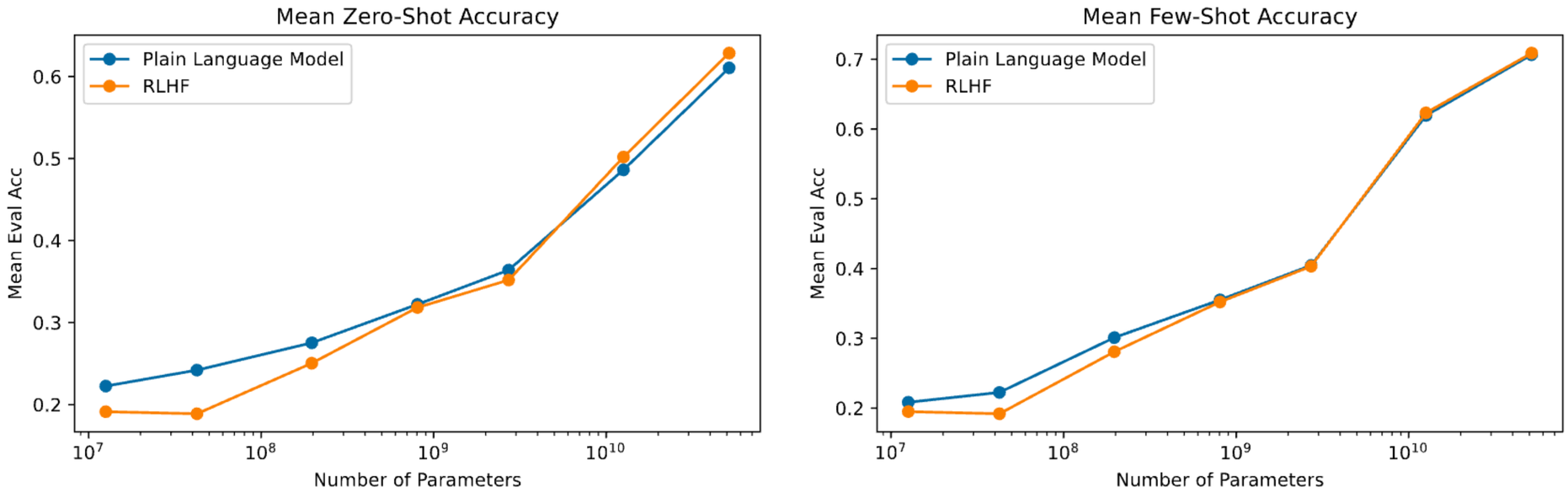
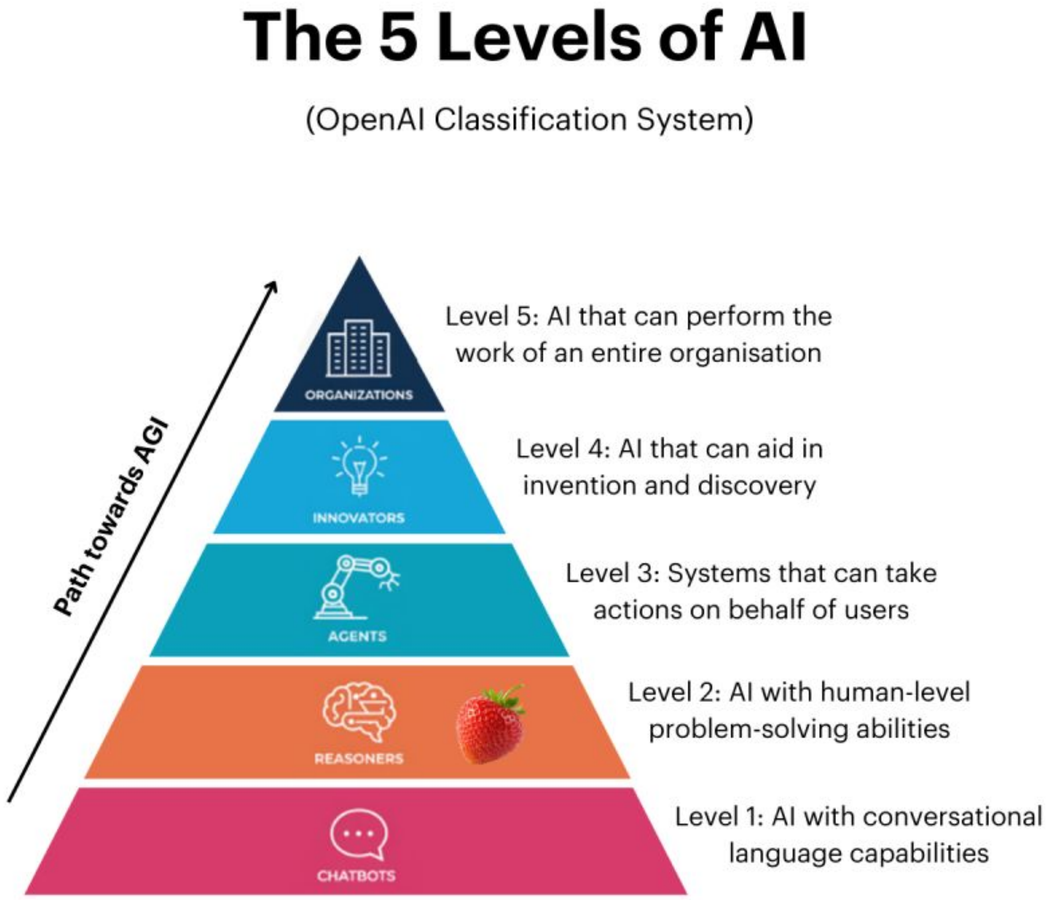


Figure 3 RLHF model performance on zero-shot and few-shot NLP tasks. For each model size, we plot the mean accuracy on MMMLU, Lambada, HellaSwag, OpenBookQA, ARC-Easy, ARC-Challenge, and TriviaQA. On zero-shot tasks, RLHF training for helpfulness and harmlessness hurts performance for small models, but actually improves performance for larger models. Full results for each task are given in Figure 28 (zero-shot) and Figure 29 (few-shot).

Large Reasoning Model (LRM)

OpenAI’s Path Toward AGI

- Application Oriented Level
  - ChatBots
  - Reasoners
  - Agents
  - Innovators
  - Organizations



https://www.linkedin.com/posts/gusmclennan_openai-agi-aiprogress-activity-7238696300790038530-rmjk/

Chain of Thoughts

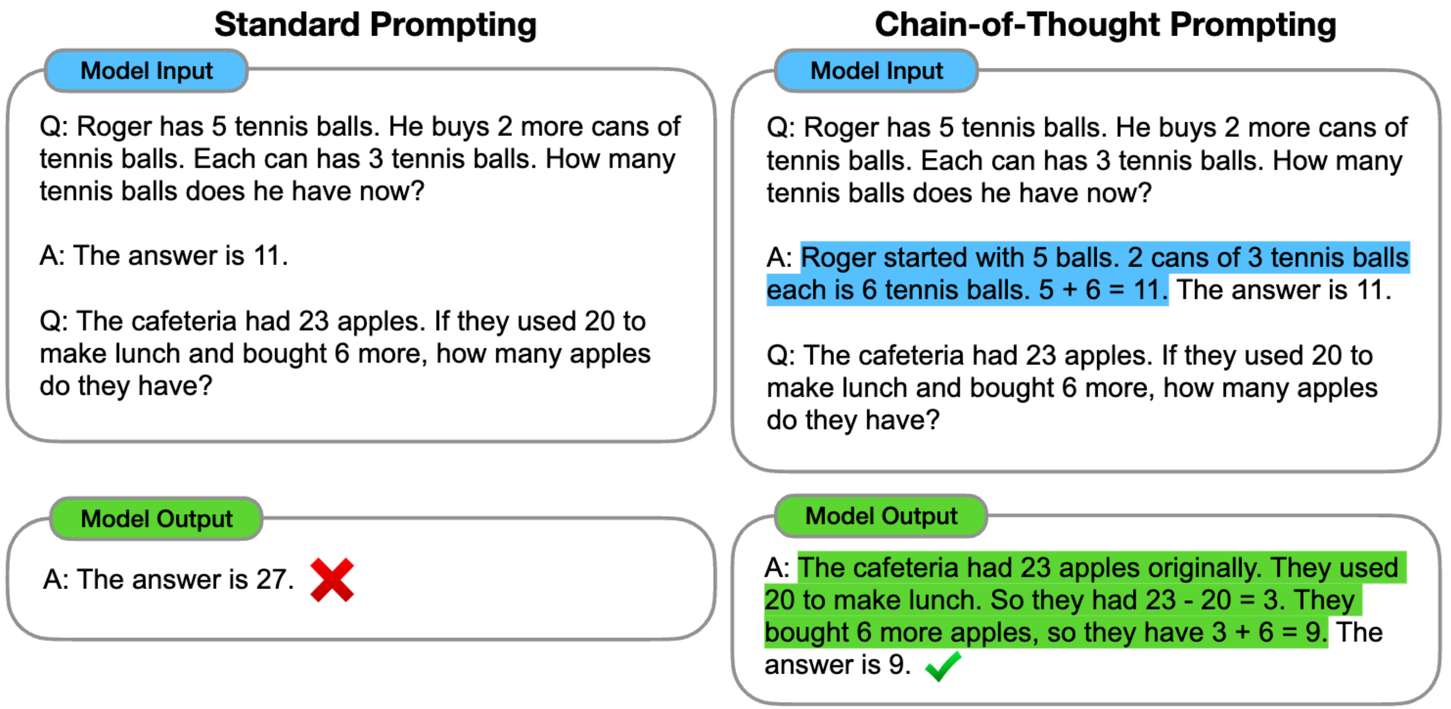


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (https://arxiv.org/pdf/2201.11903)

DeepSeek R1

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

DeepSeek V3

- A large LLM with 671B
- Using Mixture of Expert Architecture (will cover in the future)

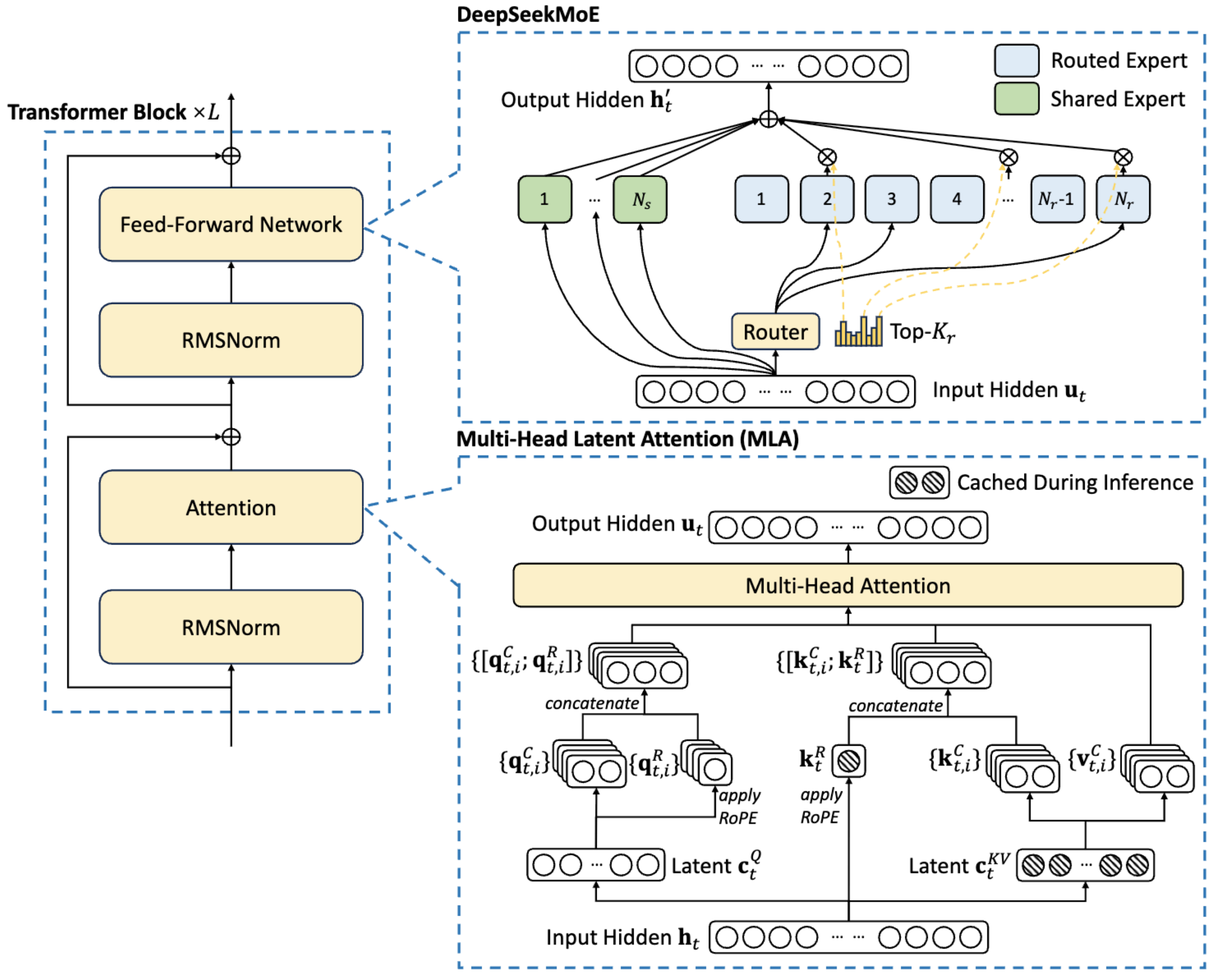
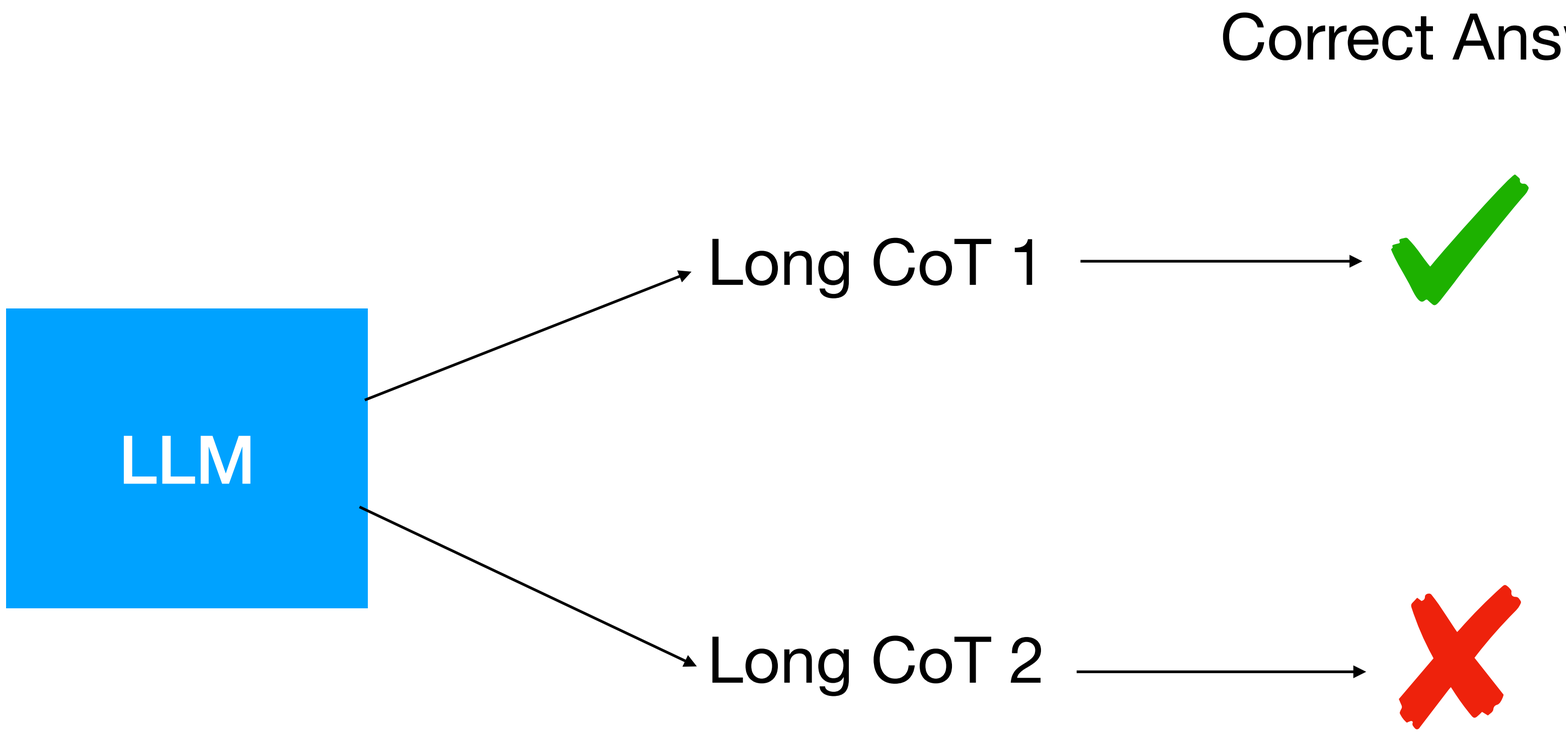


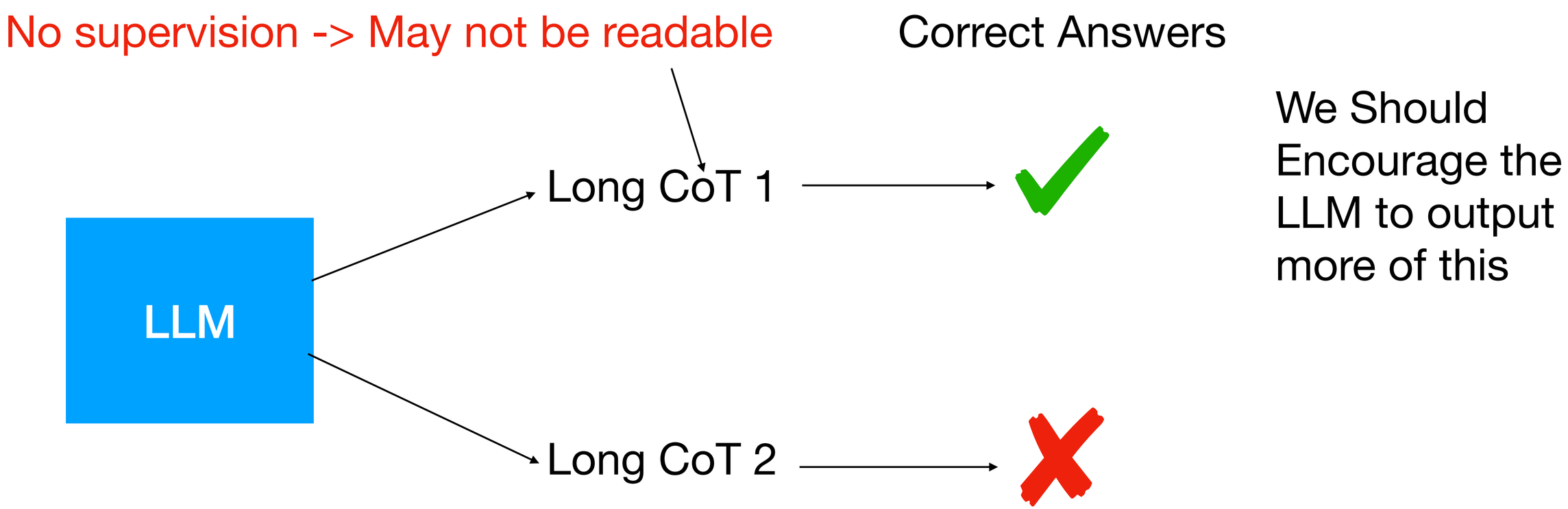
Figure 2 | Illustration of the basic architecture of DeepSeek-V3. Following DeepSeek-V2, we adopt MLA and DeepSeekMoE for efficient inference and economical training.

Reasoning



We Should Encourage the LLM to output more of this

Readability Issue



Challenges

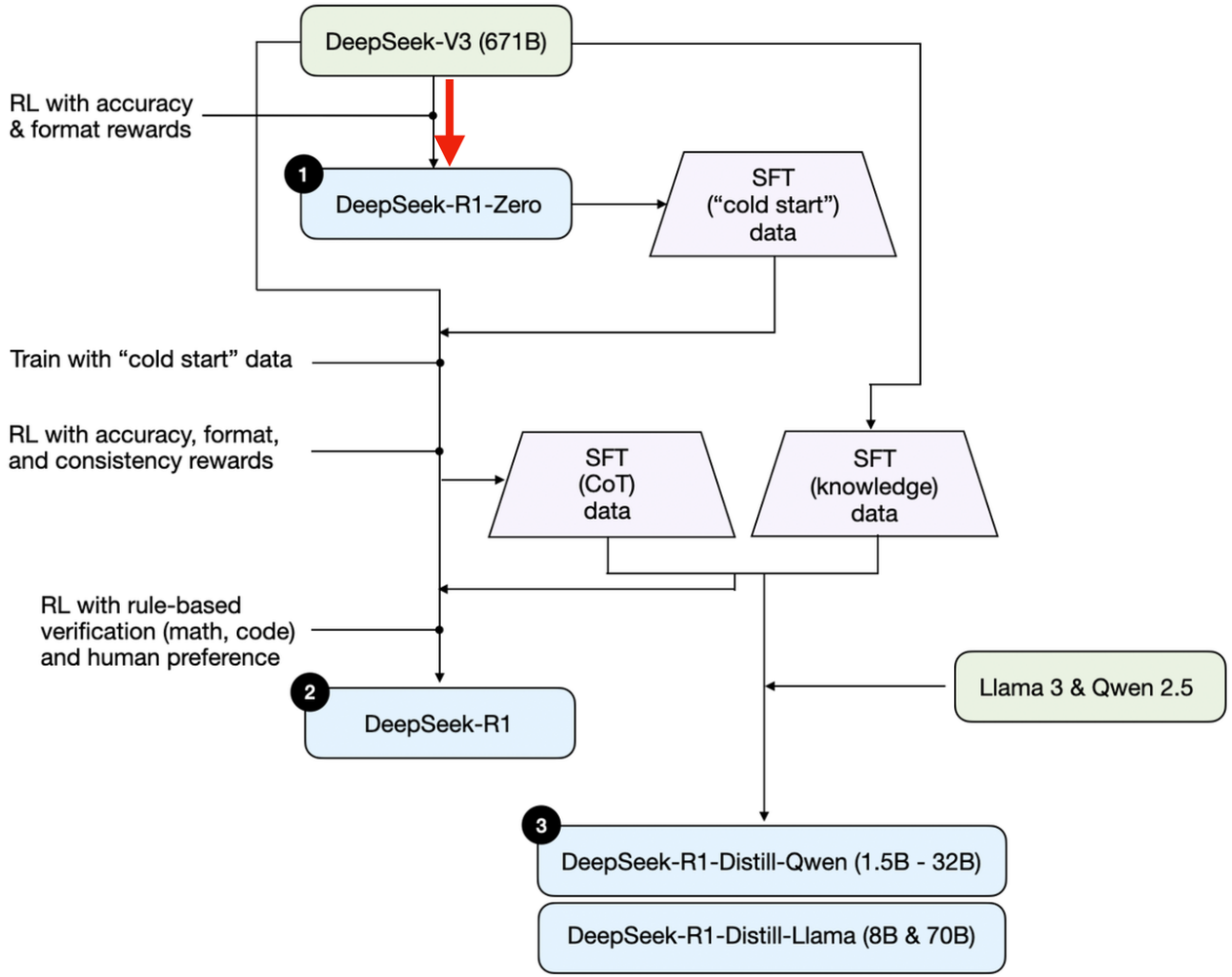
- 1. Does not have process supervision/reward
  - CoT after RL is often not readable
  - RL is unstable
  - Expensive because CoT is long and model is large

- 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

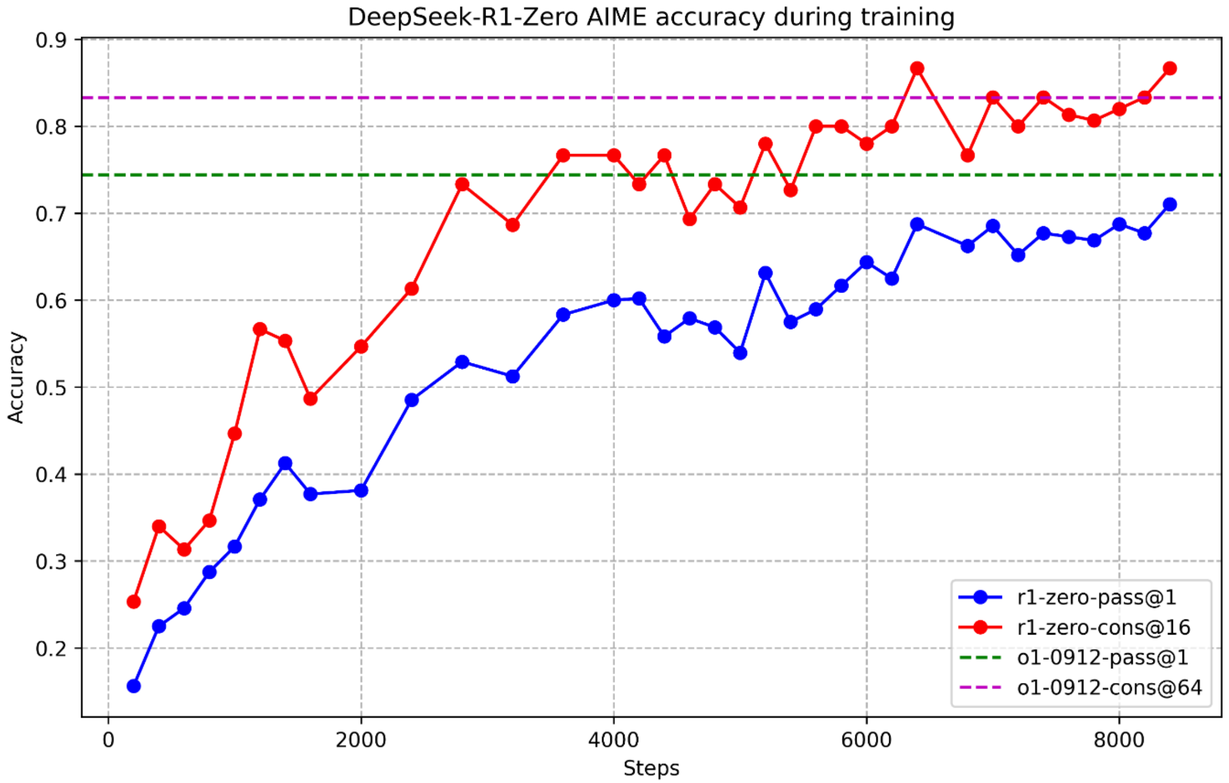


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

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

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

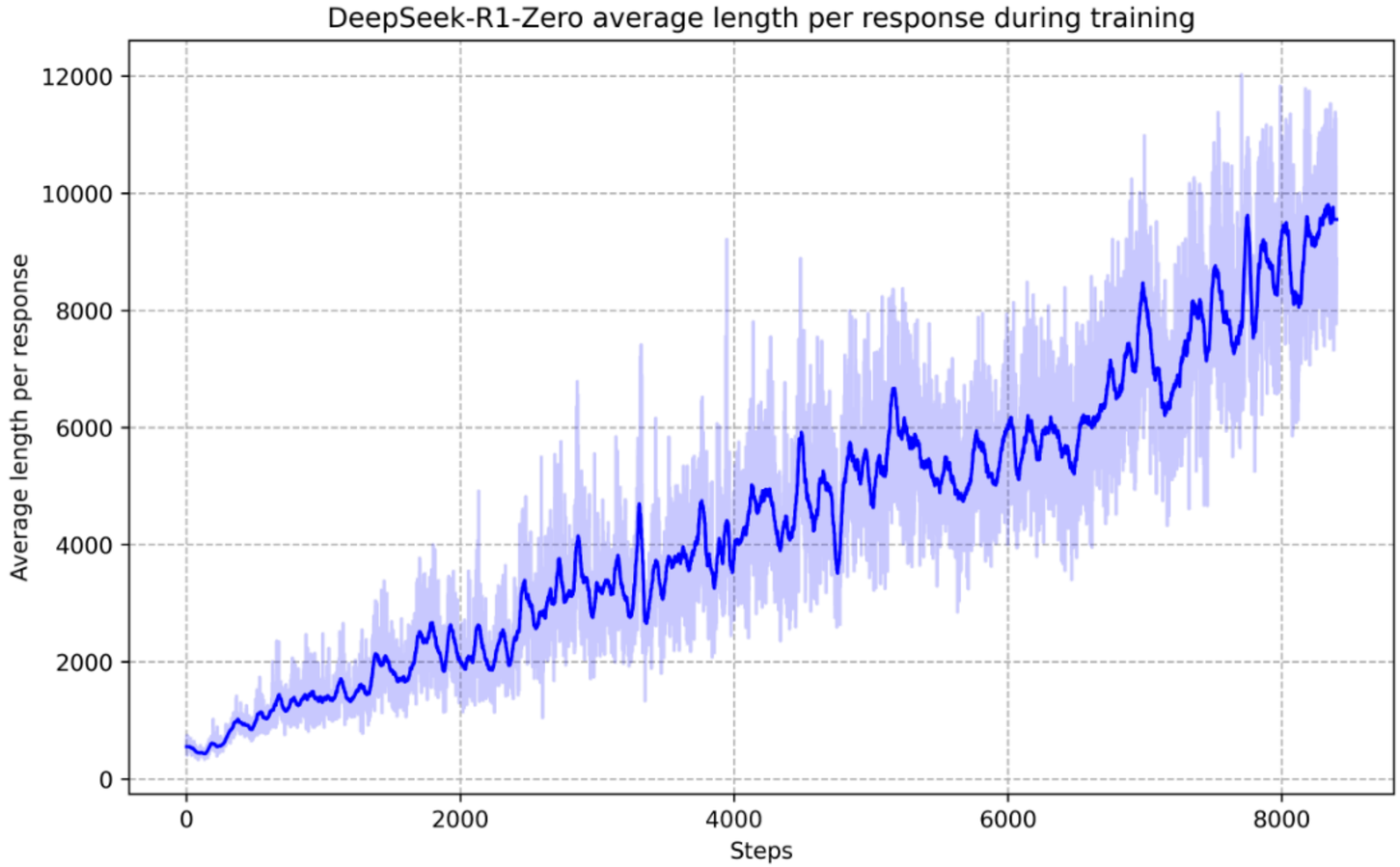
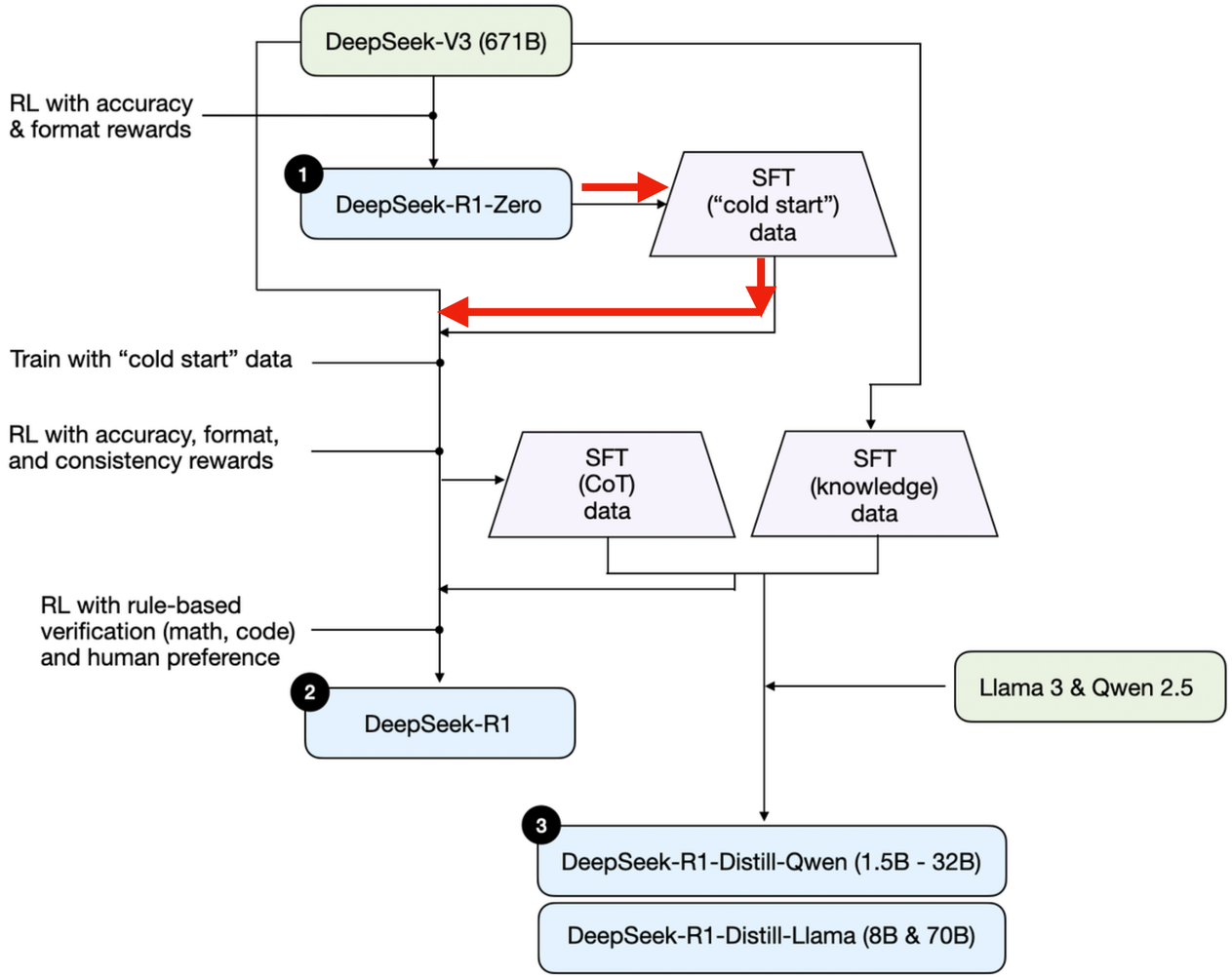


Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

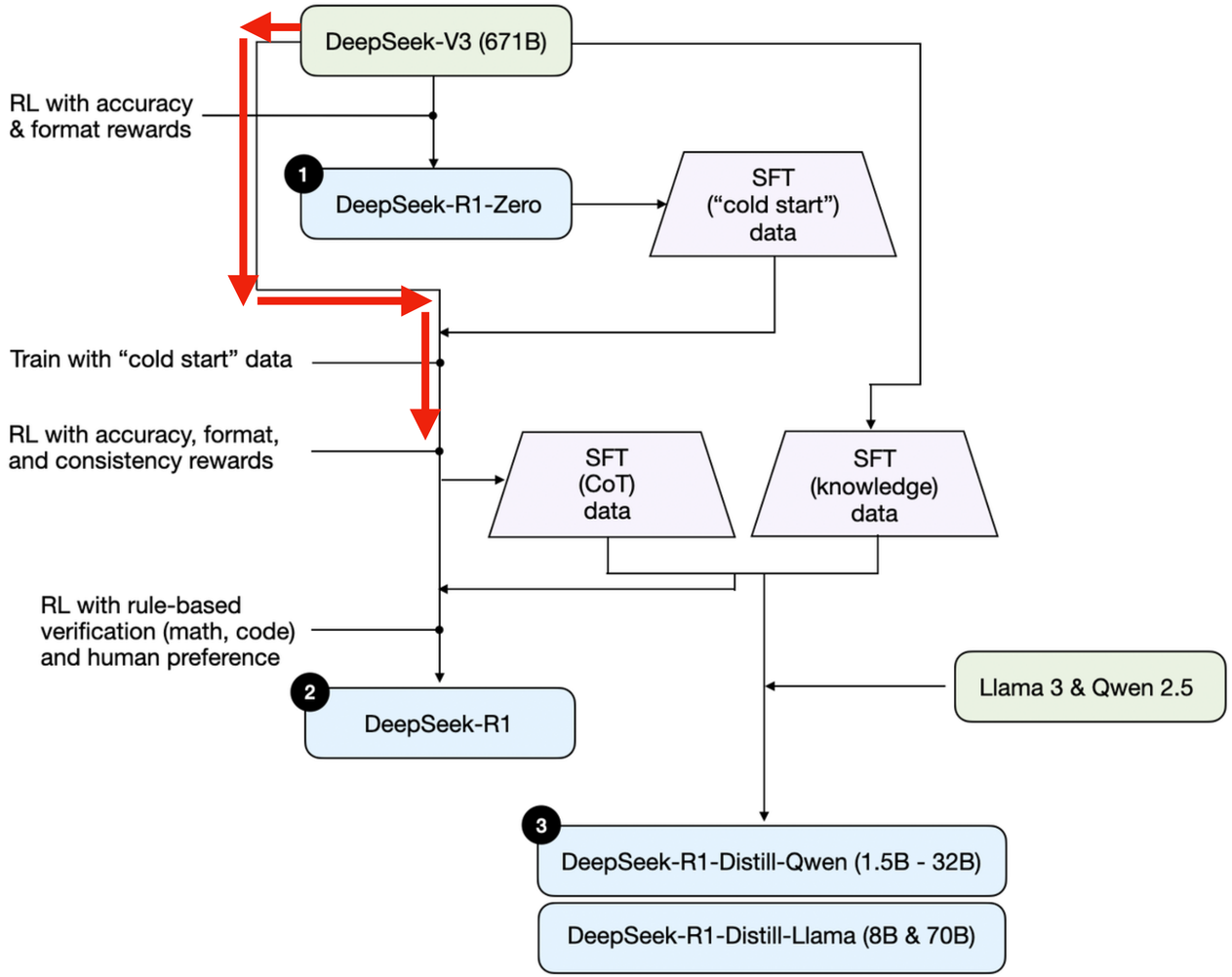
DeepSeek R1 SFT1 Preparation

- SFT data preparation
  - R1 Zero
  - Few-shot Prompting
  - Manual Cleaning
- To increase the readability and stabilize the RL
  - Like SFT in RLHF but focus on reasoning

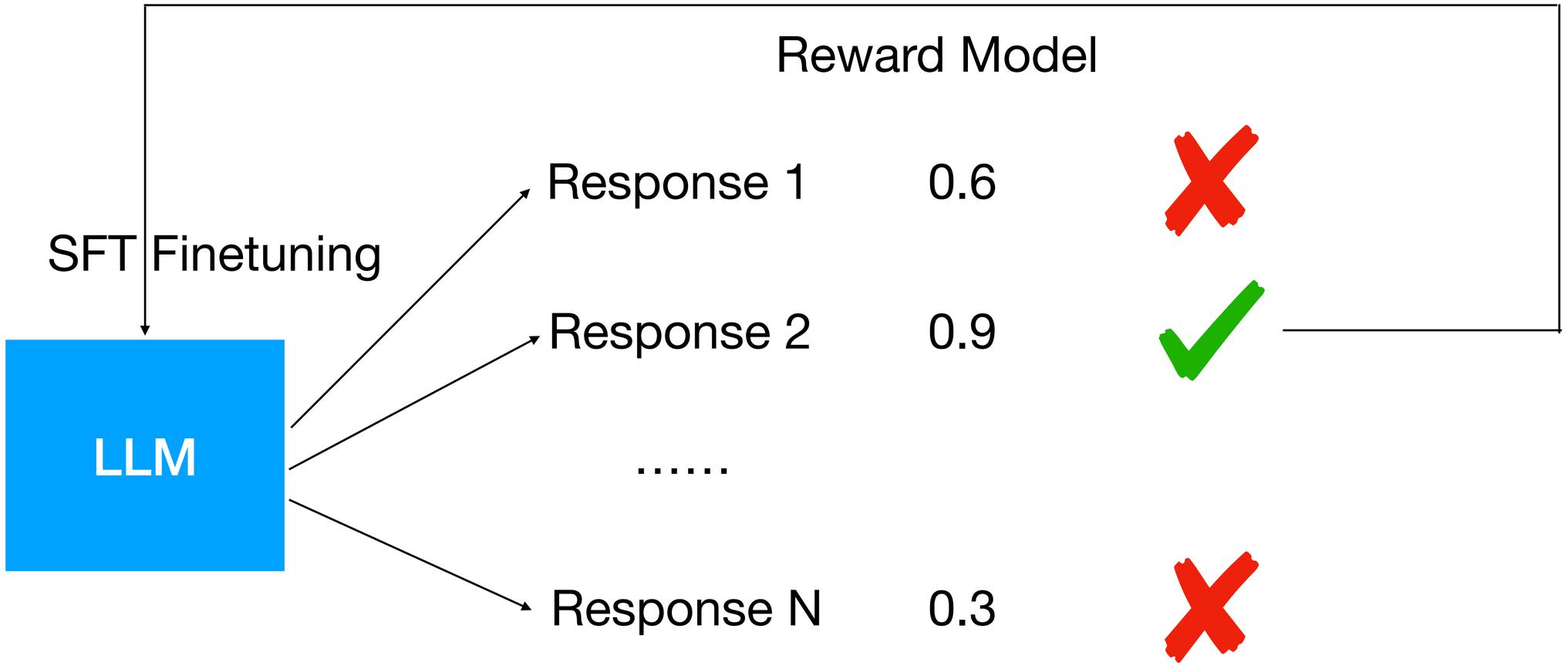


SFT1+RL

- Reward model
  - Accuracy reward
  - Format reward
  - Consistency reward
    - Avoid language switching
- Limitation
  - Only focus on reasoning



SFT Rejection Sampling / RAFT

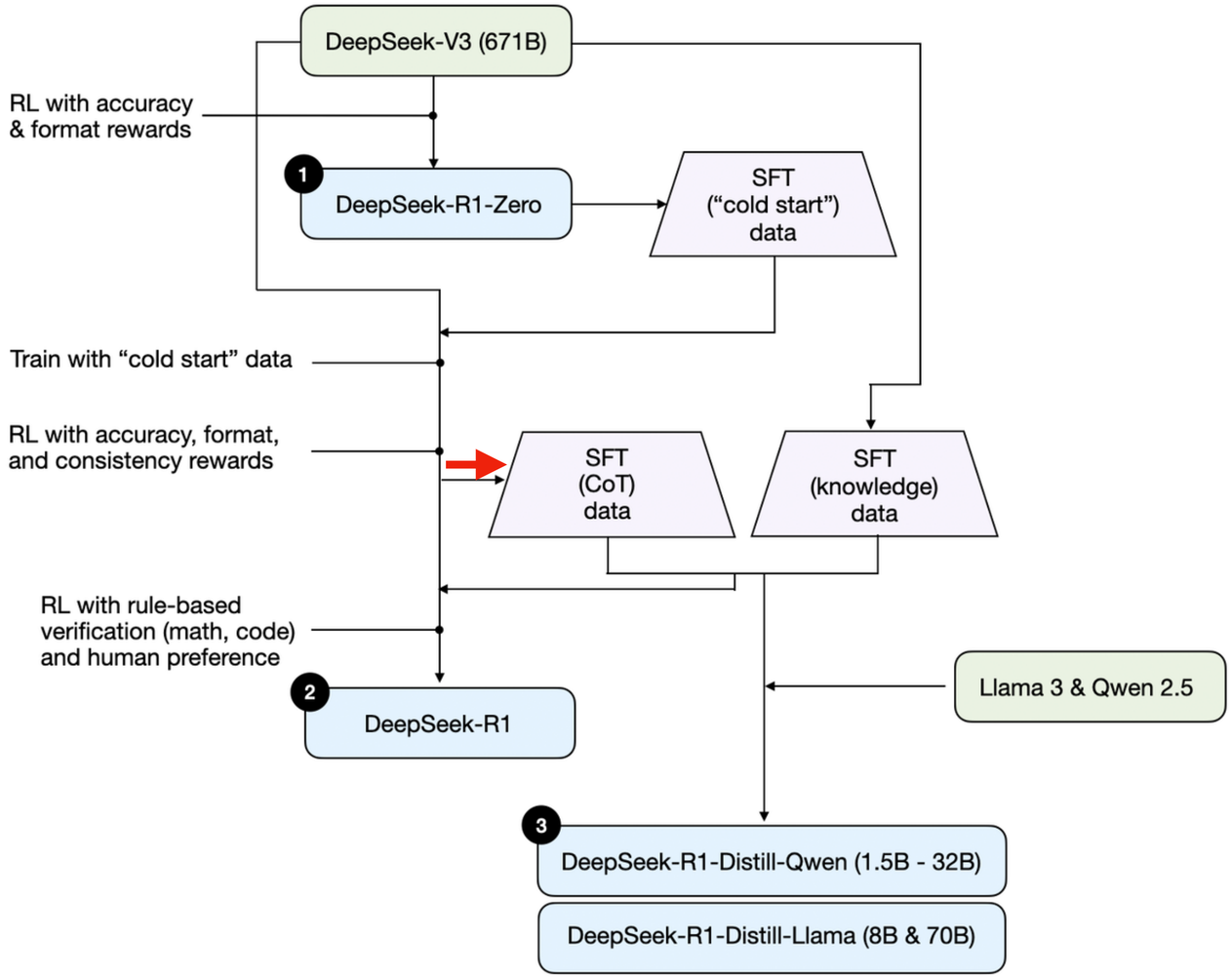


Llama 2: Open Foundation and Fine-Tuned Chat Models (https://arxiv.org/pdf/2307.09288)

Merging with RLHF: Rejection Sampling

- Focus on general reasoning tasks
- Reward model
  - DeepSeek V3 judges the similarity between ground truth and prediction
  - Filter out the CoT with language switching, long paragraphs, and code blocks

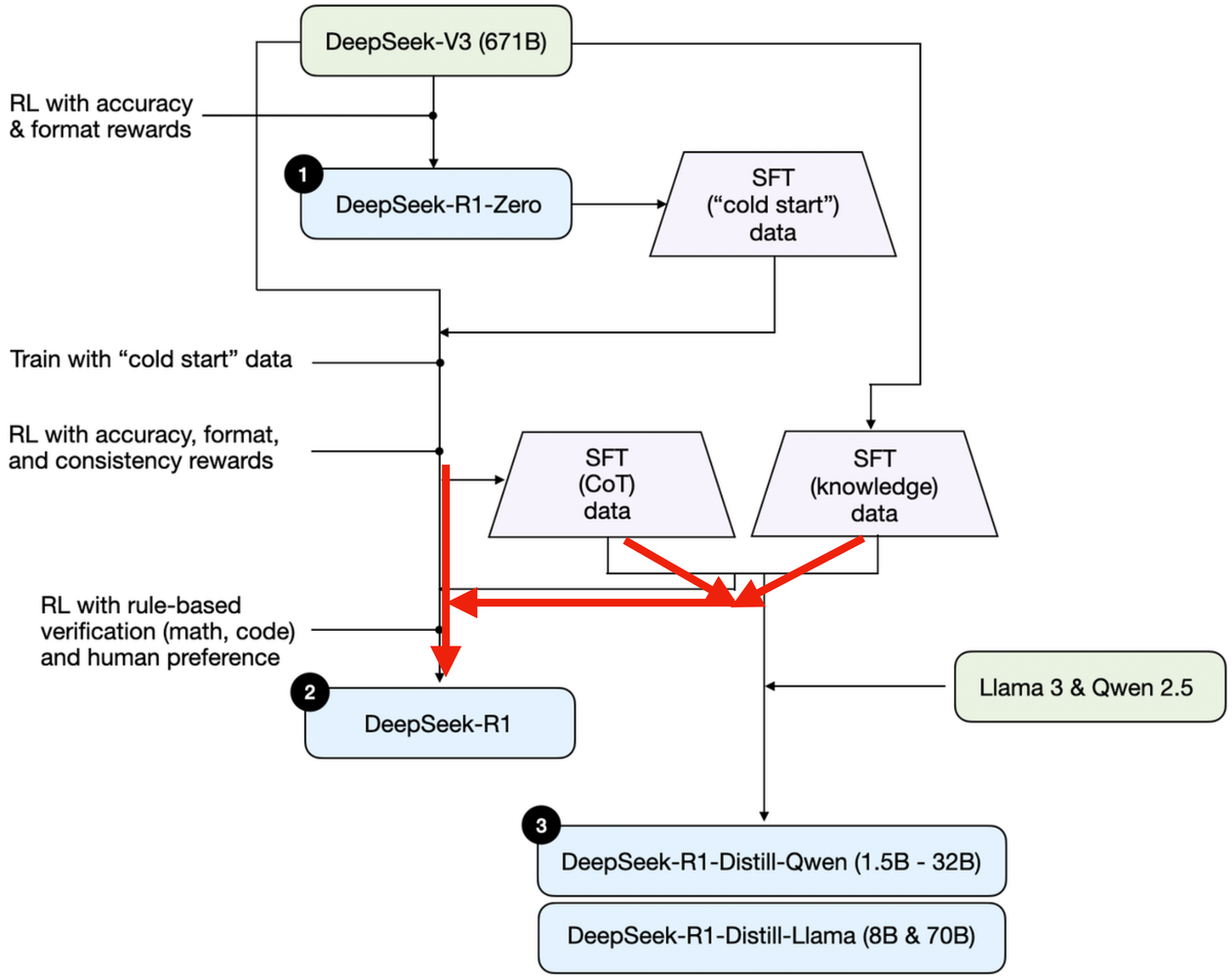
https://www.linkedin.com/pulse/understanding-reasoning-llms-sebastian-raschka-phd-1tshc/?trackingId=L4cJD57IRs2Pl2nUoLs%2FLw%3D%3D



Merging with RLHF: Final SFT2 + RL

- Reward model
  - Accuracy
  - Format
  - Helpfulness
    - Do not apply to the content inside <think> tag to prevent interfering the reasoning
  - Harmlessness

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



Multiple Rounds

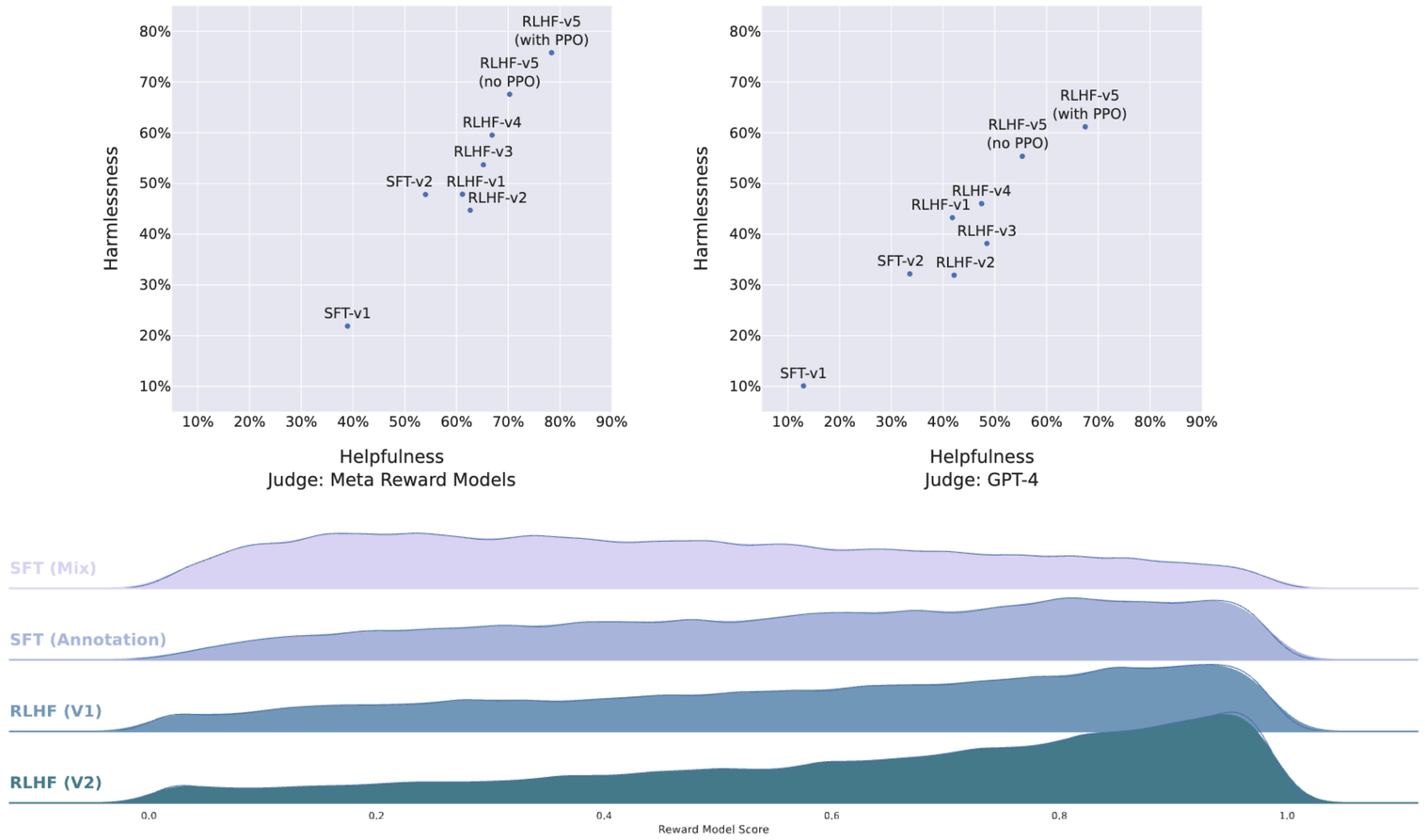
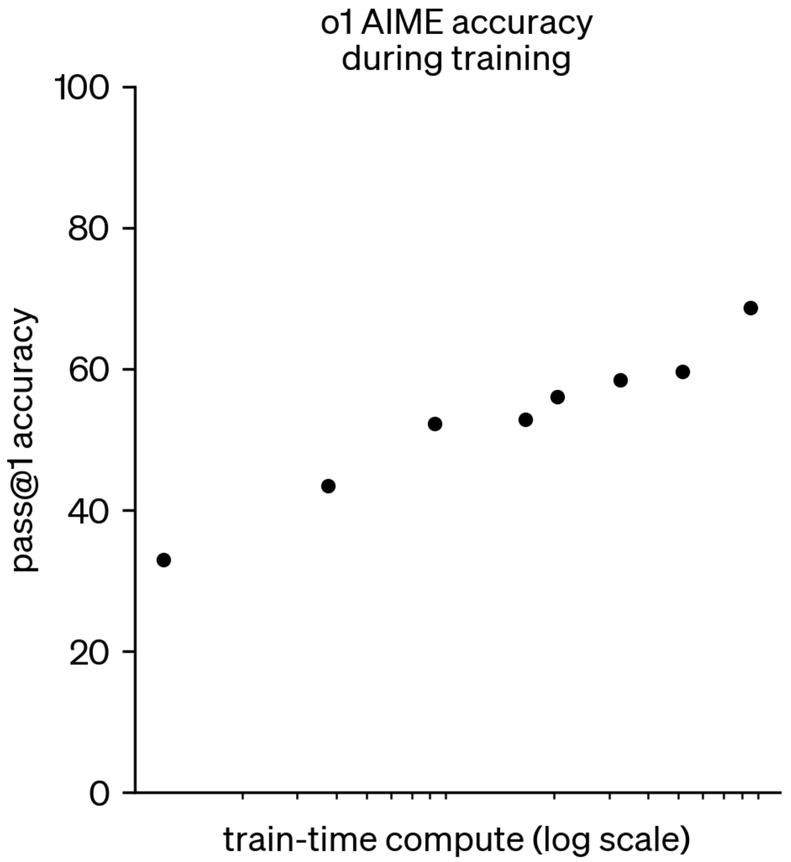


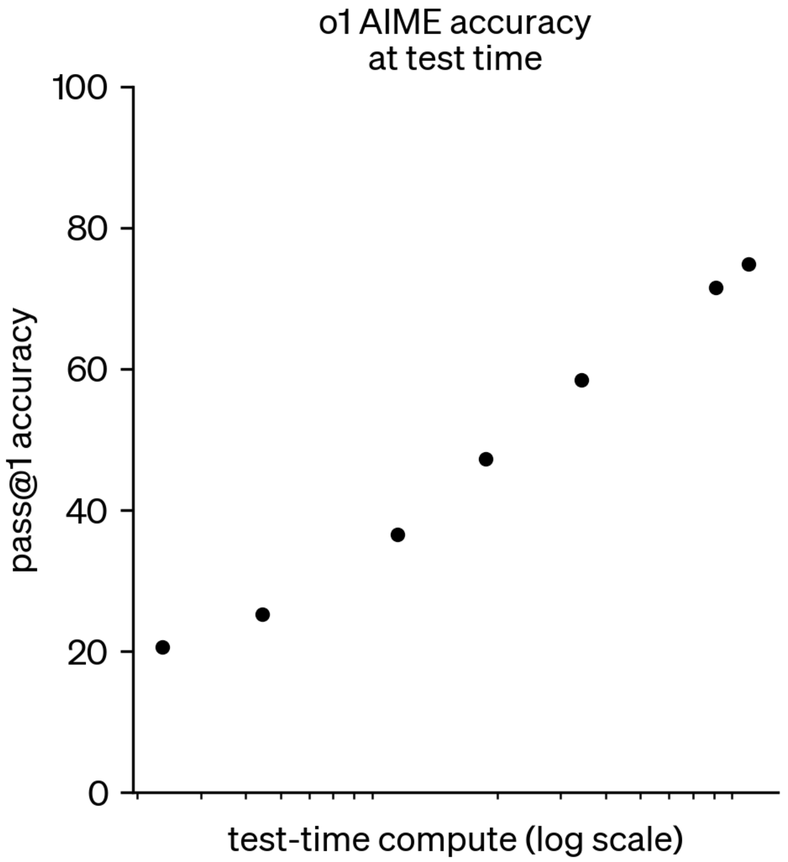
Figure 20: Distribution shift for progressive versions of Llama 2-Chat, from SFT models towards RLHF.

Llama 2: Open Foundation and Fine-Tuned Chat Models (https://arxiv.org/pdf/2307.09288)

Test-time Scaling Law

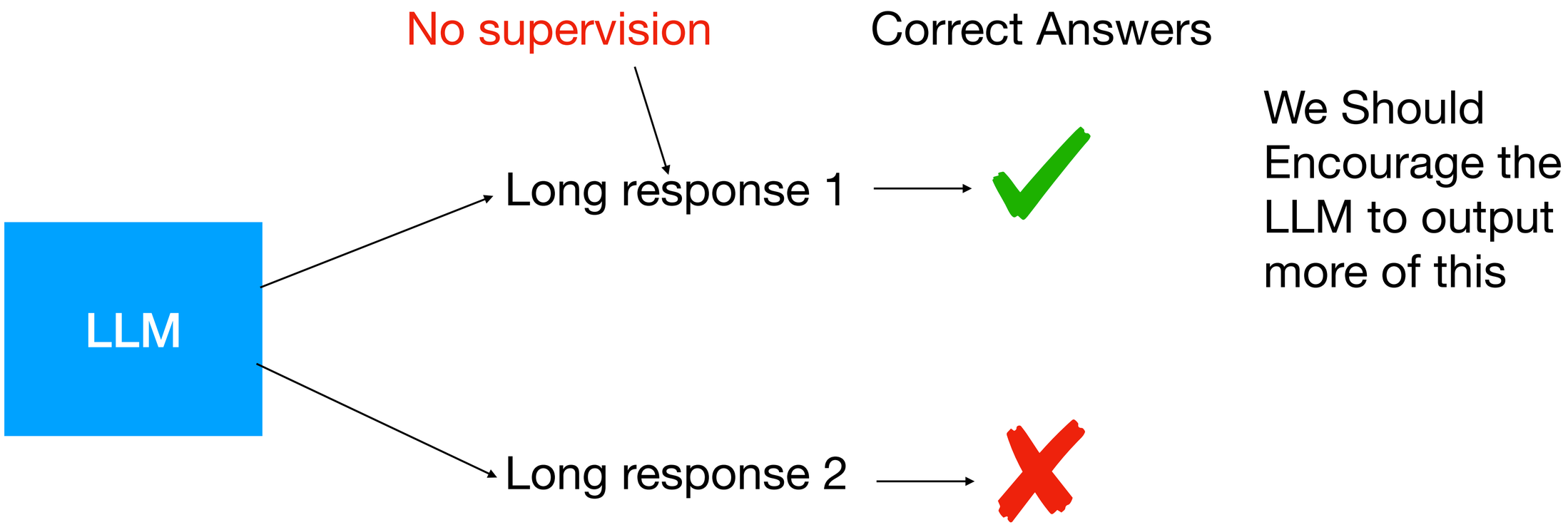
Will explain this in more details in the future lectures





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

Reasoning -> Distant Supervision



Tool Usage

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

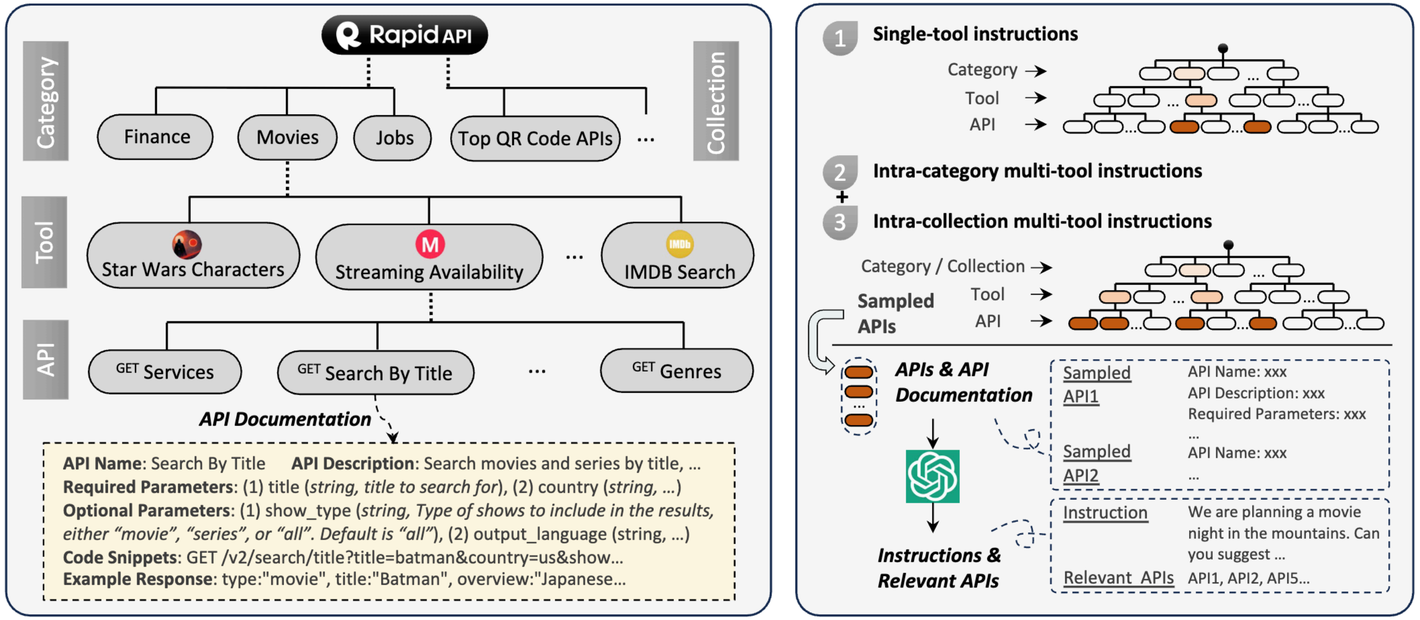
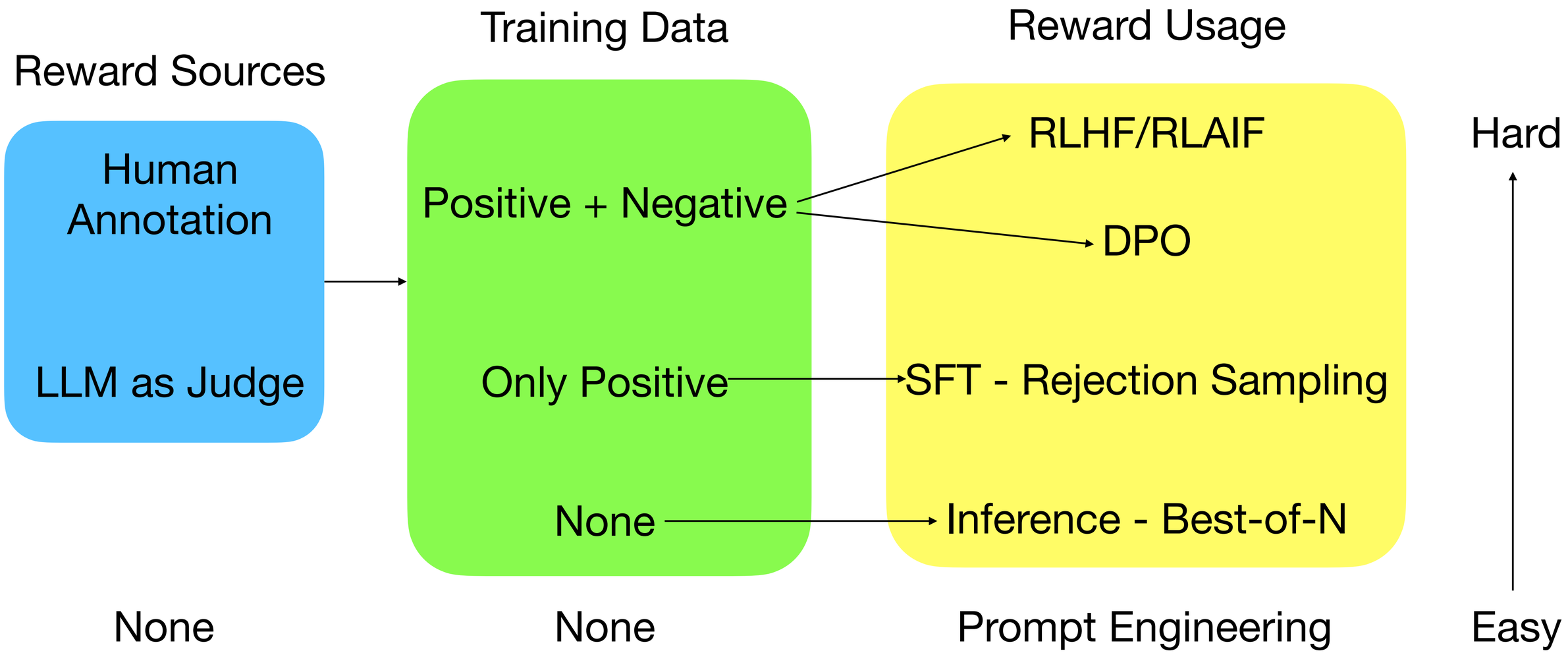


Figure 3: The hierarchy of RapidAPI (left) and the process of instruction generation (right).

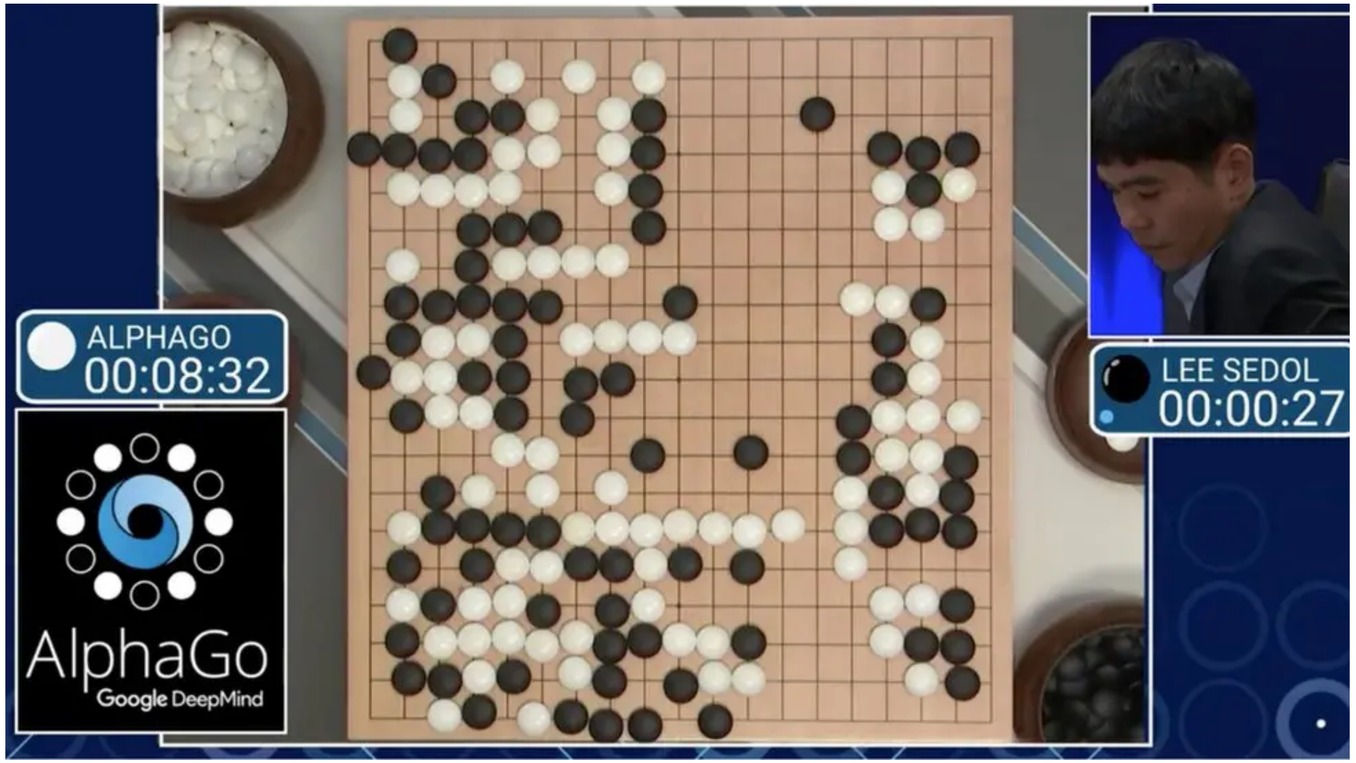
TOOLLLM: FACILITATING LARGE LANGUAGE MODELS TO MASTER 16000+ REAL-WORLD APIS (https://arxiv.org/pdf/2307.16789)

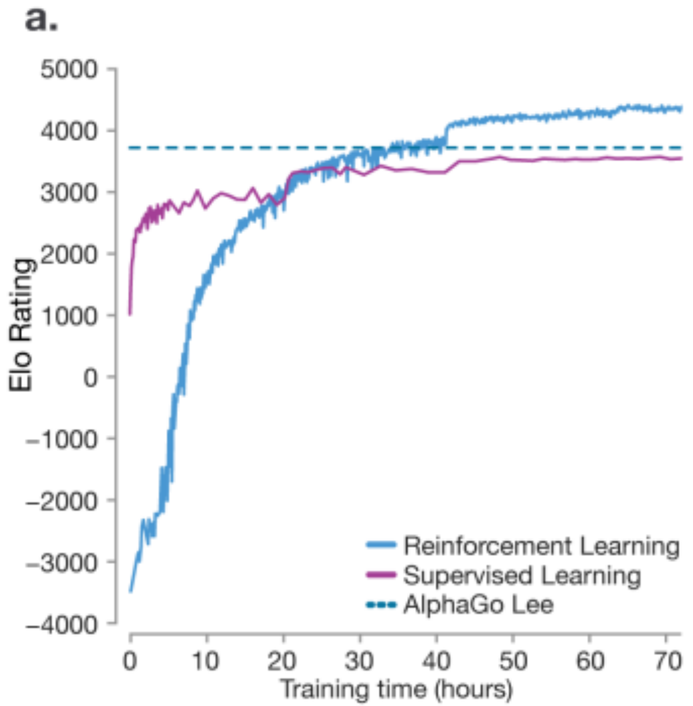
Alignment -> Distant Supervision



Do LRMs really Learn to Think like Humans?

AlphaGo and AlphaZero





Why can AlphaGo be better than top human players, but LRM cannot?

https://www.bbc.com/news/technology-35785875

https://www.science.org/doi/10.1126/science.aar6404?__cf_chl_tk=67Ig3VWHBOjaw3ybBhVGn2gbtd2QZ4UXUxDS21EBct4-1742742238-1.0.1.1-jx7XtvwAIV5eX51WMPAteOy04PT4tJF2e28qsLvXeTc