

Chain-of-Thought Reasoning in Large Language Model Prompting

Zheyu Zhang



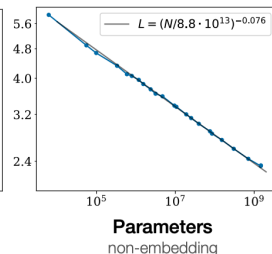
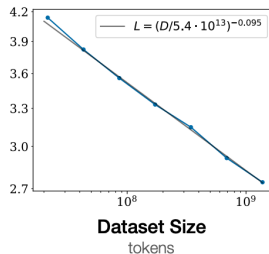
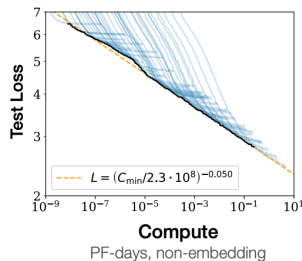
Center for Information & Language Processing (CIS), LMU Munich

June 30, 2023

- 1 Background
- 2 Chain-of-Thought Prompting
 - Few-Shot CoT
 - Zero-Shot CoT
- 3 Language Models Are Greedy Reasoners
 - Dataset: PRONTOQA
 - Evaluation Methods
 - Experimental Results
- 4 Conclusion

Recent Trends

- Larger and larger models
- Trained on more and more data

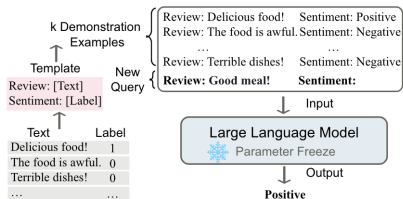


Kaplan et al. (2020)

Emergent Abilities

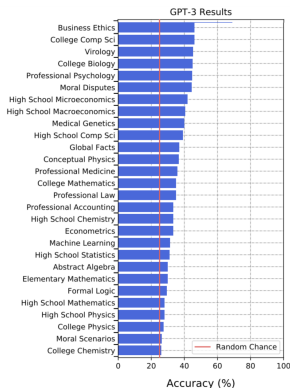
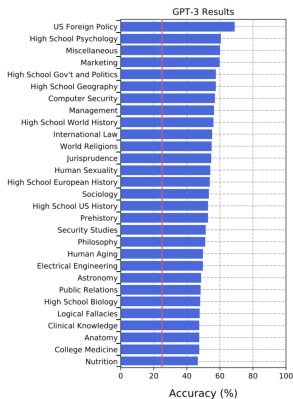
An ability is emergent if it is not present in smaller models but is present in larger models [Wei et al. (2022a)].

- *Few-Shot Prompting*: Specify a task by simply prepending examples of the task before your example
- Also called **in-context learning**, to stress that no gradient updates are performed when learning a new task



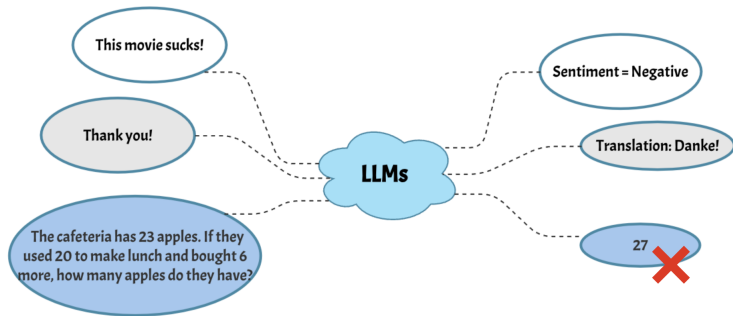
Dong et al. (2022)

- Does not require a large training dataset
- Allows a single model checkpoint to handle multiple tasks without sacrificing generality



Hendrycks et al. (2020)

- Some tasks seem too hard for even LLMs to learn through prompting alone
- Especially tasks involving **richer, multi-step reasoning**

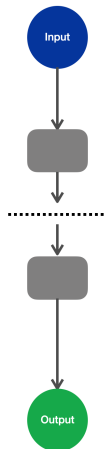


- 1 Background
- 2 Chain-of-Thought Prompting
 - Few-Shot CoT
 - Zero-Shot CoT
- 3 Language Models Are Greedy Reasoners
 - Dataset: PRONTOQA
 - Evaluation Methods
 - Experimental Results
- 4 Conclusion

Chain-of-Thought (CoT) prompting enables LLMs to generate *intermediate reasoning steps* before inferring an answer

- With a few demonstrations or just a prompt sentence
- Without gradient updates
- Additional computation can be allocated to problems that require more reasoning steps

CHAIN OF THOUGHT PROMPTING (CoT)



Paradigm Shift of Task Format:

- Standard Format: <input → output>
- CoT Format: <input → rationale → output>

Standard Prompting

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

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

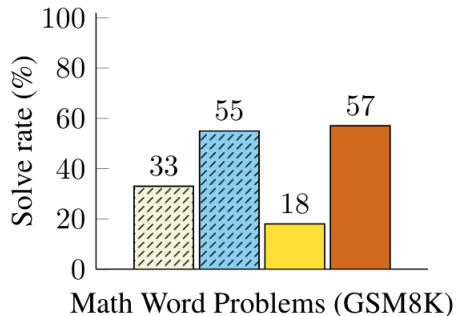
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Question

Rationale

Answer

- Finetuned GPT-3 175B
- Prior best
- PaLM 540B: standard prompting
- PaLM 540B: chain-of-thought prompting

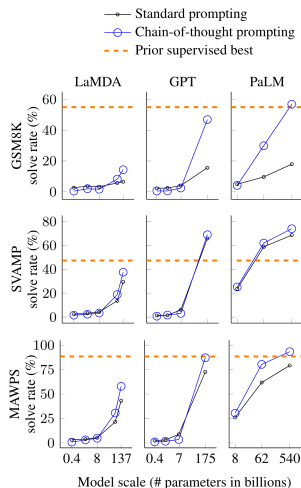


Wei et al. (2022b)

How does CoT Prompting work?

Chain-of-Thought prompting is an emergent property of **model scale**

- CoT prompting does not improve performance for small models
- Only enhances performance with models of $\sim 100\text{B}$ parameters
- Smaller-scale models produce fluent but illogical CoT
⇒ lower performance



Wei et al. (2022b)

Chain-of-Thought Prompting is applicable to *a wide range of problems*, e.g. arithmetic, commonsense, symbolic reasoning, etc.:

Math Word Problems (free response)

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.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm³, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

Wei et al. (2022b)

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)


Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

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?

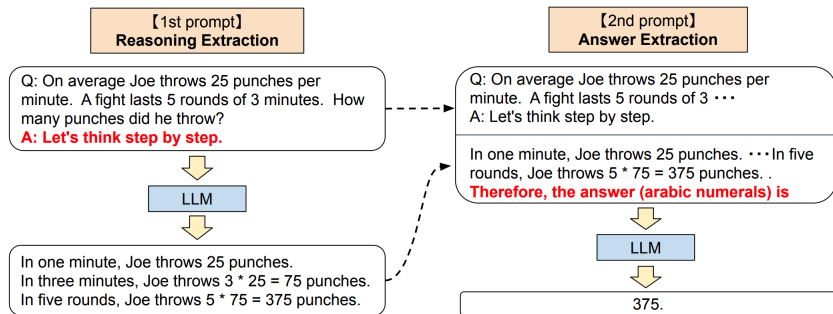
Model Output

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. 

Wei et al. (2022b)

- Do we even need examples of reasoning?
- Can we just ask the model to reason through things?

With a **trigger hint**, e.g., “let’s think step by step” after the question (question + hint):



Kojima et al. (2022)

Tasks: multi-step reasoning tasks, e.g., math word problems, commonsense reasoning, symbolic reasoning, and logical reasoning

Model	MultiArith		GSM8K		AddSub		AQUA-RAT		SingleEq		SVAMP	
	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT
<i>Zero-Shot Performance</i>												
text-davinci-002	22.7	78.7	12.5	40.7	77.0	74.7	22.4	33.5	78.7	78.7	58.8	63.7
text-davinci-003	24.2	83.7	12.6	59.5	87.3	81.3	28.0	40.6	82.3	86.4	64.7	73.6
ChatGPT	30.3	96.0	14.7	75.4	89.6	89.9	23.6	47.6	83.1	91.3	68.1	82.8
<i>Few-Shot Performance</i>												
UL2	5.0	10.7	4.1	4.4	18.5	18.2	20.5	23.6	18.0	20.2	10.1	12.5
LaMDA	7.6	44.9	6.5	14.3	43.0	51.9	25.5	20.6	48.8	58.7	29.5	37.5
text-davinci-002	33.8	91.7	15.6	46.9	83.3	81.3	24.8	35.8	82.7	86.6	65.7	68.9
Codex	44.0	96.2	19.7	63.1	90.9	90.9	29.5	45.3	86.8	93.1	69.9	76.4
PaLM	42.2	94.7	17.9	56.9	93.9	91.9	25.2	35.8	86.5	92.3	69.4	79.0

Model	CSQA		Commonsense StrategyQA		COPA		Symbolic Last Letter		Symbolic Coin Flip		Logical Date		Logical Object	
	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT	N/A	CoT
<i>Zero-Shot Performance</i>														
text-davinci-002	72.6	64.6	54.3	54.8	-	-	0.2	57.6	53.8	91.4	49.3	67.5	31.3	52.9
text-davinci-003	74.9	70.0	57.2	61.1	93.0	63.0	0.0	54.4	49.0	97.8	56.6	77.0	27.1	39.7
ChatGPT	74.7	74.6	59.2	62.5	58.0	51.0	1.6	2.6	25.4	69.6	51.0	66.1	27.1	36.1
<i>Few-Shot Performance</i>														
UL2	34.2	51.4	59.0	53.3	-	-	0.6	18.8	70.4	67.1	13.5	14.0	-	-
LaMDA	53.6	57.9	62.4	65.4	-	-	5.8	77.5	49.0	99.6	21.5	26.8	-	-
text-davinci-002	79.5	73.5	65.9	65.4	-	-	0.2	59.0	57.2	97.2	43.8	52.1	-	-
Codex	82.3	77.9	67.1	73.2	-	-	-	-	-	-	49.0	64.8	-	-
PaLM	78.1	79.9	68.6	77.8	95.0	-	7.6	99.4	98.1	100.0	49.0	65.3	23.9	-

Accuracy (%) of different models without CoT (N/A) and with CoT - [Qin et al. (2023)]

- How these models obtain the answers?
- What extent these models can reason?
- Whether they rely on simple heuristics rather than the generated chain-of-thought?

- 1 Background
- 2 Chain-of-Thought Prompting
 - Few-Shot CoT
 - Zero-Shot CoT
- 3 Language Models Are Greedy Reasoners
 - Dataset: PRONTOQA
 - Evaluation Methods
 - Experimental Results
- 4 Conclusion

Published as a conference paper at ICLR 2023

LANGUAGE MODELS ARE GREEDY REASONERS: A SYSTEMATIC FORMAL ANALYSIS OF CHAIN-OF- THOUGHT

Abulhair Saparov & He He

Center for Data Science, New York University, New York, NY 10011, USA

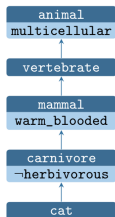
{as17582, hhe}@nyu.edu

Investigating the **reasoning capability** of LLMs:

- Focus on InstructGPT and GPT-3
- Construct a new synthetic QA dataset called **PRONTOQA** for analysis
- Directly evaluating their predicted *chains-of-thought* (the interpretable proof steps), rather than the *predicted label*

Proof and Ontology-Generated Question-Answering (PRONTOQA):

Step 1:
Generate ontology



Step 2: Generate proof from ontology

$$\frac{\frac{\text{cat}(fae) \text{ Ax} \quad \frac{\forall x(\text{cat}(x) \rightarrow \text{carnivore}(x)) \text{ Ax}}{\text{carnivore}(fae)} \text{ Hop}}{\text{carnivore}(fae)} \text{ Hop} \quad \frac{\forall x(\text{carnivore}(x) \rightarrow \neg \text{herbivorous}(x)) \text{ Ax}}{\neg \text{herbivorous}(fae)} \text{ Hop}}{\neg \text{herbivorous}(fae)} \text{ Hop}$$

Step 3: Translate ontology to natural language context

"Q: Each cat is a carnivore. Every carnivore is not herbivorous. Carnivores are mammals. All mammals are warm-blooded. Mammals are vertebrates. Every vertebrate is an animal. Animals are multicellular."

Step 4: Translate proof into query, chain-of-thought, and label

"Fae is a cat. True or false: Fae is not herbivorous."

A: Fae is a cat. Cats are carnivores. Fae is a carnivore. Every carnivore is not herbivorous. Fae is not herbivorous. True"

Q: Each cat is a carnivore. Every carnivore is not herbivorous. Carnivores are mammals. All mammals are warm-blooded. Mammals are vertebrates. Every vertebrate is an animal.

Animals are multicellular. Fae is a cat. True or false: Fae is not herbivorous. — context

A: Fae is a cat. Cats are carnivores. Fae is a carnivore. Every carnivore is not herbivorous. — query

Fae is not herbivorous. True — chain-of-thought

True — label

The evaluation focuses on checking if the model derives the right answer for the right reason:

- Each sentence of the predicted CoT is parsed into logical form
- The provability of the logical form is computed based on previous steps using deduction rules

Proof steps are categorized based on three dimensions:

- **Validity** (*strictly-valid, broadly-valid, or invalid*):
Is the current proof step provable from previous steps?
- **Atomicity** (*atomic or non-atomic*):
Is the current proof step provable from previous steps with exactly one application of a deduction rule?
- **Utility** (*correct or misleading*):
Is the current proof step's premises part of the gold proof?

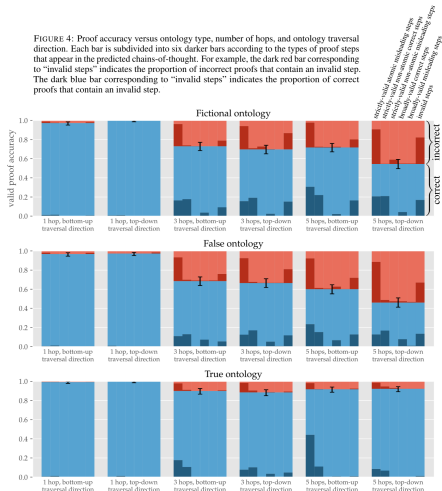
Step type	Example (the conclusion of each step is highlighted green)
Strictly-valid atomic correct step, or canonical step	" <i>Fae is a cat. Cats are carnivores. Fae is a carnivore. Every carnivore is not herbivorous. Fae is not herbivorous. True</i> " (this is the gold CoT for this example)
Strictly-valid atomic misleading step	" <i>Fae is a cat. Cats are carnivores. Fae is a carnivore. Every carnivore is a mammal. Fae is a mammal...</i> "
Strictly-valid non-atomic correct step	" <i>Fae is a cat. Fae is a carnivore. Every carnivore is not herbivorous. Fae is not herbivorous. True</i> "
Strictly-valid non-atomic misleading step	" <i>Fae is a cat. Cats are carnivores. Fae is a carnivore. Fae is a mammal. Every mammal is a vertebrate...</i> "
Broadly-valid correct step	" <i>Fae is a cat. Every cat is not herbivorous. Fae is not herbivorous...</i> "
Broadly-valid misleading step	" <i>Fae is a cat. Every cat is a mammal. Fae is a mammal...</i> "
Invalid step	" <i>Fae is a cat. Cats are carnivores. Fae is a carnivore. Every carnivore is a cat. Fae is a cat...</i> "

Saparov and He (2022)

Each experiment involves generating QA examples, performing CoT prompting on LLMs, and analyzing predicted CoTs

- Run experiments on InstructGPT and original GPT-3 with greedy decoding
 - 8-shot in-context learning
- Examples generated independently from PRONTOQA, controlling variables
 - Number of hops (1, 3, and 5)
 - Ordering of sentences from the ontology (top-down or bottom-up traversal)
 - ⇒ controls the difficulty of the generated example
 - Type of ontology (fictional, true, and false)
 - *Fictional*: use fictional concept names
 - ⇒ avoid pretraining knowledge confounding effects
 - *True*: use real concept names and are consistent with the real-world
 - *False*: use real concept names but generate false statements

FIGURE 4: Proof accuracy versus ontology type, number of hops, and ontology traversal direction. Each bar is subdivided into six darker bars according to the types of proof steps that appear in the predicted chains-of-thought. For example, the dark red bar corresponding to "invalid steps" indicates the proportion of incorrect proofs that contain an invalid step. The dark blue bar corresponding to "invalid steps" indicates the proportion of correct proofs that contain an invalid step.



Real-world knowledge significantly impacts reasoning

- Performance similar in fictional and "false" ontology settings but much better with a "true" ontology
- Performance remains stable even with an increased number of hops in the "true" ontology

Longer proofs are still challenging

- Performance more limited as the number of hops increases

Traversal direction affects reasoning

- The number of hops increases, the model becomes sensitive to the traversal direction of the ontology

Observations:

- Most predicted proof steps are strictly-valid
- LLMs tend to skip steps by producing non-atomic steps, just as humans do when they verbalize their reasoning
- Main Source of Reasoning Error: misleading steps
⇒ LLMs are able to produce valid proof steps with high probability but have difficulty with proof planning

- 1 Background
- 2 Chain-of-Thought Prompting
 - Few-Shot CoT
 - Zero-Shot CoT
- 3 Language Models Are Greedy Reasoners
 - Dataset: PRONTOQA
 - Evaluation Methods
 - Experimental Results
- 4 Conclusion

- Large language models (LLMs) have shown remarkable reasoning capabilities given chain-of-thought prompts
- LLMs exhibit competence in generating accurate deduction steps, showcasing their overall reasoning proficiency, even within fictional contexts
- LLMs have challenges in the domain of proof planning, particularly when confronted with multiple valid deduction steps

Thanks for your attention!

- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.
- Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.