

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

Author: Wei et al.

created by paper2slides

January 28, 2022

<https://arxiv.org/abs/2201.11903>

Executive Summary

- **Objective:** Explore how generating a *chain of thought* (a series of intermediate reasoning steps) improves complex reasoning in LLMs.
- **Method:** Chain-of-thought prompting provides few-shot exemplars with intermediate reasoning steps.
- **Results:** Significant performance improvements on arithmetic, commonsense, and symbolic reasoning tasks.
- **Key Findings:** Emergence of reasoning abilities at model scales 100B parameters and above.

Introduction

- Language models have transformed NLP, with improved performance and sample efficiency as size increases.
- However, scaling size alone is insufficient for tasks like arithmetic, commonsense, and symbolic reasoning.
- This work combines natural language generation and prompting to improve reasoning in LLMs.
- Chain-of-thought prompting provides models with intermediate reasoning steps.

Standard Prompting

Model Input

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

A: The answer is 11.

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

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

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

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

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

Model Output

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

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

Key Innovations in Chain-of-Thought Prompting

- **Natural Language Rationales:** Models generate intermediate natural language rationales leading to final answers.
- **Few-Shot Prompting:** Exemplars in prompting provide input, chain of thought, and output triples.
- **Decomposition:** Allows models to decompose problems into intermediate steps, allocating computation as needed.
- **Interpretability:** Chains of thought provide insight into model behavior, aiding debugging and understanding.

Experimental Setup: Arithmetic Reasoning

- **Benchmarks:** GSM8K, SVAMP, ASDiv, AQuA, MAWPS
- **Prompting Methods:**
 - Standard prompting (input-output pairs)
 - Chain-of-thought prompting (input, chain of thought, output triples)
- **Language Models Evaluated:** GPT-3, LaMDA, PaLM, UL2, Codex

Results: Arithmetic Reasoning

- Chain-of-thought prompting yields substantial performance improvements.
- Examples:
 - GSM8K: PaLM 540B achieved state-of-the-art performance with a significant margin.
 - MAWPS: Performance gains for difficult subsets.
- Emergent abilities observed: Major improvements for models above 100B parameters.

<p>Math Word Problems (free response)</p> <p>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?</p> <p>A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.</p>	<p>Math Word Problems (multiple choice)</p> <p>Q: How many keystrokes are needed to type the numbers from 1 to 500?</p> <p>Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788</p> <p>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).</p>	<p>CSQA (commonsense)</p> <p>Q: Sammy wanted to go to where the people were. Where might he go?</p> <p>Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock</p> <p>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).</p>
<p>StrategyQA</p> <p>Q: Yes or no: Would a pear sink in water?</p> <p>A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float. So the answer is no.</p>	<p>Date Understanding</p> <p>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?</p> <p>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.</p>	<p>Sports Understanding</p> <p>Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."</p> <p>A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.</p>
<p>SayCan (Instructing a robot)</p> <p>Human: How would you bring me something that isn't a fruit?</p> <p>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.</p> <p>Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().</p>	<p>Last Letter Concatenation</p> <p>Q: Take the last letters of the words in "Lady Gaga" and concatenate them.</p> <p>A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.</p>	<p>Coin Flip (state tracking)</p> <p>Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?</p> <p>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.</p>

Figure: Examples of chain-of-thought prompting for arithmetic, commonsense, and symbolic reasoning benchmarks.

Ablation Study: Arithmetic Reasoning

■ Baseline Comparison:

- Standard prompting
- Equation-only
- Variable compute-only
- Reasoning after the answer

■ Results:

- Chain-of-thought prompting provides the highest gains.
- Equation-only and variable-compute-only approaches performed worse than chain-of-thought prompting.

Experimental Setup: Commonsense Reasoning

- **Benchmarks:** CSQA, StrategyQA, BIGBench Date and Sports, SayCan
- **Prompting Methods:**
 - Standard prompting
 - Chain-of-thought prompting
- **Language Models Evaluated:** GPT-3, LaMDA, PaLM

Results: Commonsense Reasoning

- Chain-of-thought prompting improves performance across all tasks.
- Examples:
 - StrategyQA: PaLM 540B achieved 75.6% (vs. 69.4% prior SOTA)
 - Sports Understanding: PaLM 540B achieved 95.4% (vs. 84% unaided sports enthusiast)

**Types of errors made by
a 62B language model:**

Semantic understanding

(62B made 20 errors of this type,
540B fixes 6 of them)

One step missing

(62B made 18 errors of this type,
540B fixes 12 of them)

Other

(62B made 7 errors of this type,
540B fixes 4 of them)

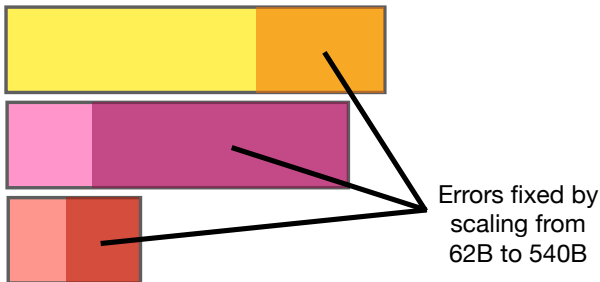


Figure: Error analysis showing the effect of chain-of-thought prompting on commonsense reasoning tasks.

Experimental Setup: Symbolic Reasoning

■ **Tasks:**

- Last letter concatenation
- Coin flip

■ **Prompting Methods:**

- Standard prompting
- Chain-of-thought prompting

■ **In-Domain vs. Out-of-Domain:**

- In-domain: Examples with the same number of steps as exemplars.
- Out-of-domain: Examples with more steps than exemplars.

Results: Symbolic Reasoning

- Chain-of-thought prompting enables models to perform symbolic reasoning tasks effectively.
- In-domain performance:
 - PaLM 540B achieved 100% solve rates for both tasks.
- OOD performance:
 - Standard prompting fails; chain-of-thought prompting shows successful length generalization.

<p>Question</p> <p>Tracy used a piece of wire 4 feet long to support tomato plants in the garden. The wire was cut into pieces 6 inches long. How many pieces did she obtain?</p>	<p>62B Model Output</p> <p>She cut the wire into 6 inch pieces. This means she got $4 \times 6 = 24$ pieces. The answer is 24. (Incorrect: semantic understanding error)</p>	<p>540B Model Output</p> <p>The wire was 4 feet long. This means it was $4 \times 12 = 48$ inches long. It was cut into pieces 6 inches long. This means she obtained $48 / 6 = 8$ pieces. The answer is 8. (Correct)</p>
<p>Question</p> <p>Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4 PM. He then travels back at a rate of 6 mph. How long does it take him to get back?</p>	<p>62B Model Output</p> <p>He travels from 1 to 4 PM. That means he travels for 3 hours. He travels back at 6 mph. That means he travels for $3 \times 6 = 18$ miles. The answer is 18. (Incorrect: semantic understanding error)</p>	<p>540B Model Output</p> <p>He travels at 10 mph from 1 to 4 PM. This means he travels 3 hours. 3 hours at 10 mph means he travels $3 \times 10 = 30$ miles. He then travels back at 6 mph. This means he travels 6 miles per hour. He has to travel 30 miles, so it takes him $30 / 6 = 5$ hours. The answer is 5. (Correct)</p>
<p>Question</p> <p>Stephen placed an online order for groceries. His final bill came to \$40.00. Because this was through a delivery vendor, they tacked on a 25% fee to his final total and charged him \$3.00 in delivery fees. Stephen also added a \$4.00 tip. After the extra fees, what was the final price of Stephen's groceries?</p>	<p>62B Model Output</p> <p>The final price of Stephen's groceries was \$40.00. The delivery vendor added 25% to the final total, so the final total was 125% of \$40.00. This means the final total was $125\% \times \\$40.00 = \\50.00. The delivery vendor also charged \$3.00 in delivery fees. So the final price of Stephen's groceries was $\\$50.00 + \\$3.00 = \\$53.00$. The answer is \$53.00. (Incorrect: one step missing error)</p>	<p>540B Model Output</p> <p>Stephen's bill was \$40.00. The delivery vendor added 25% to the bill. This means the bill was increased by 25% of \$40.00. This is $0.25 \times \\$40.00 = \\10.00. The delivery fee was \$3.00. The tip was \$4.00. So the total bill was $\\$40.00 + \\$10.00 + \\$3.00 + \\$4.00 = \\$57.00$. The answer is 57. (Correct)</p>
<p>Question</p> <p>There are four schools competing at a basketball tournament. Each school has sent a girls' basketball team and a boys' basketball team and each team has 5 players each. Each school has also sent a coach for each team. In total, how many people have all of the schools sent?</p>	<p>62B Model Output</p> <p>There are 4 schools. Each school has sent a girls' basketball team and a boys' basketball team. Each team has 5 players each. That means there are $4 \times 2 \times 5 = 40$ players. Each team has a coach. That means there are $40 + 4 = 44$ coaches. The answer is 44. (Incorrect: one step missing error)</p>	<p>540B Model Output</p> <p>Each school has sent 2 teams, each with 5 players. This means each school has sent $2 \times 5 = 10$ players. Each school has also sent 2 coaches. This means each school has sent $10 + 2 = 12$ people. There are 4 schools, so in total all of the schools have sent $4 \times 12 = 48$ people. The answer is 48. (Correct)</p>

Figure: Chains-of-thought prompt facilitates length generalization to longer sequence lengths.

Conclusion

- Chain-of-thought prompting is a versatile method for eliciting reasoning in LLMs.
- It significantly improves performance on arithmetic, commonsense, and symbolic reasoning tasks.
- The method is especially effective at large model scales (100B params and above).
- Future directions:
 - Explore other prompting methods to further expand the capabilities of LLMs.
 - Investigate ways to induce reasoning in smaller models.