

[BLOG](#) ›

Language Models Perform Reasoning via Chain of Thought

WEDNESDAY, MAY 11, 2022

Posted by Jason Wei and Denny Zhou, Research Scientists, Google Research, Brain team

In recent years, [scaling up the size of language models](#) has been shown to be a reliable way to improve performance on a range of natural language processing (NLP) tasks. Today's language models at the scale of 100B or more parameters achieve strong performance on tasks like [sentiment analysis](#) and machine translation, [even with little or no training examples](#). Even the [largest language models](#), however, can still struggle with certain multi-step reasoning tasks, such as math word problems and [commonsense reasoning](#). How might we enable language models to perform such reasoning tasks?

In "[Chain of Thought Prompting Elicits Reasoning in Large Language Models](#)," we explore a prompting method for improving the reasoning abilities of language models. Called *chain of thought prompting*, this method enables models to decompose multi-step problems into intermediate steps. With chain of thought prompting, language models of sufficient scale (~100B parameters) can solve complex reasoning problems that are not solvable with standard prompting methods.

Comparison to Standard Prompting

With *standard prompting* (popularized by [GPT-3](#)) the model is given examples of input–output pairs (formatted as questions and answers) before being asked to predict the answer for a test-time example (shown below on the left). In *chain of thought prompting* (below, right), the model is prompted to produce intermediate reasoning steps before giving the final answer to a multi-step problem. The idea is that a model-generated chain of thought would mimic an intuitive thought process when working through a multi-step reasoning problem. While producing a thought process has been [previously accomplished](#) via fine-tuning, we show that such thought processes can be elicited by

Standard Prompting	Chain of Thought Prompting
<p>Input</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: The answer is 11.</p> <p>Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?</p>	<p>Input</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>Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?</p>
<p>Model Output</p> <p>A: The answer is 27. ❌</p>	<p>Model Output</p> <p>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. ✅</p>

Whereas standard prompting asks the model to directly give the answer to a multi-step reasoning problem, chain of thought prompting induces the model to decompose the problem into intermediate reasoning steps, in this case leading to a correct final answer.

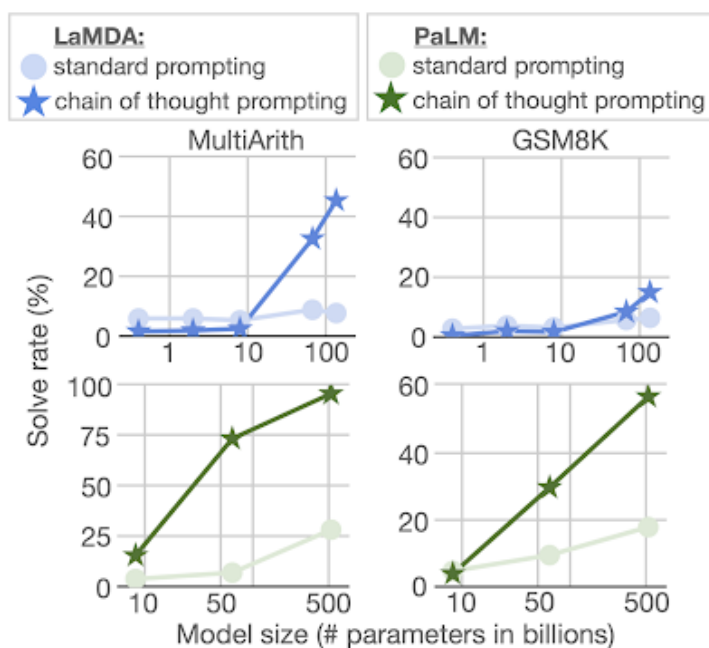
Chain of thought reasoning allows models to decompose complex problems into intermediate steps that are solved individually. Moreover, the language-based nature of chain of thought makes it applicable to any task that a person could solve via language. We find through empirical experiments that chain of thought prompting can improve performance on various reasoning tasks, and that successful chain of thought reasoning is an emergent property of model scale – that is, the benefits of chain of thought prompting only materialize with a sufficient number of model parameters (around 100B).

Arithmetic Reasoning

One class of tasks where language models typically struggle is arithmetic reasoning (i.e., solving math word problems). Two benchmarks in arithmetic reasoning are [MultiArith](#) and [GSM8K](#), which test the ability of language models to solve multi-step math problems similar to the one shown in the figure above. We evaluate both the [LaMDA collection](#) of language models ranging from 422M to 137B parameters, as well as the [PaLM collection](#) of language models ranging from 8B to 540B parameters. We manually compose chains of thought to include in the examples for chain of thought prompting.

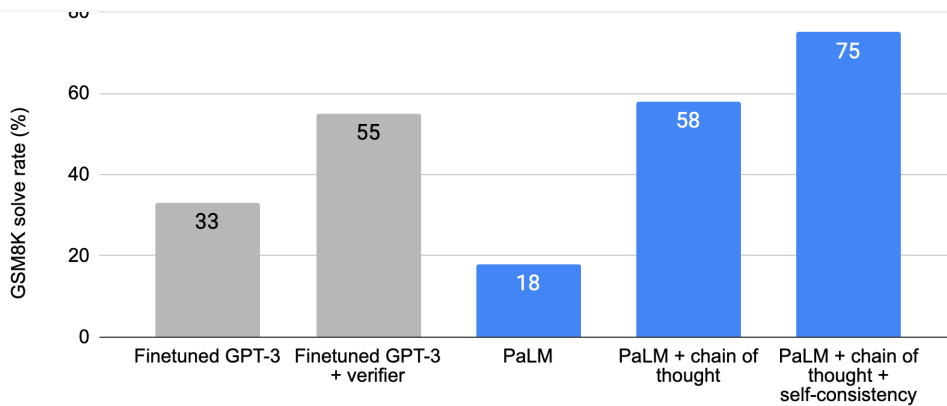
For these two benchmarks, using standard prompting leads to relatively flat scaling curves: increasing the scale of the model does not substantially

that substantially outperforms standard prompting for large model sizes.



Employing chain of thought prompting enables language models to solve arithmetic reasoning problems for which standard prompting has a mostly flat scaling curve.

On the GSM8K dataset of math word problems, PaLM shows remarkable performance when scaled to 540B parameters. As shown in the table below, combining chain of thought prompting with the 540B parameter PaLM model leads to new state-of-the-art performance of 58%, surpassing the prior state of the art of 55% achieved by fine-tuning GPT-3 175B on a large training set and then ranking potential solutions via a specially trained verifier. Moreover, [follow-up work](#) on self-consistency shows that the performance of chain of thought prompting can be improved further by taking the majority vote of a broad set of generated reasoning processes, which results in 74% accuracy on GSM8K.



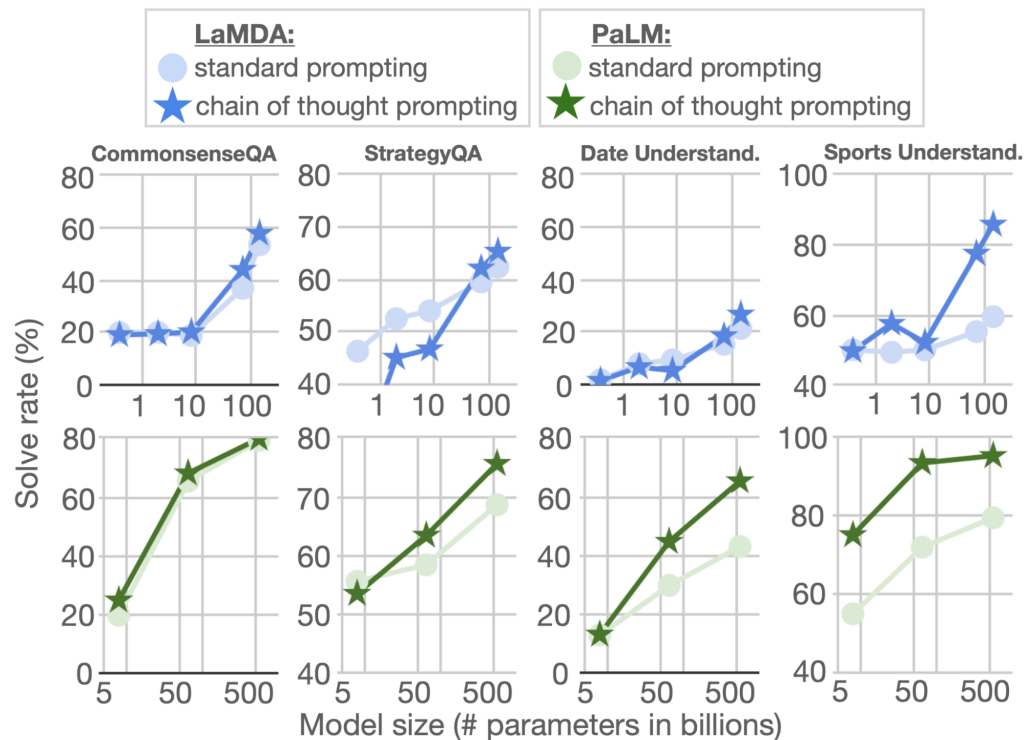
Chain of thought prompting with **PaLM** achieves a new state of the art on the **GSM8K** benchmark of math word problems. For a fair comparison against **fine-tuned GPT-3 baselines**, the chain of thought prompting results shown here also use an external calculator to compute basic arithmetic functions (i.e., addition, subtraction, multiplication and division).

Commonsense Reasoning

In addition to arithmetic reasoning, we consider whether the language-based nature of chain of thought prompting also makes it applicable to commonsense reasoning, which involves reasoning about physical and human interactions under the presumption of general background knowledge. For these evaluations, we use the **CommonsenseQA** and **StrategyQA** benchmarks, as well as two domain-specific tasks from **BIG-Bench collaboration** regarding **date understanding** and **sports understanding**. Example questions are below:

CommonsenseQA	StrategyQA	Date Understanding	Sports Understanding
<p>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</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>Q: Yes or no: Would a pear sink in water?</p> <p>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.</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>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>

As shown below, for CommonsenseQA, StrategyQA, and Date Understanding, performance improved with model scale, and employing chain of thought prompting led to additional small improvements. Chain of thought prompting had the biggest improvement on sports understanding, for which PaLM 540B's



Chain of thought prompting also improves performance on various types of commonsense reasoning tasks.

Conclusions

Chain of thought prompting is a simple and broadly applicable method for improving the ability of language models to perform various reasoning tasks. Through experiments on arithmetic and commonsense reasoning, we find that chain of thought prompting is an emergent property of model scale. Broadening the range of reasoning tasks that language models can perform will hopefully inspire further work on language-based approaches to reasoning.

Acknowledgements

It was an honor and privilege to work with Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Quoc Le on this project.



Previous posts

MAY 11, 2022

Unlocking Zero-Resource Machine Translation to

[→](#)

MAY 5, 2022

Learning Locomotion Skills Safely in the Real

[→](#)

MAY 4, 2022

GraphWorld: Advances in Graph Benchmarking

[→](#)