Reinforcement Learning for Post-Training LLMs

Slido: https://app.sli.do/event/3T9niVy29Fvve7iXwTJFAt

Yu Meng
University of Virginia
yumeng5@virginia.edu

Nov 12, 2025

Overview of Course Contents

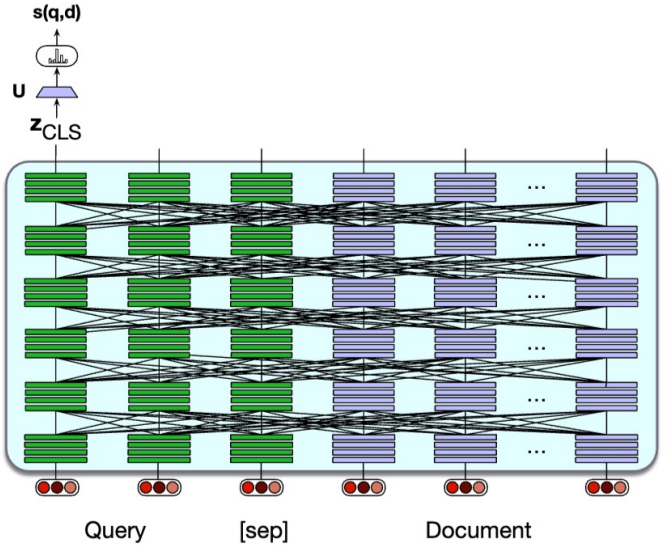
- Week 1: Logistics & Overview
- Week 2: N-gram Language Models
- Week 3: Word Senses, Semantics & Classic Word Representations
- Week 4: Word Embeddings
- Week 5: Sequence Modeling & Recurrent Neural Networks (RNNs)
- Week 6: Language Modeling with Transformers
- Week 8: Transformer and Pretraining
- Week 9: Large Language Models (LLMs) & In-context Learning
- Week 10: Knowledge in LLMs and Retrieval-Augmented Generation (RAG)
- Week 11: LLM Reasoning
- Week 12: Reinforcement Learning for Post-Training LLMs
- Week 13: LLM Agents + Course Summary
- Week 15 (after Thanksgiving): Project Presentations

(Recap) Dense Retrieval

- Motivation: sparse retrieval (e.g., TF-IDF) relies on the exact overlap of words between the query and document without considering semantic similarity
- Solution: use a language model to obtain (dense) distributed representations of query and document
- The retriever language model is typically a small text encoder model (e.g., BERT)
  - Retrieval is a natural language understanding task
  - Encoder-only models are more efficient than LLMs for this purpose
- Both query and document representations are computed by text encoders

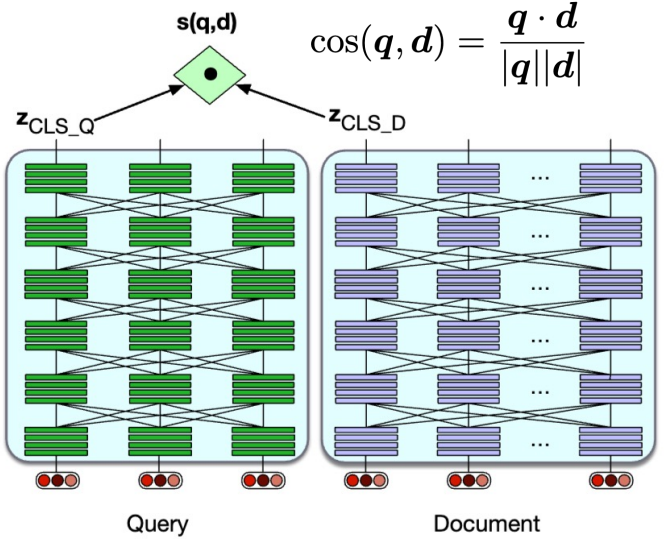
(Recap) Dense Retrieval: Cross-encoder

- Process query-document pairs together
- Relevance score produced directly by the model output
- (+) Capture intricate interactions between the query and the document
- (-) Not scalable to large retrieval corpus
- Good for small document sets



(Recap) Dense Retrieval: Bi-encoder

- Independently encode the query and the document using two separate (but often identical) encoder models
- Use cosine similarity between the query and document vectors as relevance score
- (+) Document vectors can be precomputed
- (-) Cannot capture query-document interactions
- Common choice for large-scale retrieval



(Recap) Evaluation of IR Systems

- Assume that each document returned by the IR system is either **relevant** to our purposes or **not relevant**
- Given a query, assume the system returns a set of ranked documents $\boldsymbol{T}$
  - A subset $\boldsymbol{R}$ of these are relevant (The remaining $\boldsymbol{N} = \boldsymbol{T} - \boldsymbol{R}$ is irrelevant)
  - There are $\boldsymbol{U}$ documents in the entire retrieval collection that are relevant to this query
- **Precision:** the fraction of the returned documents that are relevant

$$\text{Precision} = \frac{|R|}{|T|}$$

- **Recall:** the fraction of all relevant documents that are returned

$$\text{Recall} = \frac{|R|}{|U|}$$

(Recap) Precision & Recall @ k

- We hope to build a retrieval system that ranks the relevant documents higher
- Use precision & recall @ k (among the top-k items in the ranked list) to reflect this

Rank	Judgment	Precision_{Rank}	Recall_{Rank}
1	R	1.0	.11
2	N	.50	.11
3	R	.66	.22
4	N	.50	.22
5	R	.60	.33
6	R	.66	.44
7	N	.57	.44
8	R	.63	.55
9	N	.55	.55
10	N	.50	.55

Assume there are 9 total relevant documents in the retrieval corpus

Figure source: https://web.stanford.edu/~jurafsky/slp3/14.pdf

(Recap) Average Precision

Average precision (AP): mean of the precision values at the points in the ranked list where a relevant document is retrieved

Indicator function of whether the document is relevant

$$\text{AP} = \frac{1}{|R|} \sum_{k=1}^{|T|} (\text{Precision@}k \times \mathbb{1}(d_k \text{ is relevant}))$$

Rank	Judgment	Precision_{Rank}	Recall_{Rank}
1	R	1.0	.11
2	N	.50	.11
3	R	.66	.22
4	N	.50	.22
5	R	.60	.33
6	R	.66	.44
7	N	.57	.44
8	R	.63	.55
9	N	.55	.55
10	N	.50	.55

(Recap) RAG vs. Direct Prompting

- Prompting relies on LM's parametric knowledge to directly answer the question:

$P(w|$ `Q: Who wrote the book ‘‘The Origin of Species"? A:` $)$ prompt

- RAG prepends the set of retrieved passages to the question

Schematic of a RAG Prompt

```
retrieved passage 1

retrieved passage 2

...

retrieved passage n
```

Returned by the retriever

```
Based on these texts, answer this question: Q: Who wrote
the book ‘‘The Origin of Species"? A:
```

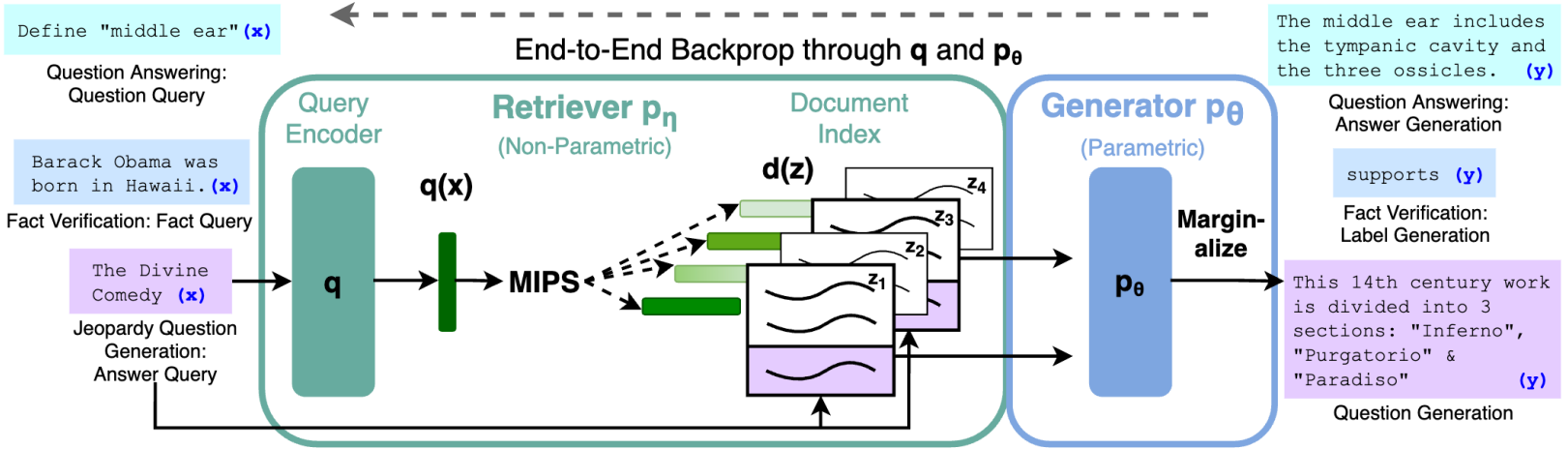
(Recap) A Latent Variable Model

The retrieved documents are treated as latent variables (z) for generation

$$p(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{z}\in\mathcal{D}} p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})$$

Retrieve document (z) based on query (x)

Generate answer (y) based on retrieved docs (z) and query (x)



(Recap) RAG-Sequence Model

- Use the same retrieved document to generate the complete sequence
- Treat the retrieved document as a single latent variable
- Marginalize to get the generation probability $p(\boldsymbol{y}|\boldsymbol{x})$ via a top-K approximation

$$p_{\text{RAG-sequence}}(\boldsymbol{y}|\boldsymbol{x}) \approx \sum_{\boldsymbol{z} \in \text{top-K}(p(\cdot|\boldsymbol{x}))} p_\eta(\boldsymbol{z}|\boldsymbol{x}) p_\theta(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) = \sum_{\boldsymbol{z} \in \text{top-K}(p(\cdot|\boldsymbol{x}))} p_\eta(\boldsymbol{z}|\boldsymbol{x}) \prod_{i=1}^{N} p_\theta(y_i|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}_{<i})$$

Top-K approximation
(only consider the top-K retrieved docs)

The same retrieved doc (z) is used to generate all tokens in the sequence

(Recap) RAG-Token Model

- Can use different retrieved documents to generate different tokens in a sequence
- Marginalization is performed for each generated token (rather than at sequence level)

$$p_{\text{RAG-token}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{N} p_\theta(y_i|\boldsymbol{x}, \boldsymbol{y}_{<i}) \approx \prod_{i=1}^{N} \sum_{\boldsymbol{z} \in \text{top-K}(p(\cdot|\boldsymbol{x}, \boldsymbol{y}_{<i}))} p_\eta(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}_{<i}) p_\theta(y_i|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}_{<i})$$

Different retrieved doc (z) can be used to generate different tokens in the sequence

(Recap) Reasoning: Overview

- **Reasoning** (rough definition): perform deductive, inductive, commonsense, or logical reasoning via generating or analyzing text with language models
- Deductive reasoning: draw specific conclusions from general principles or premises
  - E.g.: “All humans are mortal” + “Socrates is a human” => “Socrates is mortal”
- Inductive reasoning: make generalizations based on specific observations
  - E.g.: “The sun has risen in the east every day” => “The sun will rise in the east tomorrow”
- Commonsense reasoning: rely on world knowledge or commonsense understanding to make predictions or answer questions
  - E.g.: “If I drop a ball, what will happen?” => “It will fall”
- Mathematical/logical reasoning: follow specific rules or procedures to arrive at a correct answer
  - E.g.: “If 3 apples cost $6, how much do 5 apples cost?” => “$10”

(Recap) Frontier LLMs Are Reasoning Models

One unified system

GPT-5 is a unified system with a **smart, efficient model** that answers most questions, a **deeper reasoning model** (GPT-5 thinking) for harder problems, and a **real-time router** that quickly decides which to use based on conversation type, complexity, tool needs, and your explicit intent (for example, if you say "think hard about this" in the prompt). The router is continuously trained on real signals, including when users switch models, preference rates for responses, and measured correctness, improving over time. Once usage limits are reached, a mini version of each model handles remaining queries. In the near future, we plan to integrate these capabilities into a single model.

Screenshot source: https://openai.com/index/introducing-gpt-5/

(Recap) Chain-of-thought (CoT) Prompting

- **Chain-of-thought (CoT)**: the model breaks down complex problems into a step-by-step reasoning process
- Instead of directly providing an answer to a question or task, the model is prompted to explain its reasoning or thought process in a logical sequence

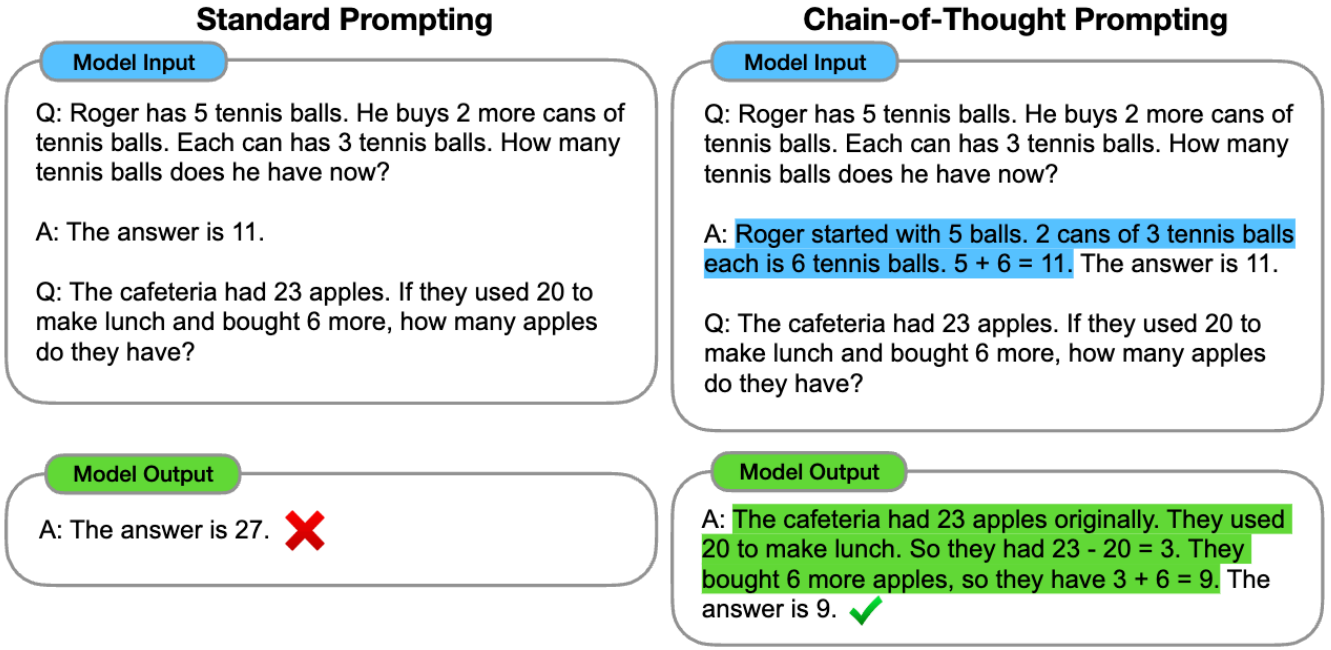
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

Paper: https://arxiv.org/pdf/2201.11903

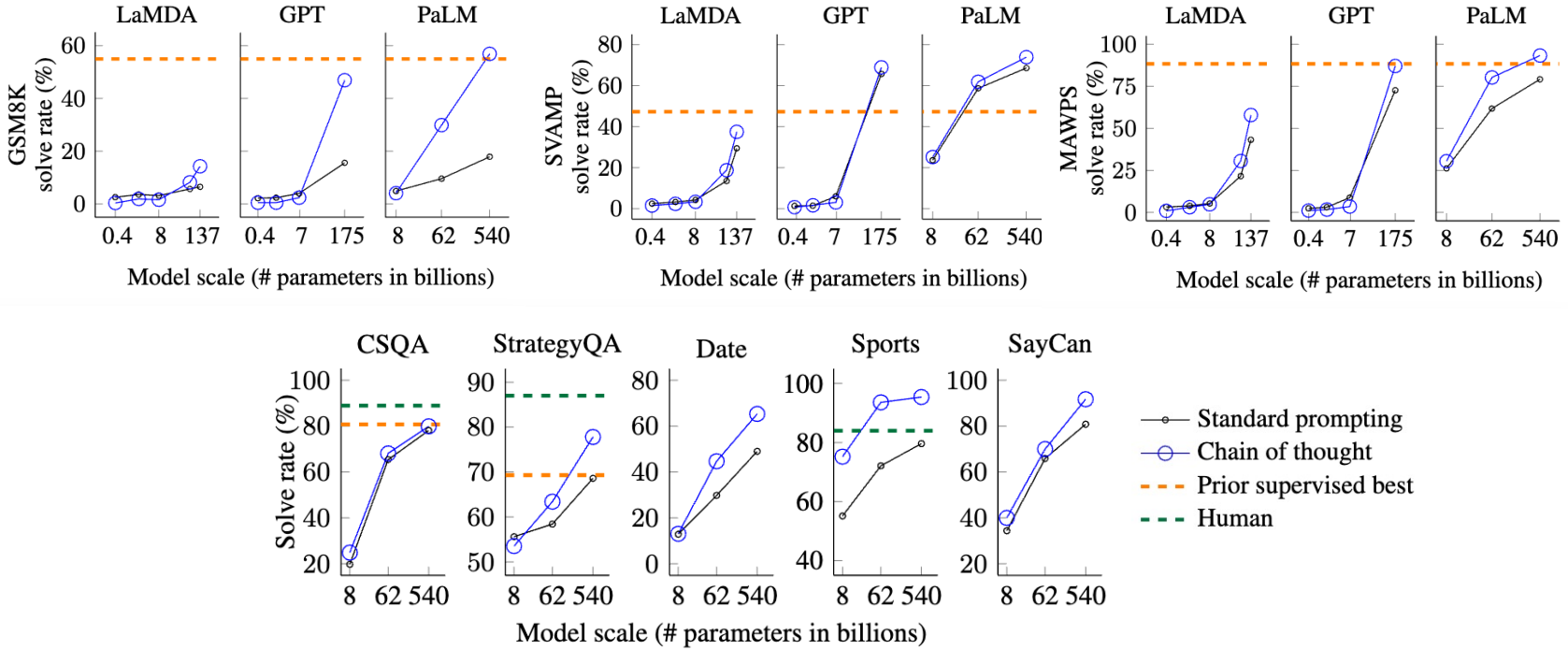
(Recap) Standard Prompting vs. CoT Prompting



Paper: https://arxiv.org/pdf/2201.11903

(Recap) Standard vs. CoT Prompting Performance

CoT prompting is especially effective for large models



CoT Can Be Triggered Zero-shot

Just add "Let's think step by step" at the beginning of the answer

(a) Few-shot

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

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: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

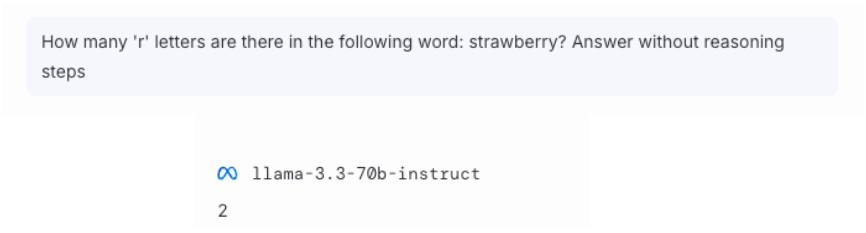
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Paper: https://arxiv.org/pdf/2205.11916

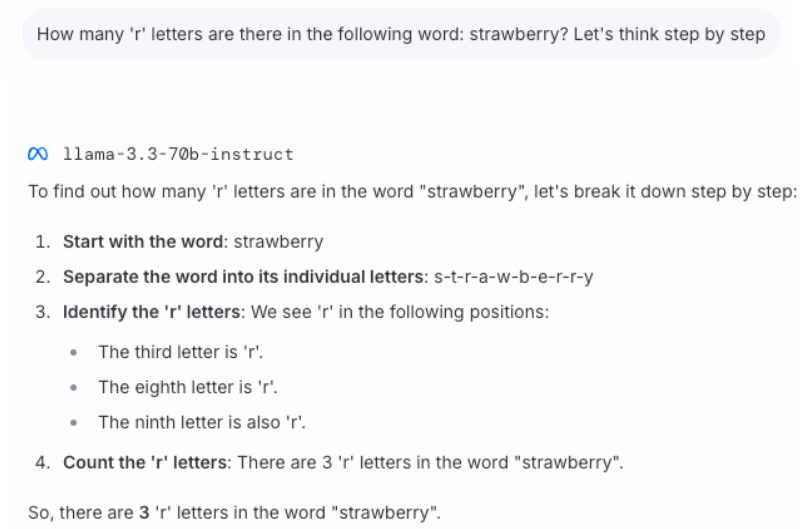
CoT Demo

No-CoT prompt: How many 'r' letters are there in the following word: strawberry? Answer without reasoning steps



Wrong result

CoT prompt: How many 'r' letters are there in the following word: strawberry? Let's think step by step

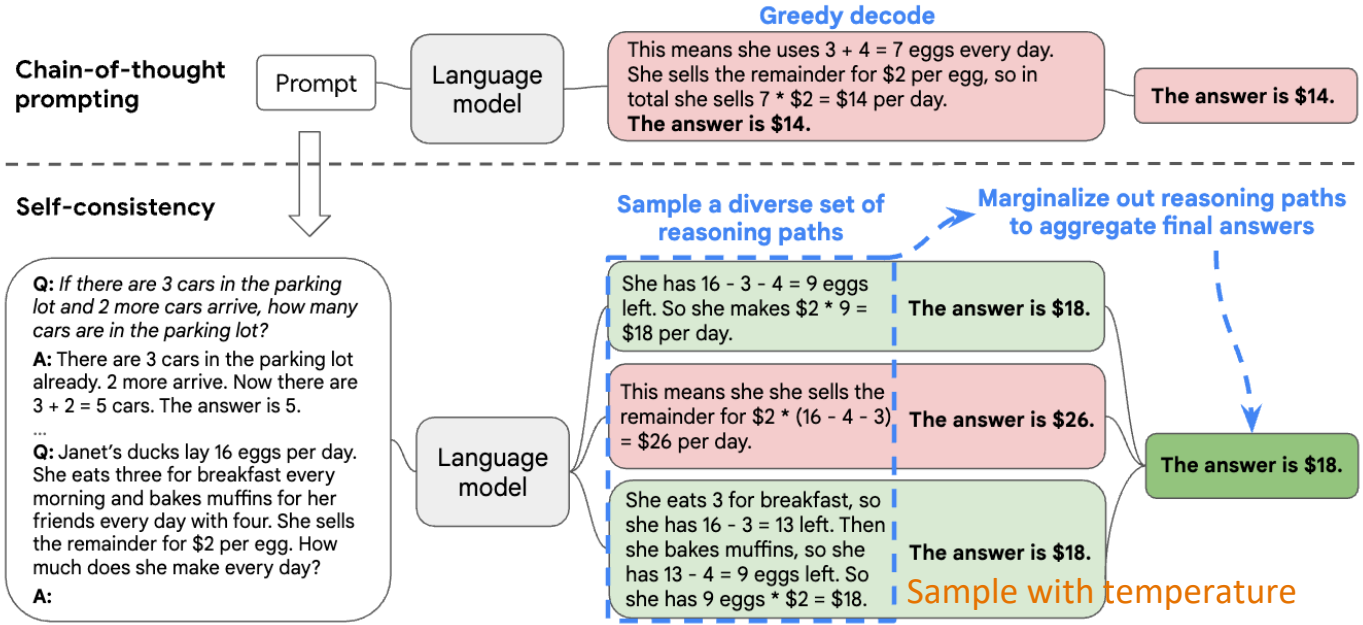


Correct result

Figure source: https://lmarena.ai

Self-consistency CoT

Intuition: if multiple different ways of thinking lead to the same answer, one has greater confidence that the final answer is correct



Paper: https://arxiv.org/pdf/2203.11171

Further Reading on LLM Reasoning

- Least-to-Most Prompting Enables Complex Reasoning in Large Language Models [Zhou et al., 2022]
- Large Language Models Can Self-Improve [Huang et al., 2022]
- Tree of Thoughts: Deliberate Problem Solving with Large Language Models [Yao et al., 2023]
- Let's Verify Step by Step [Lightman et al., 2023]

Agenda

- Reasoning Benchmarks
- Reinforcement Learning with Verifiable Rewards (RLVR)
- Introduction to LLM Alignment
- Instruction Tuning

Grade School Math (GSM8K)

8.5K high quality grade school math problems created by human problem writers

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

Paper: https://arxiv.org/pdf/2110.14168

MATH

12.5K challenging competition mathematics problems

Problem: Suppose a and b are positive real numbers with $a > b$ and $ab = 8$. Find the minimum value of $\frac{a^2+b^2}{a-b}$.

Ground truth solution: We can write $\frac{a^2+b^2}{a-b} = \frac{a^2+b^2-2ab+16}{a-b} = \frac{(a-b)^2+16}{a-b} = a - b + \frac{16}{a-b}$. By AM-GM, $a - b + \frac{16}{a-b} \geq 2\sqrt{(a-b) \cdot \frac{16}{a-b}} = 8$. Equality occurs when $a - b = 4$ and $ab = 8$. We can solve these equations to find $a = 2\sqrt{3} + 2$ and $b = 2\sqrt{3} - 2$. Thus, the minimum value is $\boxed{8}$.

Problem: Right ΔABC has legs measuring 8 cm and 15 cm. The triangle is rotated about one of its legs. What is the number of cubic centimeters in the maximum possible volume of the resulting solid? Express your answer in terms of π.

Ground truth solution: If the triangle is rotated about the shorter leg, then the radius is the longer leg and the height is the shorter leg, and the volume is $\frac{1}{3} \cdot (15^2\pi)(8) = 600\pi$ cubic centimeters. If the triangle is rotated about the longer leg, then the radius is the shorter leg and the height is the longer leg, and the volume is $\frac{1}{3}(8^2\pi)(15)$, which is $\frac{8}{15}$ of the volume we found earlier. So, the maximum possible volume is $\boxed{600\pi}$ cubic centimeters.

Paper: https://arxiv.org/pdf/2103.03874

AI2 Reasoning Challenge (ARC)

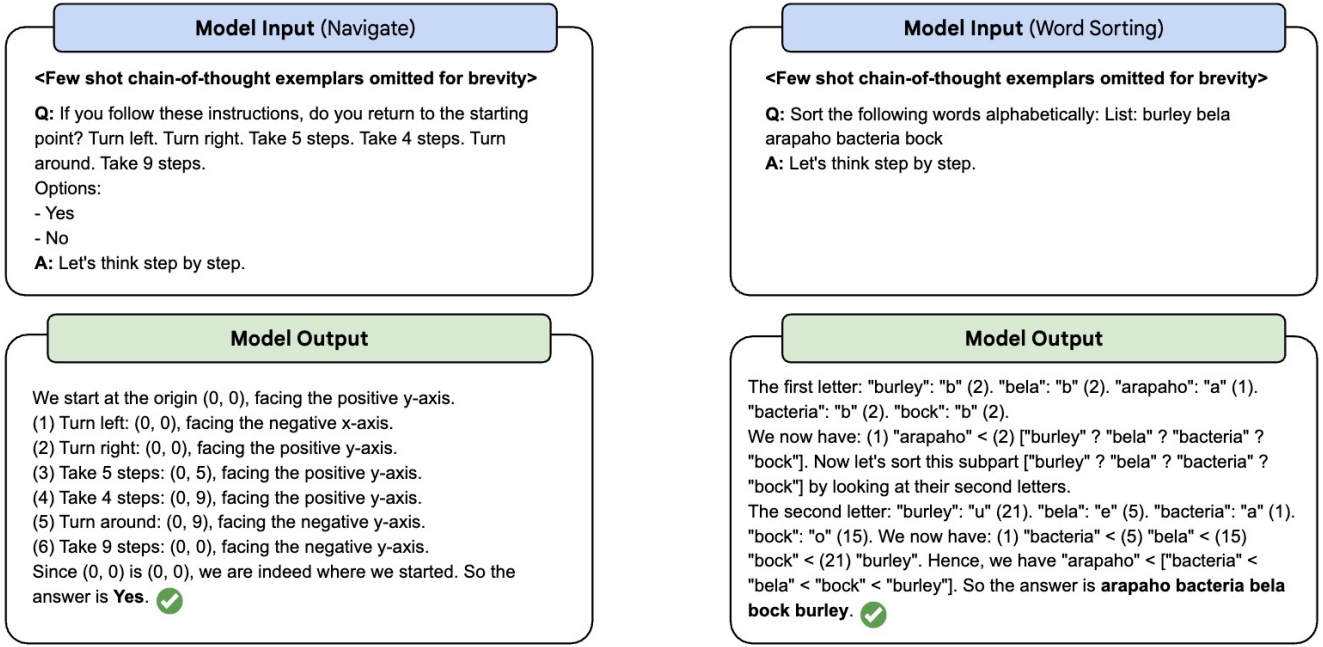
~8K natural science questions on commonsense knowledge/reasoning

Reasoning Type	Example
Question logic	Which item below is **not** made from a material grown in nature? (A) a cotton shirt (B) a wooden chair (C) a plastic spoon (D) a grass basket
Linguistic Matching	Which of the following best describes a mineral? (A) the main nutrient in all foods (B) a type of grain found in cereals (C) a natural substance that makes up rocks (D) the decomposed plant matter found in soil
Multihop Reasoning	Which property of a mineral can be determined just by looking at it? (A) luster (B) mass (C) weight (D) hardness
Comparison	Compared to the Sun, a red star most likely has a greater (A) volume. (B) rate of rotation. (C) surface temperature. (D) number of orbiting planets
Algebraic	If a heterozygous smooth pea plant (Ss) is crossed with a homozygous smooth pea plant (SS), which are the possible genotypes the offspring could have? (A) only SS (B) only Ss (C) Ss or SS (D) ss or SS
Hypothetical / Counterfactual	If the Sun were larger, what would most likely also have to be true for Earth to sustain life? (A) Earth would have to be further from the Sun. (B) Earth would have to be closer to the Sun. (C) Earth would have to be smaller. (D) Earth would have to be larger.
Explanation / Meta-reasoning	Why can steam be used to cook food? (A) Steam does work on objects. (B) Steam is a form of water. (C) Steam can transfer heat to cooler objects. (D) Steam is able to move through small spaces.

Paper: https://arxiv.org/pdf/1803.05457

BIG-Bench Hard (BBH)

23 challenging tasks covering a wide range of reasoning (e.g. arithmetic, logical, spatial…)



Paper: https://arxiv.org/pdf/2210.09261

American Invitational Mathematics Examination (AIME)

High school math competition, where each answer is an integer from 000 to 999

Problem

Let A be the set of positive integer divisors of 2025. Let B be a randomly selected subset of A. The probability that B is a nonempty set with the property that the least common multiple of its elements is 2025 is $\frac{m}{n}$, where m and n are relatively prime positive integers. Find $m+n$.

Solution 1

We split into different conditions:

Note that the numbers in the set need to have a least common multiple of 2025, so we need to ensure that the set has at least 1 number that is a multiple of 3^4 and a number that is a multiple of 5^2.

Multiples of 3^4: $81, 405, 2025$

Multiples of 5^2: $25, 75, 225, 675, 2025$

If the set B contains 2025, then all of the rest 14 factors is no longer important. The valid cases are 2^{14}.

If the set B doesn't contain 2025, but contains 405, we just need another multiple of 5^2. It could be 1 of them, 2 of them, 3 of them, or 4 of them, which has $2^4-1=15$ cases. Excluding $2025, 405, 25, 75, 225, 675,$ the rest 9 numbers could appear or not appear. Therefore, this case has a valid case of $15 \cdot 2^9$.

If set B doesn't contain 2025 nor 405, it must contain 81. It also needs to contain at least 1 of the multiples from 5^2, where it would be $15 \cdot 2^8$.

The total valid cases are $2^{14}+15 \cdot (2^9+2^8)$, and the total cases are 2^{15}.

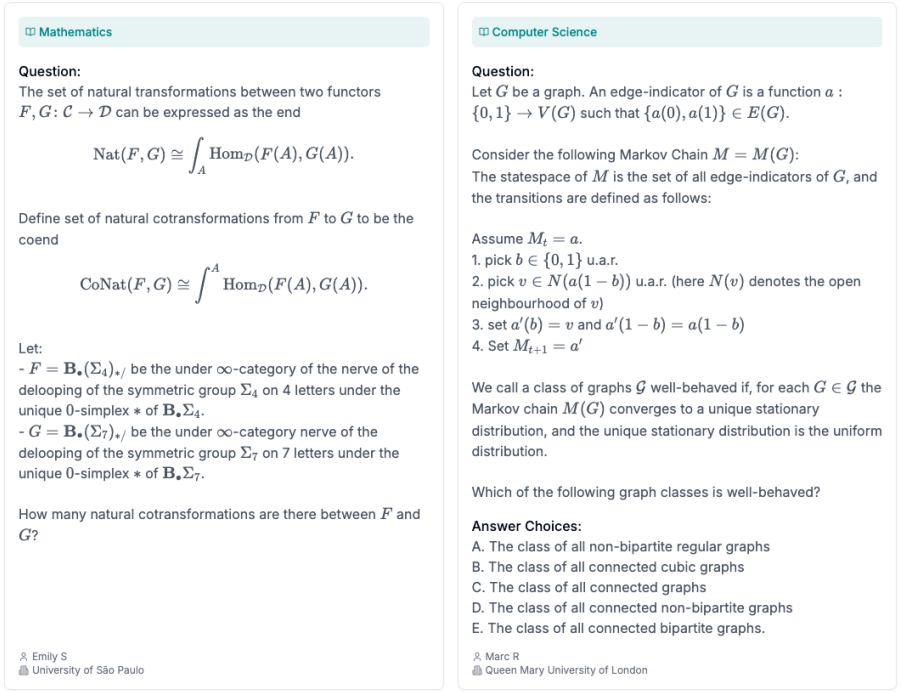
The answer is $\frac{2^8 \cdot (64+30+15)}{2^8 \cdot 2^7} = \frac{109}{128}$.

Desired answer: $109+128=\boxed{237}$.

Figure source: https://artofproblemsolving.com/wiki/index.php?title=2025_AIME_II_Problems/Problem_7

Humanity's Last Exam (HLE)

2,500 challenging questions across over a hundred subjects (created by experts)



Mathematics

Question:
The set of natural transformations between two functors $F, G \colon \mathcal{C} \to \mathcal{D}$ can be expressed as the end

$$\mathrm{Nat}(F,G) \cong \int_A \mathrm{Hom}_{\mathcal{D}}(F(A), G(A)).$$

Define set of natural cotransformations from F to G to be the coend

$$\mathrm{CoNat}(F,G) \cong \int^A \mathrm{Hom}_{\mathcal{D}}(F(A), G(A)).$$

Let:
- $F = \mathbf{B}_\bullet(\Sigma_4)_{*/}$ be the under ∞-category of the nerve of the delooping of the symmetric group Σ_4 on 4 letters under the unique 0-simplex $*$ of $\mathbf{B}_\bullet\Sigma_4$.
- $G = \mathbf{B}_\bullet(\Sigma_7)_{*/}$ be the under ∞-category nerve of the delooping of the symmetric group Σ_7 on 7 letters under the unique 0-simplex $*$ of $\mathbf{B}_\bullet\Sigma_7$.

How many natural cotransformations are there between F and G?

Emily S
University of São Paulo

Computer Science

Question:
Let G be a graph. An edge-indicator of G is a function $a : \{0,1\} \to V(G)$ such that $\{a(0), a(1)\} \in E(G)$.

Consider the following Markov Chain $M = M(G)$:
The statespace of M is the set of all edge-indicators of G, and the transitions are defined as follows:

Assume $M_t = a$.
1. pick $b \in \{0,1\}$ u.a.r.
2. pick $v \in N(a(1-b))$ u.a.r. (here $N(v)$ denotes the open neighbourhood of v)
3. set $a'(b) = v$ and $a'(1-b) = a(1-b)$
4. Set $M_{t+1} = a'$

We call a class of graphs $\mathcal{G}$ well-behaved if, for each $G \in \mathcal{G}$ the Markov chain $M(G)$ converges to a unique stationary distribution, and the unique stationary distribution is the uniform distribution.

Which of the following graph classes is well-behaved?

Answer Choices:
A. The class of all non-bipartite regular graphs
B. The class of all connected cubic graphs
C. The class of all connected graphs
D. The class of all connected non-bipartite graphs
E. The class of all connected bipartite graphs.

Marc R
Queen Mary University of London

Website: https://agi.safe.ai/

Agenda

- Reasoning Benchmarks
- Reinforcement Learning with Verifiable Rewards (RLVR)
- Introduction to LLM Alignment
- Instruction Tuning

OpenAI's o1 (2024/09)

OpenAI released o1 (a reasoning model) in 2024, with remarkable performance on AIME

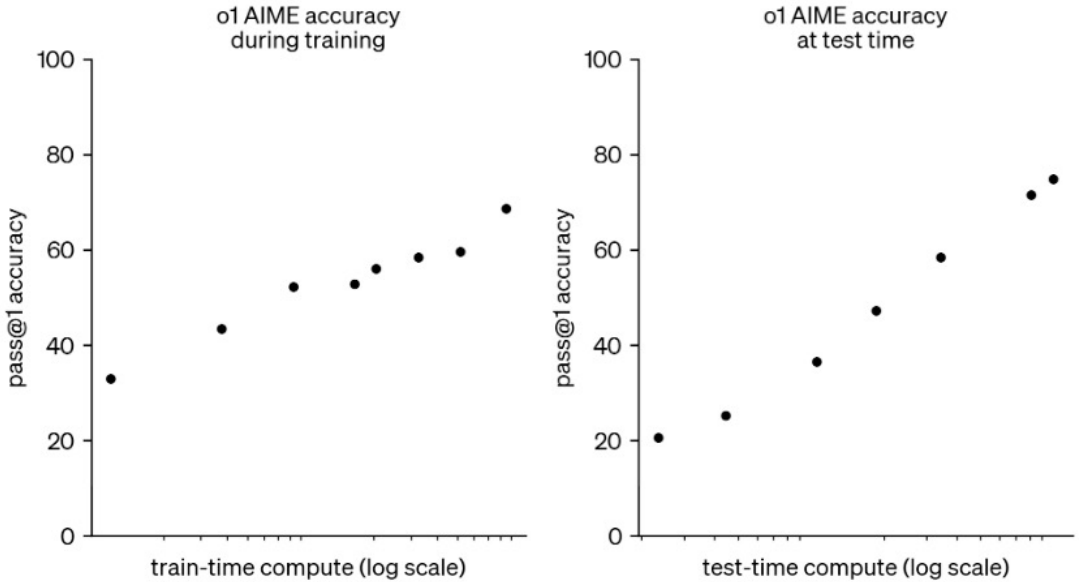


Figure source: https://openai.com/index/learning-to-reason-with-llms/

DeepSeek-R1 (2025/01)

- Open-source reproduction of OpenAI's o1
- During training, DeepSeek-R1 naturally learns to solve reasoning tasks with more thinking time

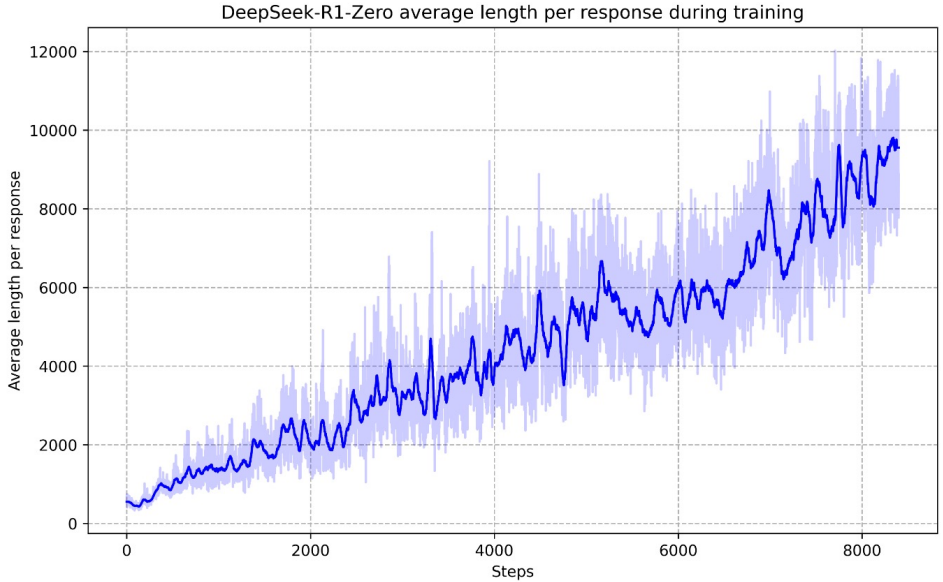


Figure source: https://arxiv.org/pdf/2501.12948

The "Aha Moment" of DeepSeek-R1

The model is not explicitly taught on how to solve a problem, but rather autonomously develops advanced problem-solving strategies (e.g., reevaluating its initial approach)

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: $\cdots$
$\ldots$

Figure source: https://arxiv.org/pdf/2501.12948

Scaling Test-Time Compute

- **Test-time Scaling**: increase the compute used at inference time (e.g., generating more tokens) to generate a higher-quality answer
- **Self-consistency** (majority voting):
  - Generate multiple responses to the same prompt
  - Use majority voting to select the best answer
- **Long CoT**:
  - Longer reasoning chains (think more thoroughly) ⇢ **OpenAI o1 & DeepSeek-R1**
- **Beam Search/Tree Search**:
  - Explore multiple reasoning paths simultaneously
  - Backtrack when hitting dead ends
  - Prune bad branches
- **Iterative Refinement:**
  - Generate initial response and then improve it iteratively

Reinforcement Learning with Verifiable Rewards (RLVR)

- The (major) post-training recipe of OpenAI's o1 & DeepSeek-R1
- **RLVR**:
  - Fine-tune the policy model (LLM) using **reinforcement learning**
  - The LLM receives a reward when its generated responses are verifiably correct

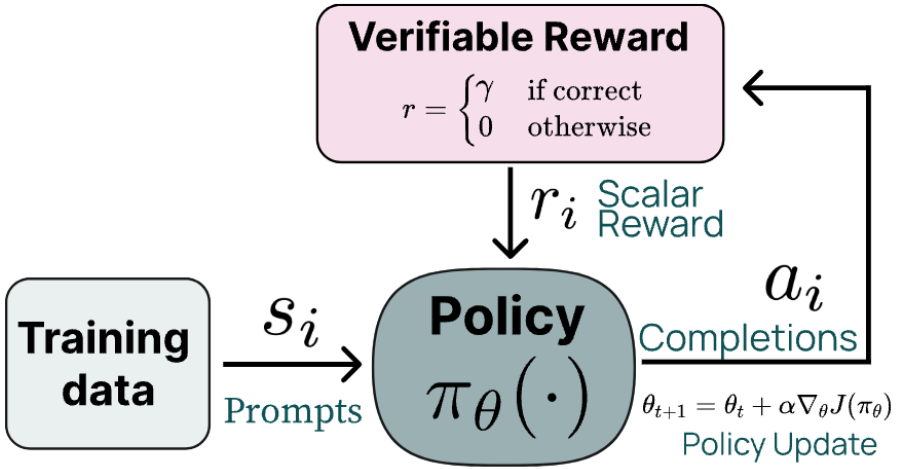


Figure source: https://arxiv.org/pdf/2411.15124

Why RLVR?

- **Data scalability**
  - Supervised learning requires human annotators to create the full correct responses
  - RLVR only requires automatic verifiers to grade the responses
- **Imitation vs. optimization**
  - Supervised learning forces models to imitate human reasoning steps, which may be suboptimal
  - RLVR optimizes directly for correct final answers, allowing the model to discover its own efficient reasoning paths
- **Distribution mismatch**
  - Supervised learning (where data are created by humans) might cause a discrepancy from the model's own distribution
  - RLVR learns from the model's own generated sequences, matching the inference distribution

RLVR Setup

- Generate responses using the LLM (the policy model)
- Assign rewards to the generated responses
- Maximize the expected reward

$$\max_{\theta} \mathbb{E}_{y \sim p_\theta(\cdot|x)} \left[r_{\text{RLVR}}(x, y) \right]$$

LLM output probability

reward of RLVR (binary)

Optimization with Reinforcement Learning (RL)

- Why reinforcement learning:
  - No supervised data available
  - Encourage the model to explore new possibilities (generations) guided by the rewards
- Optimization: policy gradient methods
  - Optimize the policy (LLM) by adjusting the parameters in the direction that increases expected rewards
- REINFORCE (simplest policy gradient method):

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a|s) R$$

- R: cumulative reward
- α: step size
- π_θ: policy model (LLM)
- a: action (generating the response)
- s: state (user prompt + conversation history)

Overview: Proximal Policy Optimization (PPO)

- A more advanced policy gradient method that improves stability and efficiency
- Clipped mechanism: PPO uses a clipped surrogate objective to ensure that policy updates are not too large, which helps maintain stability
- Advantage estimation: PPO uses Generalized Advantage Estimation (GAE) to reduce variance in the advantage estimates with a critic model, improving learning efficiency

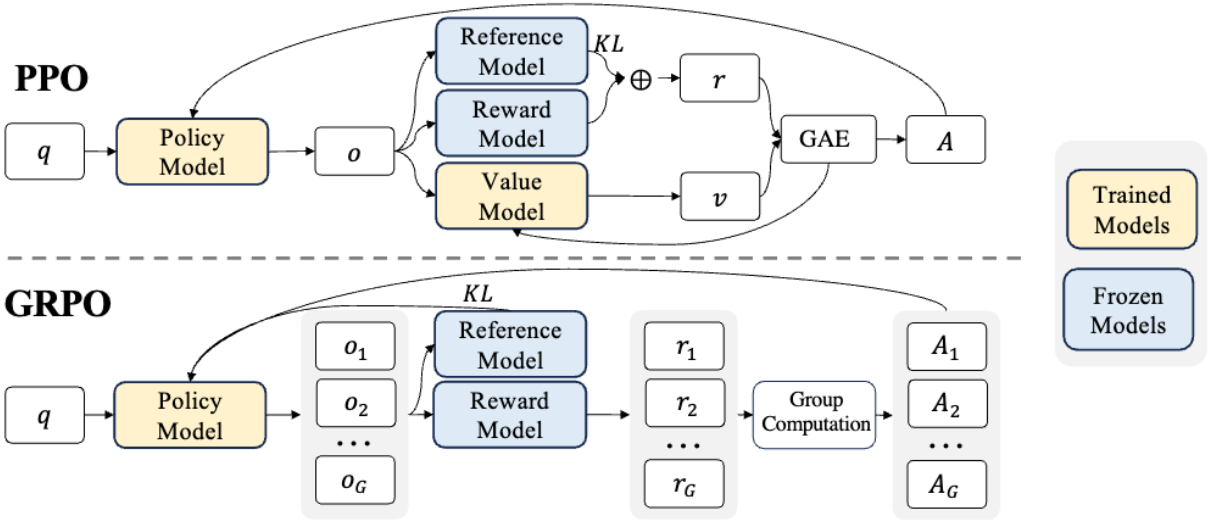
Proximal Policy Optimization Algorithms

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov
OpenAI
{joschu, filip, prafulla, alec, oleg}@openai.com

Paper: https://arxiv.org/pdf/1707.06347

Overview: Group Relative Policy Optimization (GRPO)

- A variant of PPO that foregoes the critic model
- Advantage estimation: for each question, GRPO samples a group of outputs and use the comparative rewards to estimate advantage



Paper: https://arxiv.org/pdf/2402.03300

Further Reading on RLVR

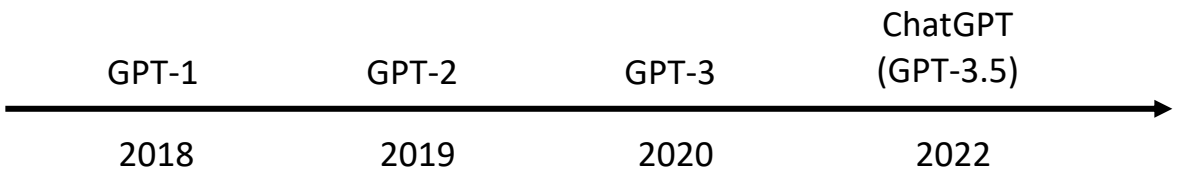
- ProRL: Prolonged Reinforcement Learning Expands Reasoning Boundaries in Large Language Models [Liu et al., 2025]
- DAPO: An Open-Source LLM Reinforcement Learning System at Scale [Yu et al., 2025]
- The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning [Zhu et al., 2025]
- Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? [Yue et al., 2025]

Agenda

- Reasoning Benchmarks
- Reinforcement Learning with Verifiable Rewards (RLVR)
- Introduction to LLM Alignment
- Instruction Tuning

The Evolution of GPT Models: ChatGPT

- GPT-1: decoder-only Transformer pretraining
- GPT-2: language model pretraining is multi-task learning
- GPT-3: scaling up & in-context learning
- ChatGPT: language model alignment



Overview: Language Model Alignment

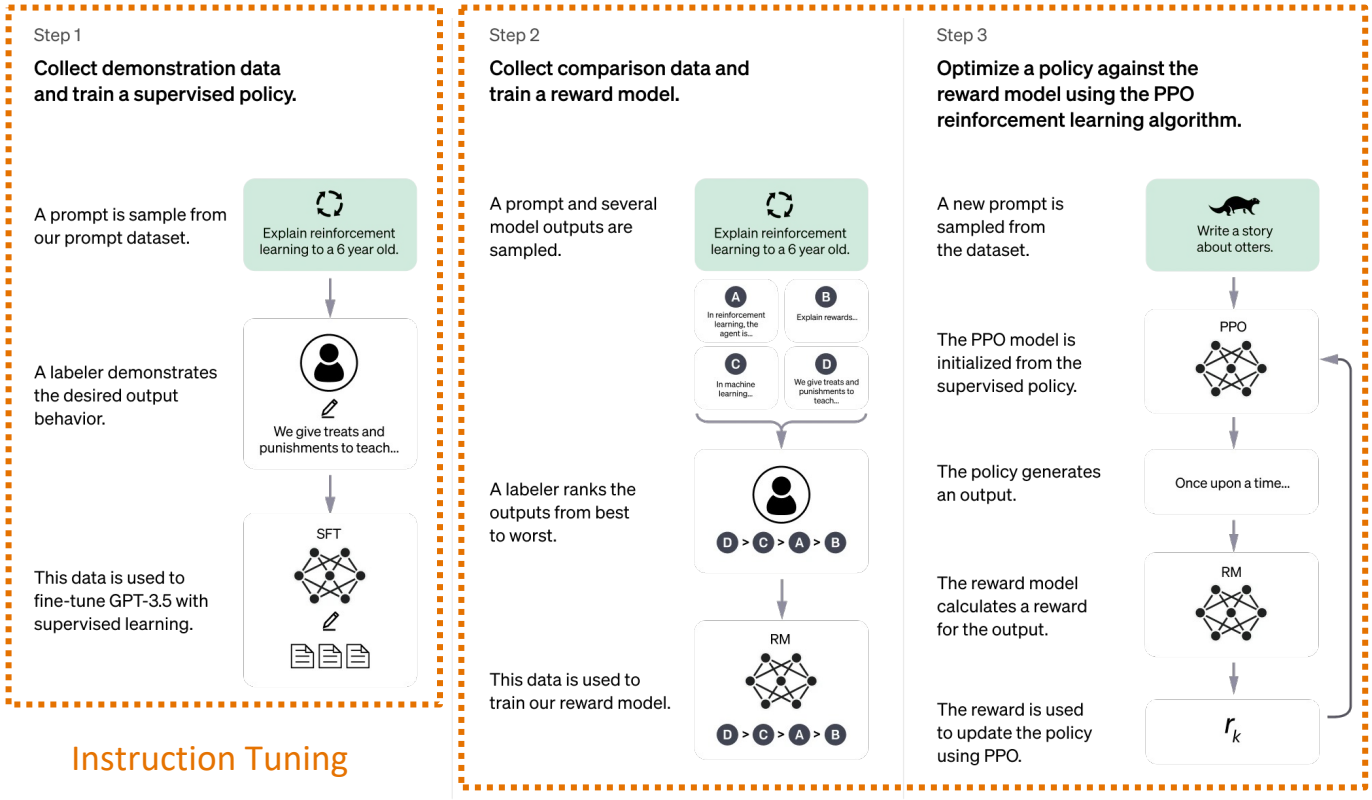
- Ensure language models behaviors are aligned with human values and intent for general tasks/applications
- "HHH" criteria (Askell et al. 2021):
  - **Helpful**: Efficiently perform the task requested by the user
  - **Honest**: Give accurate information & express uncertainty
  - **Harmless**: Avoid offensive/discriminatory/biased outputs

Paper: https://arxiv.org/pdf/2112.00861

Language Model Alignment: Post-training

- Pretrained language models are **not** aligned
- Objective mismatch
  - Pretraining is to predict the next word in a sentence
  - Does not involve understanding human intent/values
- Training data bias
  - Text from the internet can contain biased, harmful, or misleading information
  - LMs don't distinguish between good and bad behavior in training data
- (Over-)generalization issues
  - LMs' generalization can lead to outputs that are inappropriate in specific contexts
  - Might not align with intended ethics/honesty standard

Language Model Alignment Techniques



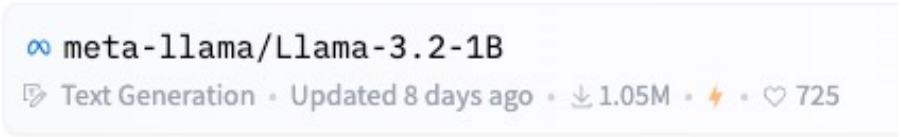
Instruction Tuning

Reinforcement Learning from Human Feedback (RLHF)

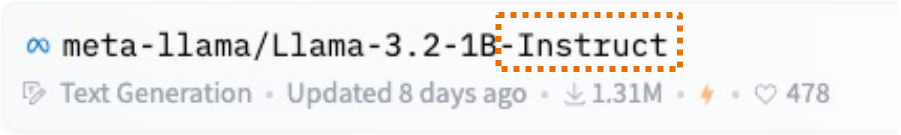
Figure source: https://openai.com/index/chatgpt/

Overview: Instruction Tuning

- Train an LM using a diverse set of tasks
  - Each task is framed as an **instruction** followed by an example of the desired output
  - The goal is to teach the model to follow specific instructions (human intent) effectively
- The resulting model can perform a variety of tasks **zero-shot** (w/o requiring in-context demonstrations)
- The instructions can also be in chat format – tuning an LM into a chatbot



Pretrained (base) model



Instruction-tuned (post-trained) model

Models: https://huggingface.co/meta-llama

Overview: RLHF

- Human feedback collection
  - Generate multiple responses using the model given the same prompt
  - Human evaluators rank responses of the model based on helpfulness/honesty/safety…
- Reward model training
  - A reward model is trained on human feedback data to predict the quality of responses
  - Higher reward = more preferred by human evaluators
- Policy optimization
  - Use reinforcement learning algorithms to further train the LM to maximize the reward predicted by the reward model
  - Encourage the model to produce outputs that align better with human preferences

Agenda

- Reasoning Benchmarks
- Reinforcement Learning with Verifiable Rewards (RLVR)
- Introduction to LLM Alignment
- Instruction Tuning

Instruction Tuning: Introduction

- **Setting**: fine-tune LLMs with task-specific instructions on diverse tasks
- **Goal**: enable LLM to better understand user prompts and generalize to a wide range of (unseen) tasks **zero-shot**

FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS

Jason Wei*, Maarten Bosma*, Vincent Y. Zhao*, Kelvin Guu*, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le

Google Research

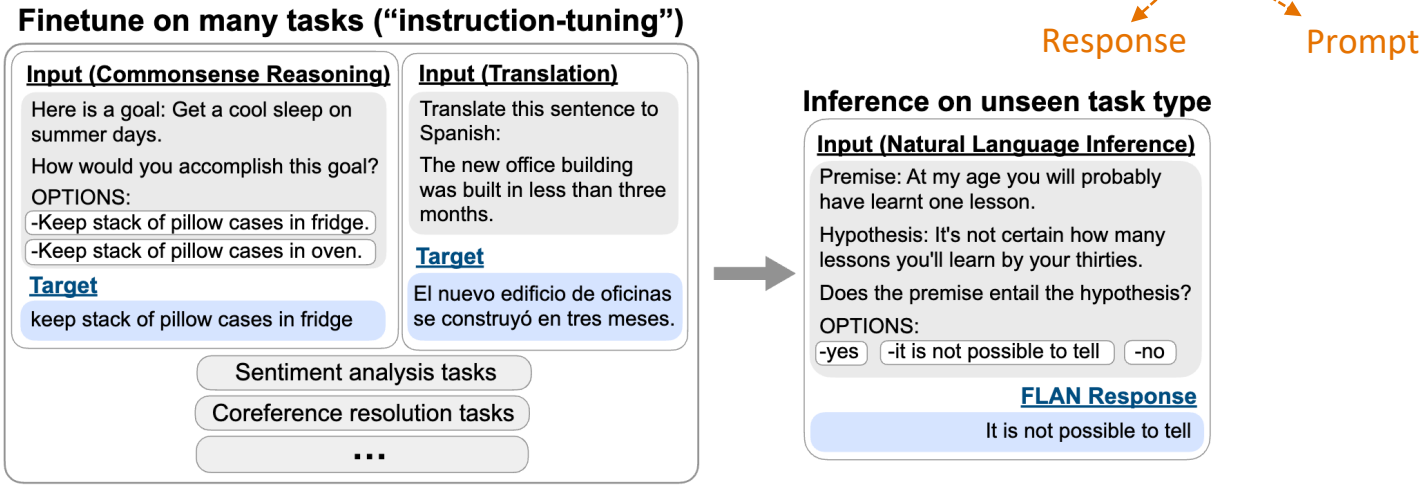
Paper: https://arxiv.org/pdf/2109.01652

Instruction Tuning: Method

- **Input**: task description
- **Output**: expected response or solution to the task
- Train LLMs to generate response tokens given prompts $\min_{\theta} -\log p_{\theta}(\boldsymbol{y}|\boldsymbol{x})$

Response

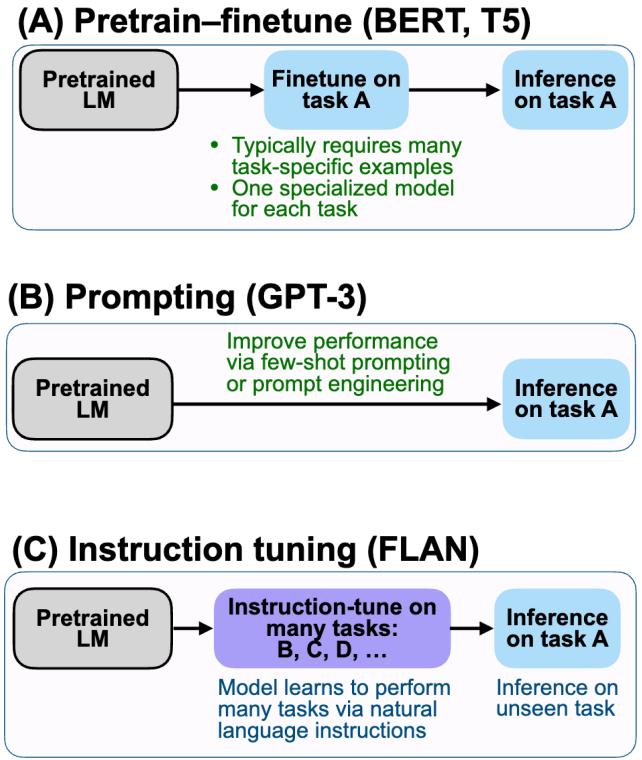
Prompt



Paper: https://arxiv.org/pdf/2109.01652

Instruction Tuning vs. Other Paradigms

- Task-specific fine-tuning does not enable generalization across multiple tasks
- In-context learning requires few-shot demonstrations
- Instruction tuning enables zero-shot cross task generalization

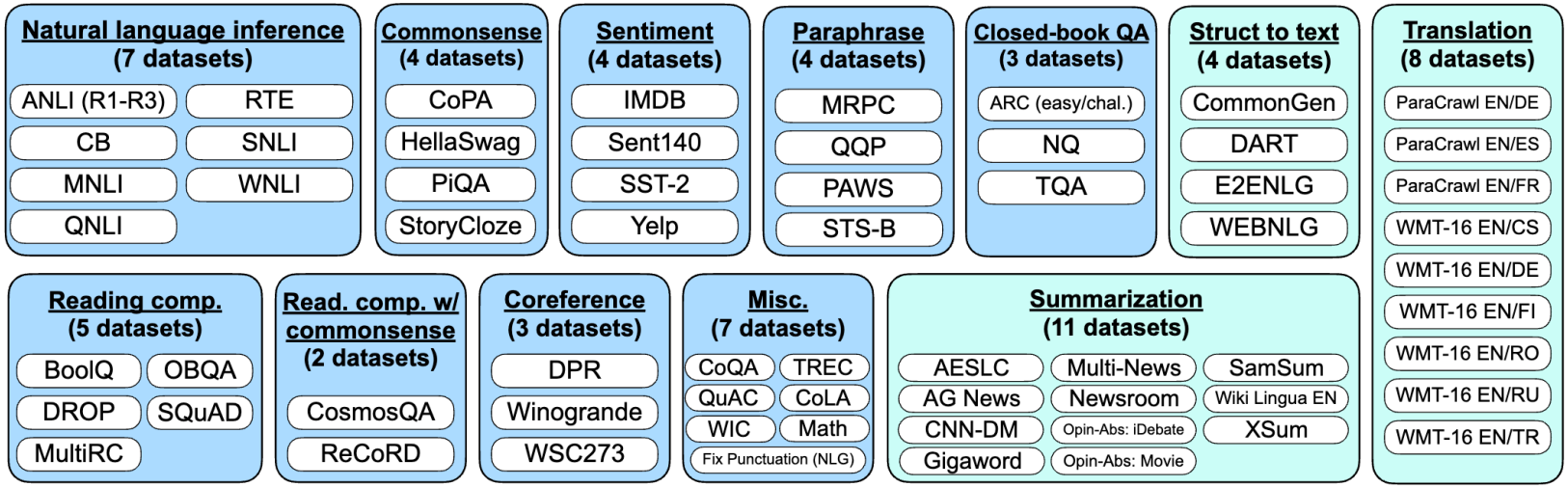


Instruction Tuning vs. Pretraining

- Both instruction tuning and pretraining are **multi-task** learning paradigms
- Supervision
  - Pretraining: self-supervised learning (raw data w/o human annotation)
  - Instruction tuning: supervised learning (human annotated responses)
- Task format
  - Pretraining: tasks are implicit (predicting next tokens)
  - Instruction tuning: tasks are explicit (defined using natural language instructions)
- Goal
  - Pretraining: teach LMs a wide range of linguistic patterns & general knowledge
  - Instruction tuning: teach LMs to follow specific instructions and perform a variety of tasks

FLAN: Collection of Instruction Tuning Datasets

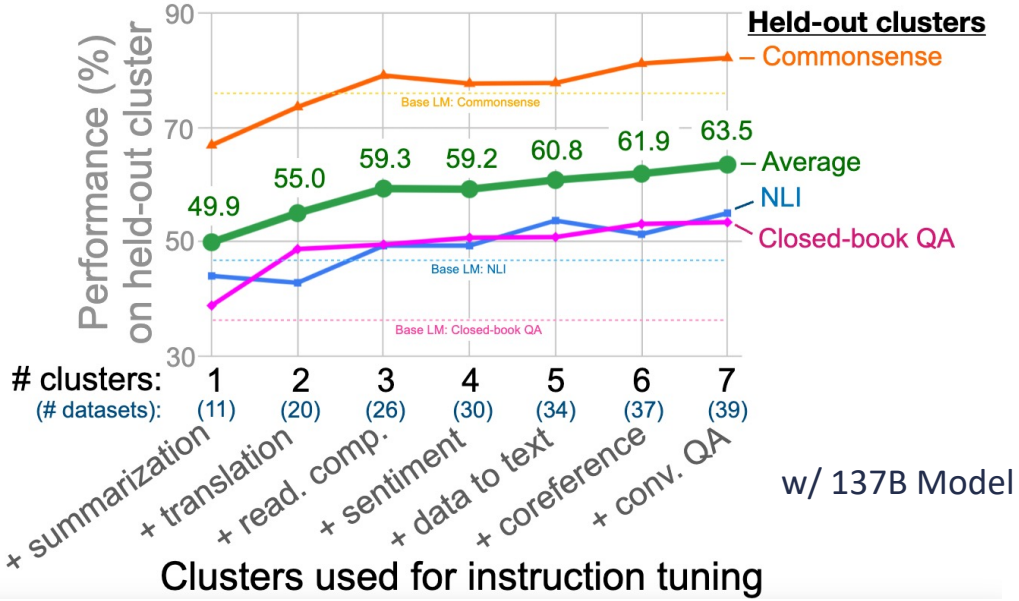
62 datasets (12 task clusters) covering a wide range of understanding + generation tasks



Paper: https://arxiv.org/pdf/2109.01652

Generalization Improves with More Clusters

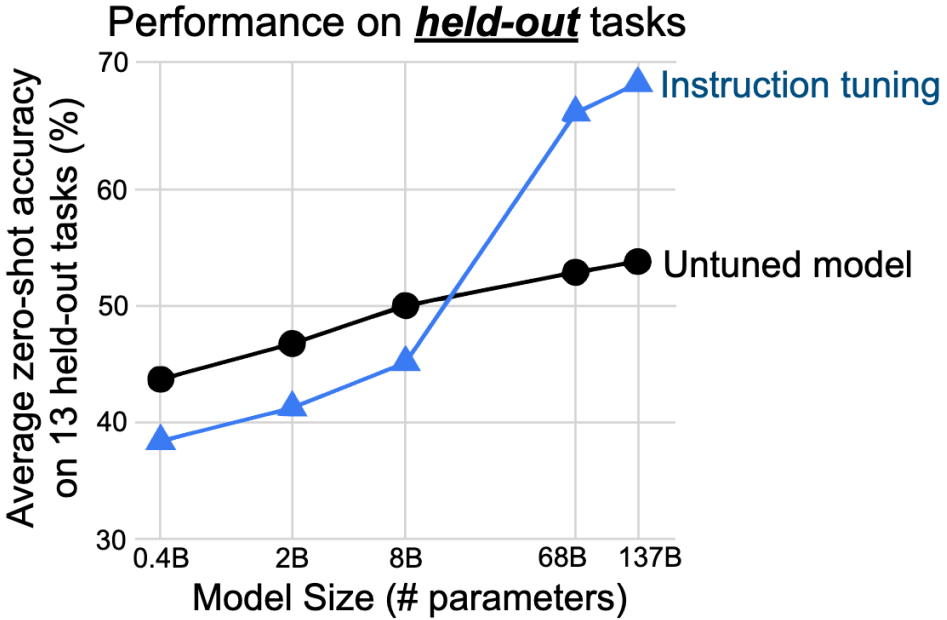
- Held out three clusters from instruction tuning: Commonsense, NLI, Closed-book QA
- More clusters and tasks used in instruction tuning => better generalization to unseen clusters



w/ 137B Model

Instruction Tuning with Different Model Sizes

- Instruction tuning can hurt small model (< 8B) generalization
- Instruction tuning substantially improves generalization for large models



Chat-style Instruction Tuning

- Instruction tuning can also be used to build chatbots for multi-turn dialogue
- Instructions may not correspond strictly to one NLP task, but mimic a human-like dialogue
- Multi-turn instruction tuning training data example:

```
{"role": "user", "content": "What's the weather like today?"},
{"role": "assistant", "content": "It's sunny with a high of 75 degrees."},
{"role": "user", "content": "Great! What about tomorrow?"},
{"role": "assistant", "content": "Tomorrow will be partly cloudy with a high of 72 degrees."}
```

Further Reading on Instruction Tuning

- Multitask Prompted Training Enables Zero-Shot Task Generalization [Sanh et al., 2021]
- Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks [Wang et al., 2022]
- Self-Instruct: Aligning Language Models with Self-Generated Instructions [Wang et al., 2022]
- LIMA: Less Is More for Alignment [Zhou et al., 2023]

Thank You!

Yu Meng
University of Virginia
yumeng5@virginia.edu