



LLM Reasoning

Yu Meng

University of Virginia
yumeng5@virginia.edu

Oct 23, 2024

Announcement

Project midterm report grades & feedback posted

Join at

slido.com

#4015 388



Overview of Course Contents

Join at

slido.com

#4015 388



- 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 and Neural Language Models
- Week 6-7: Language Modeling with Transformers (Pretraining + Fine-tuning)
- Week 8: Large Language Models (LLMs) & In-context Learning
- **Week 9-10: Reasoning, Knowledge, and Retrieval-Augmented Generation (RAG)**
- Week 11: LLM Alignment
- Week 12: Language Agents
- Week 13: Recap + Future of NLP
- Week 15 (after Thanksgiving): Project Presentations



(Recap) Scaling Up Pretraining Data

The Pile: 22 sub-datasets (> 800GB), a common choice for pretraining corpus

Composition of the Pile by Category

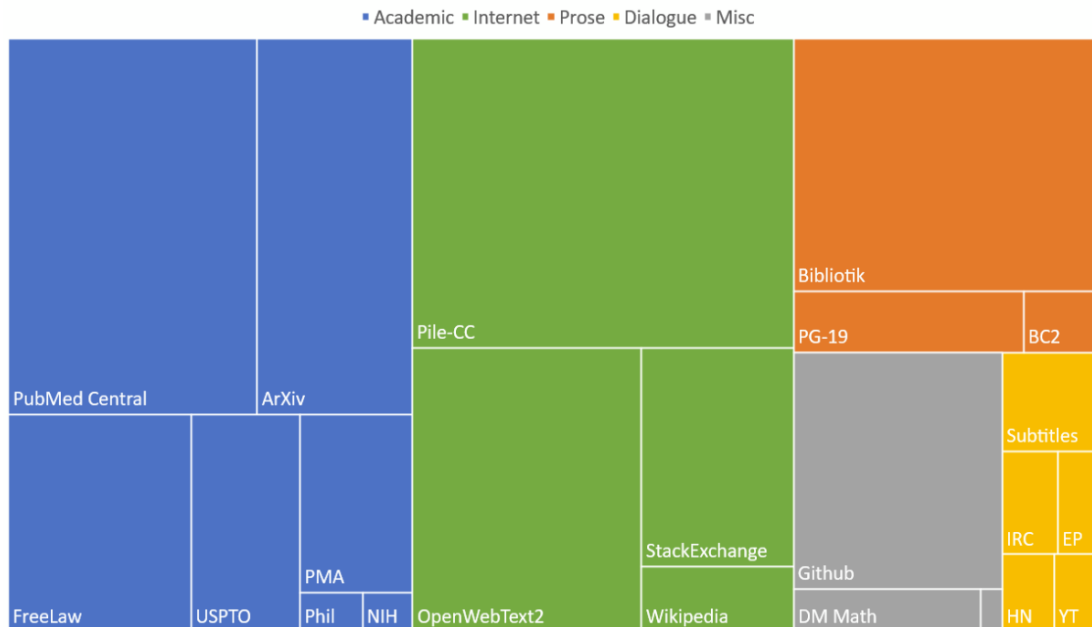
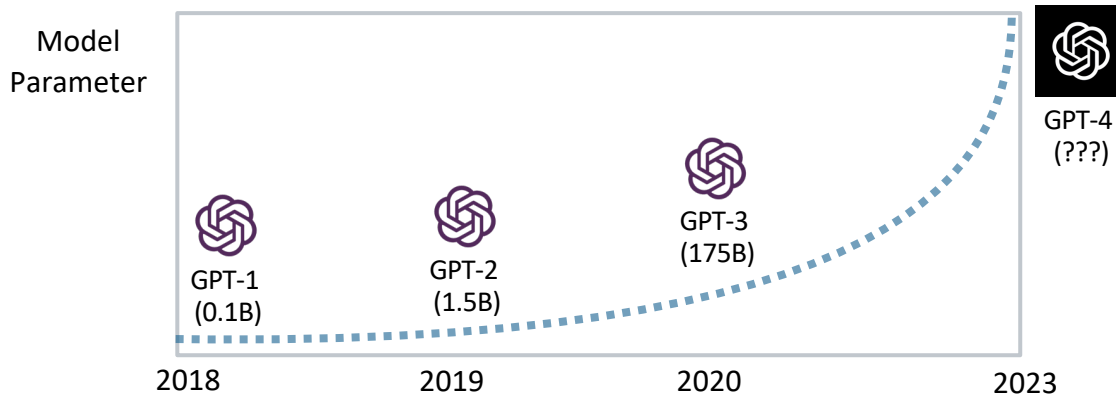


Figure source: <https://arxiv.org/pdf/2101.00027>



(Recap) Scaling Up Model Sizes

- GPT-1 (2018): 12 layers, 117M parameters, trained in ~1 week
- GPT-2 (2019): 48 layers, 1.5B parameters, trained in ~1 month
- GPT-3 (2020): 96 layers, 175B parameters, trained in several months



Papers: (GPT-1) https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

(GPT-2) https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

(GPT-3) <https://arxiv.org/pdf/2005.14165.pdf>

(Recap) Emergent Abilities

Join at

slido.com

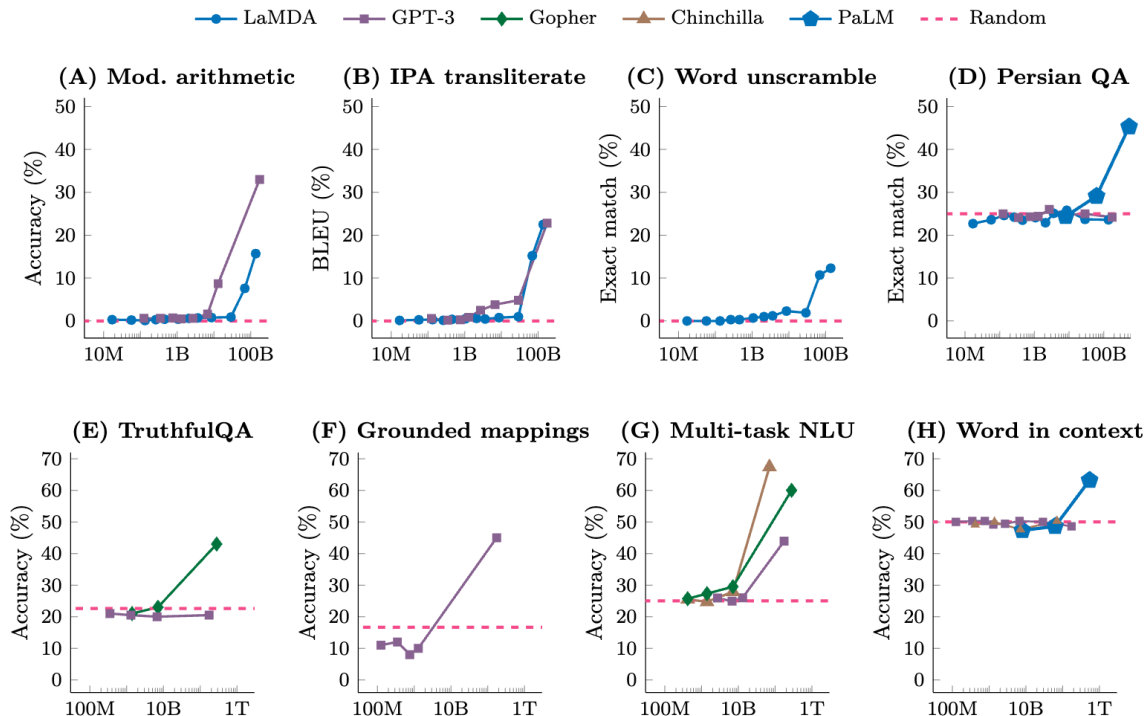
#4015 388



- Consider the **few-shot in-context learning** paradigm
- Consider an ability to be **emergent** when a model has **random** performance until a certain scale, after which performance increases to **well-above random**
- Abilities to test
 - Arithmetic: addition, subtraction, multiplication
 - Transliteration
 - Recover a word from its scrambled letters
 - Persian question answering
 - Question answering (truthfully)
 - Grounded conceptual mappings
 - Multi-task understanding (math, history, law, ...)
 - Contextualized semantic understanding

(Recap) Performance vs. Model Scale

Join at
 slido.com
 #4015 388



Model scale (number of parameters)

Models exhibit random performance until a certain scale, after which performance significantly increases



(Recap) Scaling Laws of LLMs

- (Pretrained) LLM performance is mainly determined by 3 factors
 - Model size: the number of parameters
 - Dataset size: the amount of training data
 - Compute: the amount of floating point operations (FLOPs) used for training
- Scaling up LLMs involves scaling up the 3 factors
 - Add more parameters (adding more layers or having more model dimensions or both)
 - Add more data
 - Train for more iterations
- **Scaling laws:** study the correlation between the cross-entropy language modeling loss and the above three factors
- How to optimally allocate a fixed compute budget?

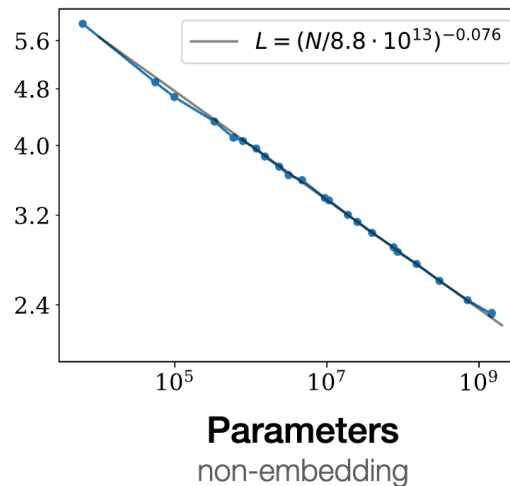


(Recap) Scaling Model Parameters

- Language model loss vs. models with a limited number of parameters (N)
 - Only count non-embedding parameters
 - Infinite compute: trained to convergence
 - Infinite dataset: trained with sufficiently large datasets
- Performance depends strongly on scale, weakly on model shape (depth vs. width)

$$\mathcal{L}(N) = \left(\frac{N_c}{N} \right)^{\alpha_N}, \quad \alpha_N \approx 0.076, \quad N_c \approx 8.8 \times 10^{13}$$


 Model parameters
 (non-embedding)



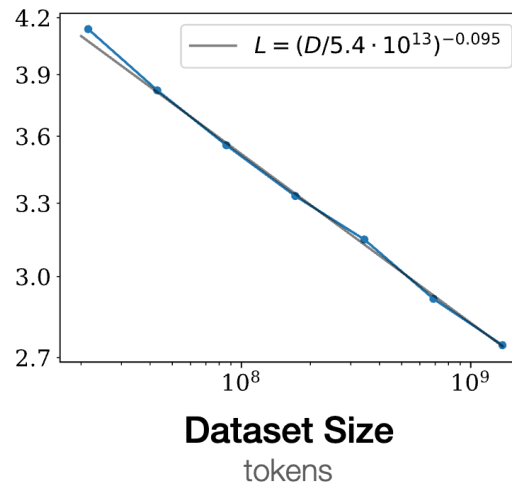


(Recap) Scaling Dataset Size

- Language model loss vs. a limited dataset size (D)
 - Infinite model size: sufficiently large model
 - With appropriate early stopping: avoid overfitting to the training data

$$\mathcal{L}(D) = \left(\frac{D_c}{D} \right)^{\alpha_D}, \quad \alpha_D \approx 0.095, \quad D_c \approx 5.4 \times 10^{13}$$


 Dataset size
 (# of tokens)



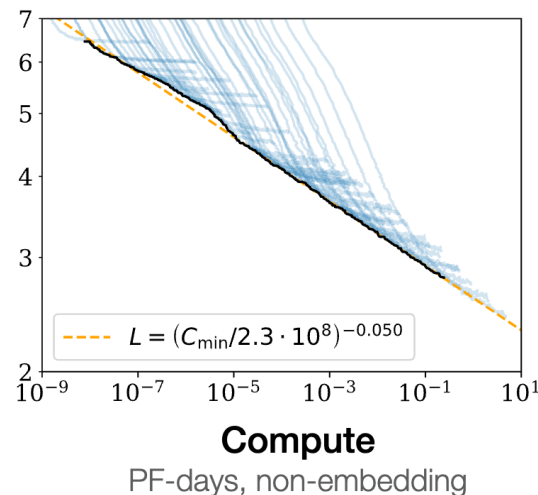


(Recap) Scaling Training Compute

- Language model loss vs. a limited amount of compute (C)
 - Infinite dataset size: sufficiently large training corpus
 - Optimal model size: can effectively learn the data and not excessively compute-consuming

$$\mathcal{L}(C) = \left(\frac{C_c}{C}\right)^{\alpha_C}, \quad \alpha_C \approx 0.050, \quad C_c \approx 3.1 \times 10^8$$

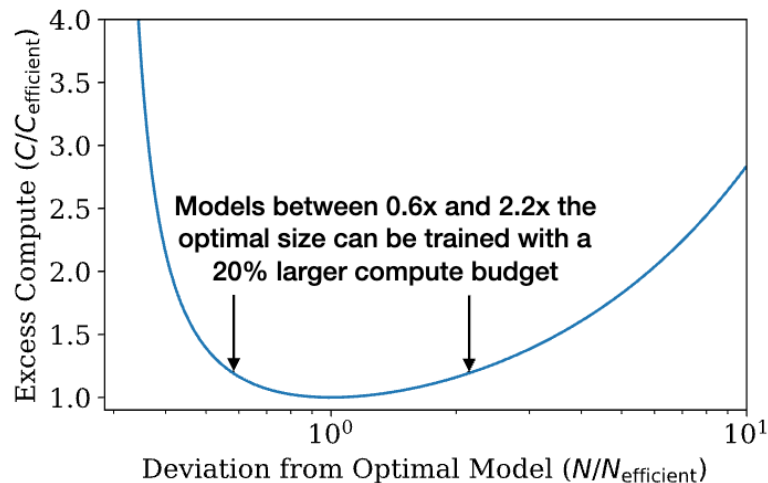
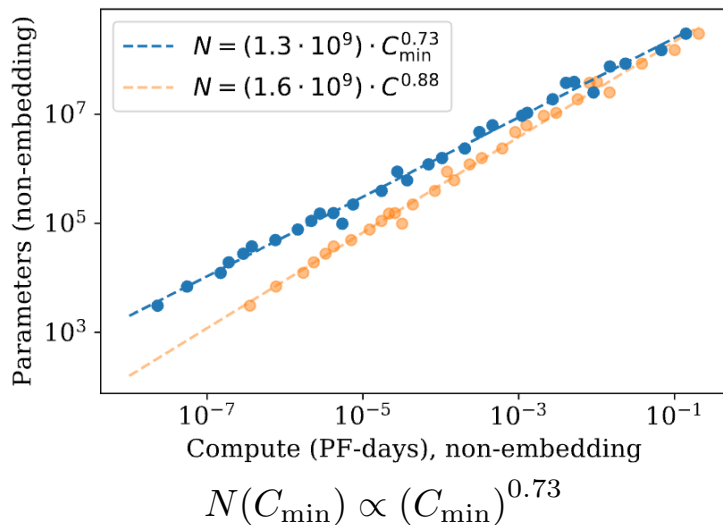

 Compute
 (# Peta-FLOP days)





(Recap) Optimal Model Size

- Given a specific amount of training compute C , what's the optimal model size $N(C)$ that leads to minimal language modeling loss?
- $N(C)$ can be fit with a power-law wrt C
- Additional compute needs to be used when model size is suboptimal



Agenda

- Chain-of-thought Reasoning
- Reasoning Benchmarks
- Question Answering & Hallucination

Join at

slido.com

#4015 388





Reasoning: Overview

- **Reasoning** (rough definition): perform deductive, inductive, commonsense, or logical reasoning via generating or analyzing text
- 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”

Latest LLMs for Reasoning: OpenAI o1

Join at
slido.com
#4015 388



September 12, 2024

Learning to Reason with LLMs

We are introducing OpenAI o1, a new large language model trained with reinforcement learning to perform complex reasoning. o1 thinks before it answers—it can produce a long internal chain of thought before responding to the user.

Contributions

OpenAI o1 ranks in the 89th percentile on competitive programming questions (Codeforces), places among the top 500 students in the US in a qualifier for the USA Math Olympiad (AIME), and exceeds human PhD-level accuracy on a benchmark of physics, biology, and chemistry problems (GPQA). While the work needed to make this new model as easy to use as current models is still ongoing, we are releasing an early version of this model, OpenAI o1-preview, for immediate use in ChatGPT and to trusted API users.

OpenAI o1: Commonsense Reasoning

Join at
slido.com
#4015 388



o1-preview

Assume laws of physics on Earth. A small strawberry is put into a normal cup and the cup is placed upside down on a table. Someone then takes the cup and puts it inside the microwave. Where is the strawberry now? Explain your reasoning step by step



Tracking the strawberry's path

To determine the current location of the strawberry, let's break down the events step by step:

1. **Initial Placement:** A small strawberry is placed into a normal cup. The strawberry is inside the cup, resting at the bottom.
2. **Cup Inversion:** The cup is then placed upside down on a table. This means the open end of the cup (where the strawberry was initially placed) is now facing down, in contact with the table surface. The strawberry is now sitting on the table, covered by the inverted cup.
3. **Cup Removal:** Someone picks up the cup to put it inside the microwave. When lifting the inverted cup off the table:



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



Standard Prompting vs. CoT Prompting

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?

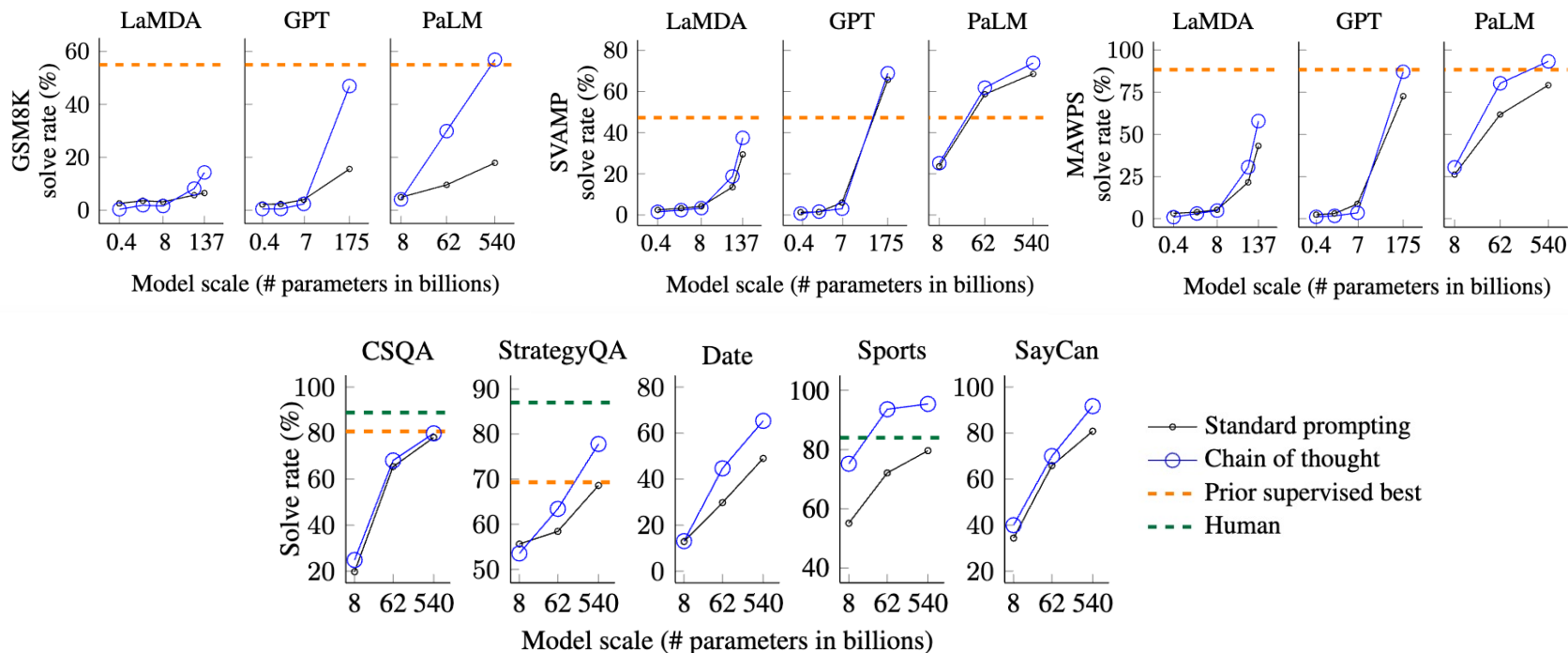
Model Output

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. ✅



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.* ✗

(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* ✗

(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.* ✓

(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.* ✓

CoT Demo

Join at
slido.com
#4015 388



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

llama-3.1-70b-instruct

Expand to see the descriptions of 75 models

Scroll down and start chatting

How many 'r' letters are there in the following word: strawberry? Answer without reasoning steps

There are 2 'r' letters in the word "strawberry". **Wrong result**

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

llama-3.1-70b-instruct

Expand to see the descriptions of 75 models

Scroll down and start chatting

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

Step 1: Write down the word
The word is: strawberry

Step 2: Break down the word into individual letters
Here are the individual letters: s-t-r-a-w-b-e-r-r-y

Step 3: Count the number of 'r' letters
Let's count the 'r' letters:

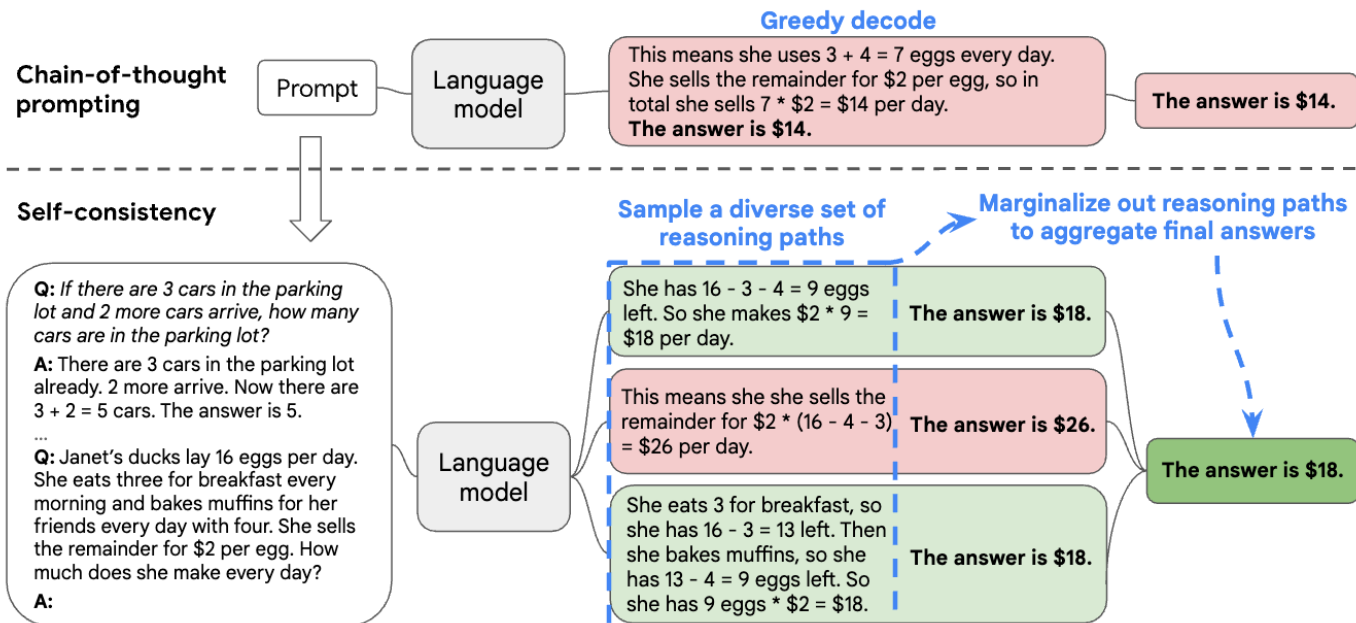
1. r (first 'r')
2. r (second 'r')
3. r (third 'r')

There are 3 'r' letters in the word "strawberry". **Correct result**



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





Self-consistency CoT Demo

- Prompt: When rolling two dice, what is the probability that you roll a total number that is at least 3?

(Previous generation cropped)

- Total of 11: (5,6), (6,5)
- Total of 12: (6,6)

There are 33 favorable outcomes.

Step 3: Calculate the Probability

The probability of rolling a total number that is at least 3 is the number of favorable outcomes divided by the total number of possible outcomes.

Probability = Number of favorable outcomes / Total number of possible outcomes
= 33 / 36
= 11/12

Answer

The probability of rolling a total number that is at least 3 when rolling two dice is 11/12.

Wrong result

(Previous generation cropped)

- Total of 10: (4,6), (5,5), (6,4) = 3 outcomes
- Total of 11: (5,6), (6,5) = 2 outcomes
- Total of 12: (6,6) = 1 outcome

There are 36 possible outcomes, and 35 of them have a total of at least 3.

Step 3: Calculate the Probability

The probability of rolling a total number that is at least 3 is the number of favorable outcomes divided by the total number of possible outcomes.

Probability = Number of favorable outcomes / Total number of possible outcomes
= 35 / 36

Answer

The probability of rolling a total number that is at least 3 with two dice is 35/36, which is approximately 0.9722 (or 97.22%).

Correct result

Generated twice with temperature = 0.1, top-p = 0.7
Figure source: <https://lmarena.ai/?model=llama-3.1-70b-instruct>



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

- Chain-of-thought Reasoning
- Reasoning Benchmarks
- Question Answering & Hallucination

Join at

slido.com

#4015 388



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 $4 \times 2 = 8$ dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = 96$ cookies
She splits the 96 cookies equally amongst 16 people so they each eat $96/16 = 6$ 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 = 50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = 200 gallons.
She was able to sell 200 gallons - 24 gallons = 176 gallons.

Thus, her total revenue for the milk is $\$3.50/\text{gallon} \times 176 \text{ gallons} = \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 $3 \times 12 = 36$ sodas
6 people attend the party, so half of them is $6/2 = 3$ people
Each of those people drinks 3 sodas, so they drink $3 \times 3 = 9$ sodas
Two people drink 4 sodas, which means they drink $2 \times 4 = 8$ sodas
With one person drinking 5, that brings the total drank to $5 + 9 + 8 + 3 = 25$ sodas
As Tina started off with 36 sodas, that means there are $36 - 25 = 11$ sodas left

Final Answer: 11



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 $\triangle ABC$ 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.

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.



~10K commonsense natural language inference problems: select the best follow-up



ACTIVITYNET

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

+


 Adversarial
Filtering

- A. rinses the bucket off with soap and blow dry the dog's head.
- B. uses a hose to keep it from getting soapy.
- C. gets the dog wet, then it runs away again.**
- D. gets into a bath tub with the dog.


 wikiHow
to do anything

 How to
determine
who has right
of way.

+


 Adversarial
Filtering

Come to a complete halt at a stop sign or red light. At a stop sign, come to a complete halt for about 2 seconds or until vehicles that arrived before you clear the intersection. If you're stopped at a red light, proceed when the light has turned green. ...

- A. Stop for no more than two seconds, or until the light turns yellow. A red light in front of you indicates that you should stop.
- B. After you come to a complete stop, turn off your turn signal. Allow vehicles to move in different directions before moving onto the sidewalk.
- C. Stay out of the oncoming traffic. People coming in from behind may elect to stay left or right.

D. If the intersection has a white stripe in your lane, stop before this line. Wait until all traffic has cleared before crossing the intersection.

BIG-Bench Hard (BBH)

Join at
slido.com
#4015 388



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

Model Input (Navigate)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: If you follow these instructions, do you return to the starting point? Turn left. Turn right. Take 5 steps. Take 4 steps. Turn around. Take 9 steps.

Options:

- Yes
- No

A: Let's think step by step.

Model Output

We start at the origin (0, 0), facing the positive y-axis.

- (1) Turn left: (0, 0), facing the negative x-axis.
- (2) Turn right: (0, 0), facing the positive y-axis.
- (3) Take 5 steps: (0, 5), facing the positive y-axis.
- (4) Take 4 steps: (0, 9), facing the positive y-axis.
- (5) Turn around: (0, 9), facing the negative y-axis.
- (6) Take 9 steps: (0, 0), facing the negative y-axis.

Since (0, 0) is (0, 0), we are indeed where we started. So the answer is **Yes**. ✓

Model Input (Word Sorting)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: Sort the following words alphabetically: List: burley bela arapaho bacteria bock

A: Let's think step by step.

Model Output

The first letter: "burley": "b" (2). "bela": "b" (2). "arapaho": "a" (1). "bacteria": "b" (2). "bock": "b" (2).

We now have: (1) "arapaho" < (2) ["burley" ? "bela" ? "bacteria" ? "bock"]. Now let's sort this subpart ["burley" ? "bela" ? "bacteria" ? "bock"] by looking at their second letters.

The second letter: "burley": "u" (21). "bela": "e" (5). "bacteria": "a" (1). "bock": "o" (15). We now have: (1) "bacteria" < (5) "bela" < (15)

"bock" < (21) "burley". Hence, we have "arapaho" < ["bacteria" < "bela" < "bock" < "burley"]. So the answer is **arapaho bacteria bela bock burley**. ✓

Agenda

- Chain-of-thought Reasoning
- Reasoning Benchmarks
- Question Answering & Hallucination

Join at

slido.com

#4015 388





Introduction to Question Answering

- **Question Answering (QA):** build systems that can automatically answer questions posed by humans in natural language
- Categorization by application domain: closed-domain vs. open-domain QA
- **Closed-domain QA:** answer questions within a specific domain
 - Example: medical, legal, technical fields
 - Models are trained on specialized knowledge to be highly accurate within their domain
- **Open-domain QA:** answer questions from any domain
 - Typically rely on vast (external) knowledge sources like the web or large text corpora
 - Most LLM applications consider open-domain QA settings



Introduction to Question Answering

- **Question Answering (QA):** build systems that can automatically answer questions posed by humans in natural language
- Categorization by modeling approach: extractive vs. abstractive QA
- **Extractive QA:** output a span of text extracted directly from a given context
 - A natural language understanding task (reading comprehension)
 - Example: context: “The human brain contains approximately 86 billion neurons” Q: “How many neurons are in the human brain?” A: “86 billion”
 - Can be done with encoder-only LMs (e.g., BERT)
- **Abstractive QA:** synthesize the answer in its own words (rephrasing/summarizing)
 - Example: context: “Albert Einstein published his theory of special relativity which introduced the famous equation $E=mc^2$, which relates energy (E) to mass (m) and the speed of light (c)” Q: “What did Einstein contribute to physics?” A: “Einstein made significant contributions to the theory of special relativity which established the relationship between energy and mass”
 - Need to use a generative LM (e.g., GPT)



Introduction to Question Answering

- **Question Answering (QA):** build systems that can automatically answer questions posed by humans in natural language
- Categorization by access to external source: closed-book vs. open-book QA
- **Closed-book QA:** answer questions without access to any external information
 - Accuracy depends heavily on how well the training data covered the relevant information
 - Similar to a human answering a question from memory without looking anything up
- **Open-book QA:** can access external knowledge source to answer the questions
 - Typically using retrieval from reliable external sources that contain
 - Similar to a human answering a question by looking it up in a book or online resource

Hallucination

Join at

slido.com

#4015 388



- **Hallucination:** LM generates information that is factually incorrect, misleading, or fabricated, even though it may sound plausible or convincing
- Why does hallucination happen?
 - Limited knowledge: LLMs are trained on finite datasets, which don't have access to all possible information; when asked about topics outside their training data, they may generate plausible-sounding but incorrect responses
 - Overgeneralization: LLMs may apply patterns they've learned from one context to another where they don't apply, leading to incorrect conclusions
 - Lack of common sense: While LLMs can process and generate human-like text, they often lack the ability to apply commonsense reasoning to their outputs
 - ...

Hallucination Examples

Join at

slido.com

#4015 388



- **(Limited knowledge)** Q: “What were the main features of the iPhone 15 Pro Max?”
LLM (trained before 2023): “The iPhone 15 Pro Max features a revolutionary holographic display, quantum computing chip, and telepathic user interface.”
- **(Overgeneralization)** Q: “How do you form the past tense in Japanese?”
LLM: “In Japanese, you typically add '-ed' to the end of verbs to form the past tense, just like in English.” (incorrect)
- **(Lack of common sense)** Q: “How many tennis balls can fit in a typical smartphone?”
LLM: “Approximately 15-20 tennis balls can fit in a typical smartphone, depending on the model and screen size.”

Concerns About Hallucination

Still a concerning issue in modern LLMs!

Attorneys Face Sanctions After Citing Case Law 'Hallucinated' by ChatGPT

BY PYMNTS | MAY 30, 2023



Join at
slido.com
#4015 388



Figure source: <https://www.pymnts.com/artificial-intelligence-2/2023/attorneys-face-sanctions-after-citing-information-hallucinated-by-chatgpt/>



Thank You!

Yu Meng

University of Virginia

yumeng5@virginia.edu