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

💡 **Paper link:** https://arxiv.org/pdf/2305.15507.pdf

💡 **Related Article:** https://open.substack.com/pub/aiguide/p/can-large-language-models-reason?r=b0ibt&utm_campaign=post&utm_medium=web

Brief Summary

Chain of thought as a way to improve the scope of what LLMs can do. What is chain of thought? Adding intermediate steps in the form of natural language to improve the model performance. Does not answer much about reasoning. The paper basically says that the capabilities of LLMs expand in various domains such as arithmetic, commonsense, symbolic reasoning as we increase the size of the model and use chain of thought.

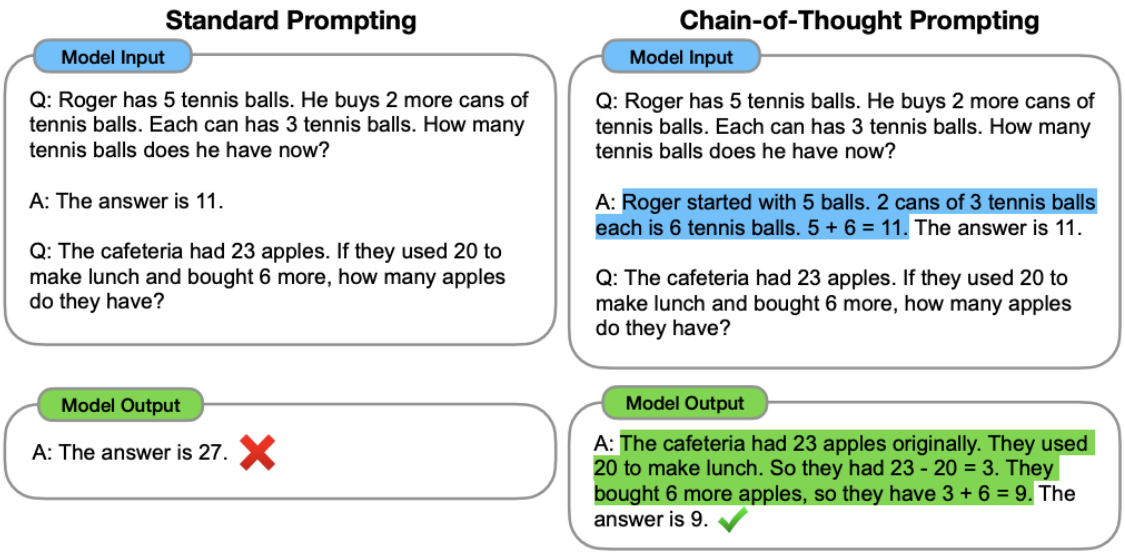


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Detailed summary

Natural Language Processing has been revolutionized by the use of LLMs. If you are using any AI product these days, such as ChatGPT, chances are that LLM is the technology behind it. It has been observed that as we scale up the model, i.e. increase the number of model parameters, increase the amount of training data, the model helps in reasoning but is still not enough for arithmetic tasks. Arithmetic reasoning is a task which is easy for humans but models suffer a lot in them. This paper makes use of natural language to improve performance on arithmetic tasks. The paper takes advantage of few-shot learning, which means that we do not need to fine tune model to a specific task (arithmetic reasoning is this case), but just providing a few sample input-output exemplar demonstrating the task will make the model learn. Few shot learning is a very helpful methodology which saves the cost of creating high set of rational datasets to be used for fine tuning the model. The method proposed by the authors consists of a prompt<input, chain of thought, output>

Chain of thought is an intermediate step that explain the process involved for taking input to the output. It has been shown in the paper that sufficiently large language models can generate chains of thought if demonstrations of chain-of-thought reasoning are provided in the exemplars for few-shot prompting. Chain of thought allows the

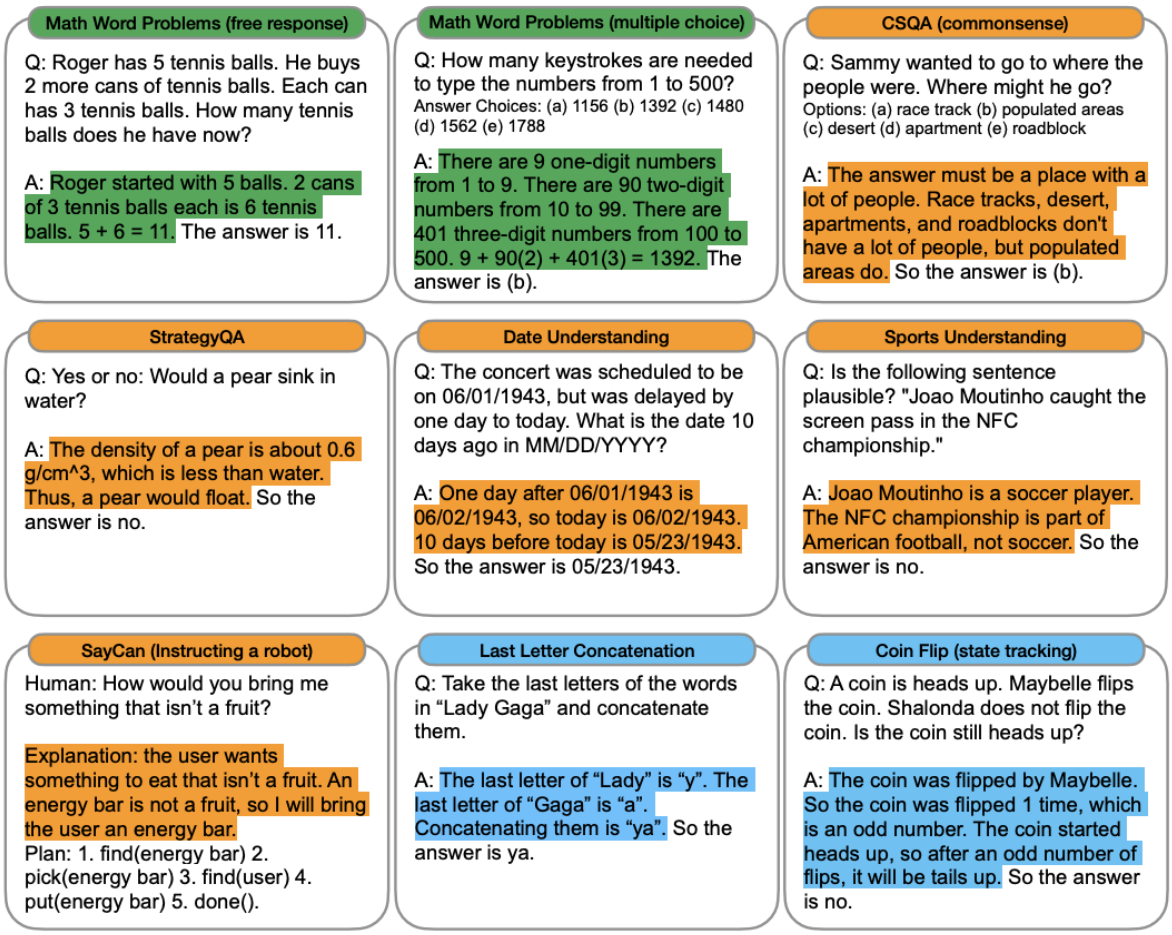
model to decompose the problem into multi step so that computation can be allotted to problems that require more reasoning steps. This is also a very interesting method as it explains the model and is applicable to any task that humans can solve using language

The benchmark mathematical problem sets considered for the paper: GSM8K, SVAMP, ASDiv, AQuA, MAWPS

The paper mentions two types of prompting -

Standard prompting: Consider standard few-shot prompting in which language model is given in-context exemplars of input–output pairs before outputting a prediction for a test-time example. There is no chain of thought input here.

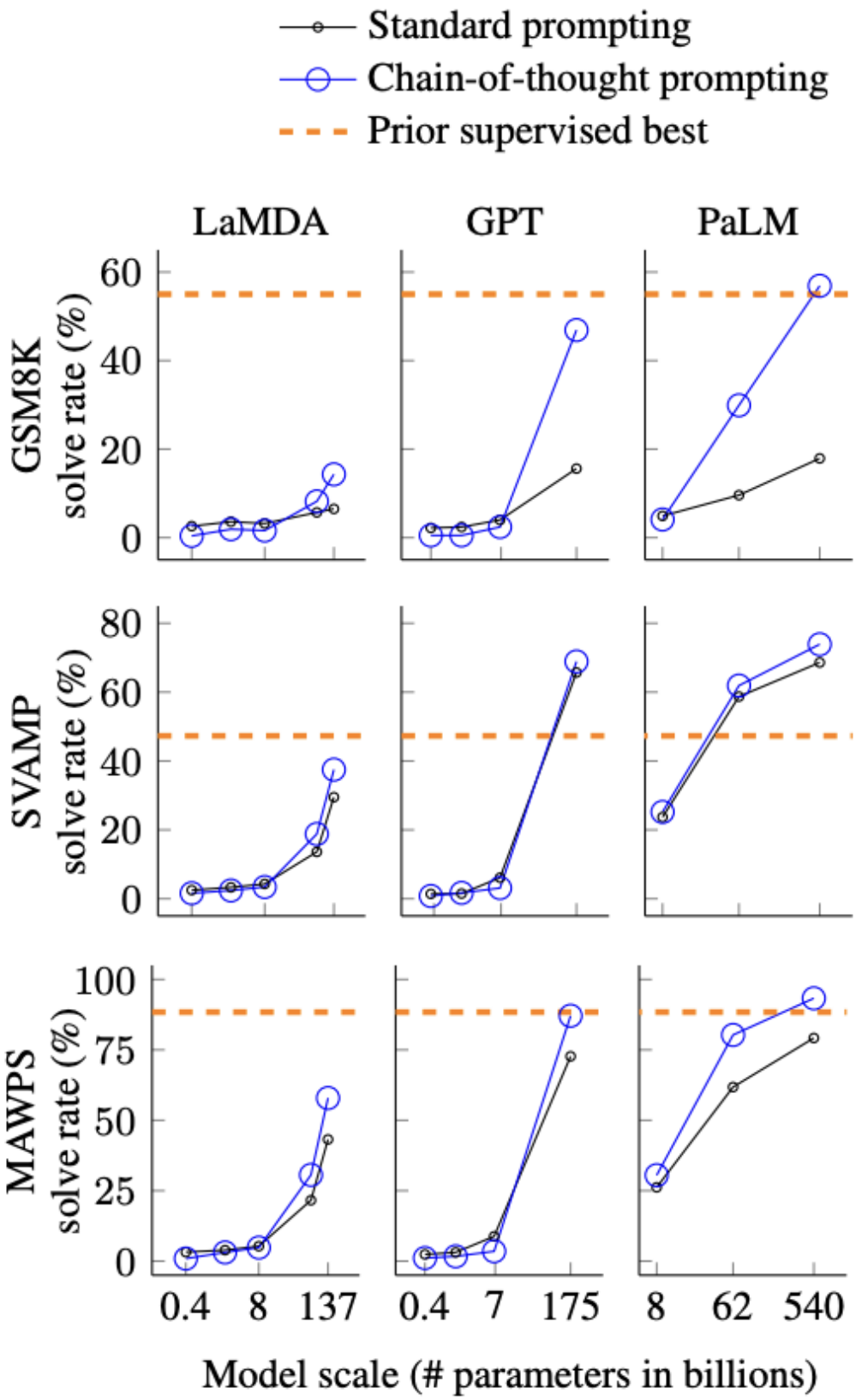
Chain of thought prompting: Augment each exemplar with chain of thought in few shot learning.



The evaluation is done on 5 large models namely - GPT-3, LaMDA, UL2 20B, Codex

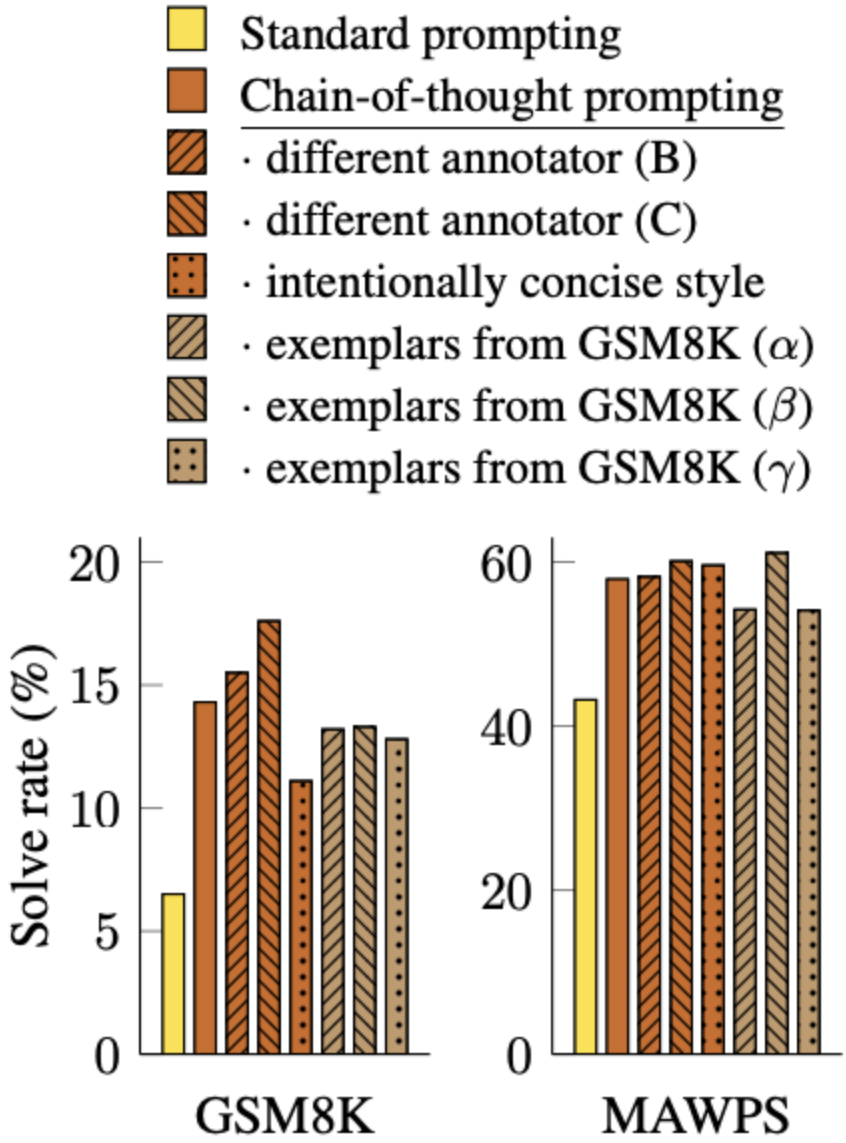
RESULTS -

1. Not useful for smaller models as the performance is not impacted, qualitatively found that models of smaller scale produced fluent but illogical chains of thought, leading to lower performance than standard prompting
2. Chain-of-thought prompting has larger performance gains for more-complicated problems.
3. Scaling improves the chain of thought i.e. larger the model size, the better quality chain of thoughts produced.


Standard prompting
Chain-of-thought prompting
Prior supervised best
LaMDA
GPT
PaLM
GSM8K solve rate (%)
SVAMP solve rate (%)
MAWPS solve rate (%)
0
20
40
60
80
100
25
50
75
0.4
8
137
0.4
7
175
8
62
540
Model scale (# parameters in billions)

The diagram above clearly shows that for larger model sizes (model scale), chain-of-thought prompting surpasses the benchmark. While standard prompting does improve performance as the size increases, it falls short of both the benchmark and the chain-of-thought prompting method. It can also be observed that when the model scale is less, the performance of chain-of-thought prompting is similar to that of standard prompting because of the quality of chain of thoughts generated by small models.

The paper also mentions different ablation studies that highlight how chain-of-thought is the only variable in this work responsible for improving the model performance as the model scales up. Feel free to refer the paper to understand the ablation study methods used.



The above graph is for LaMDA trained on 137Billion parameters. It can be clearly seen that all the chain-of-thought methods perform better than the standard prompting although they differ by some margin owing to different thought styles written by different persons. This implies that successful use of chain of thought does not depend on a particular linguistic style. In addition to robustness to annotators, independently-written chains of thought, different exemplars,
and various language models, the paper also finds that chain-of-thought prompting for arithmetic reasoning is robust to different exemplar orders and varying numbers of exemplars

Apart from mathematic reasoning tasks, the paper also shows the use of chain-of-thought prompting on Commonsense reasoning and Symbolic reasoning and shows that chain-of-thought method does take the model evaluation past the previous benchmark results.