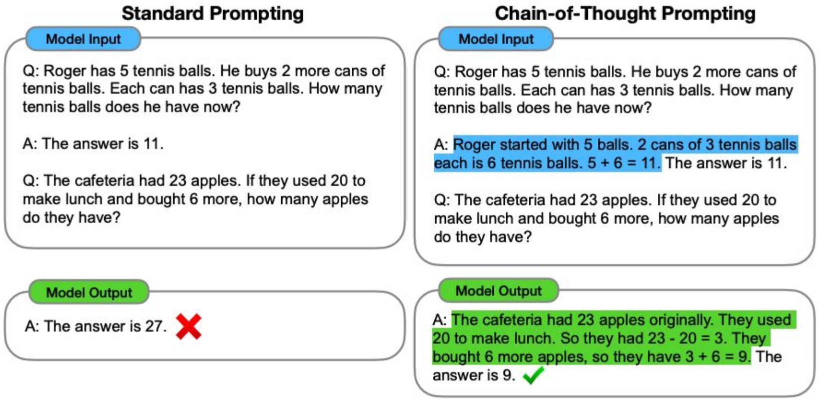
Chain-of-Thought Prompting Elicits Reasoning in LLMs

Jason Wei et.al., 2022

Chain of Thought

- A series of intermediate reasoning steps to improve reasoning capabilities
- Usually few exemplars are provided during prompting



Standard Prompting

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?

A: The answer is 27. ✗

Chain-of-Thought Prompting

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?

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. ✓

Motivation - Prior Work

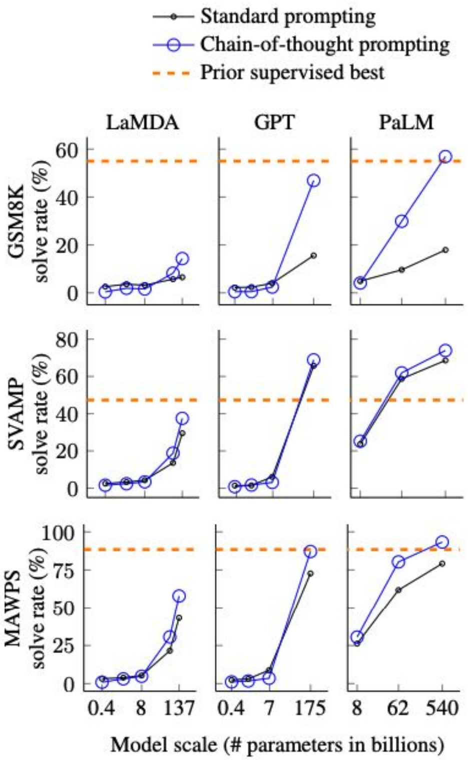
- Scaling up LLM size alone does not improve performance
- 2 main motivations:
  - Arithmetic reasoning can benefit from generating natural language rationales
  - LLMs offer the exciting prospect of in-context few-shot learning via prompting
- Combining these 2 approaches:
  - Few shot prompt that consists of triples: <input, chain of thought, output>
- Does not require a large training dataset
- Single model checkpoint can perform many tasks without loss of generality

Benefits

1. Decompose multi-step problems into intermediate steps: additional computation can be allocated to problems that require more reasoning steps.
2. Provides an interpretable window into the behavior of the model
3. Used for tasks such as math word problems, commonsense reasoning, and symbolic manipulation
4. Can be readily elicited in huge LLMs simply by including examples of CoT sequences into the exemplars of few-shot prompting

Results

- Only yields performance gains when used with models of ~100B parameters
- Models of smaller scale produced fluent but illogical chains of thought



Commonsense Reasoning

- Although chain of thought is particularly suitable for math word problems, it’s language based nature makes it applicable to a broad class of commonsense reasoning problems.
- Results:

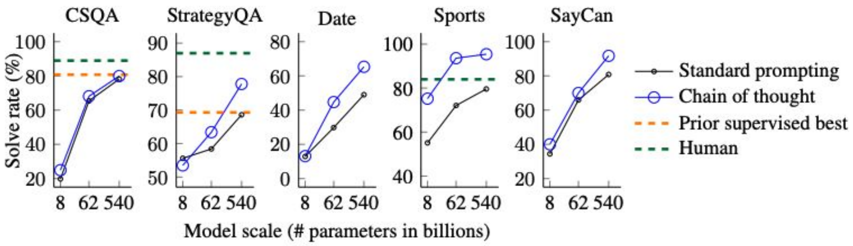


Figure 7: Chain-of-thought prompting also improves the commonsense reasoning abilities of language models. The language model shown here is PaLM. Prior best numbers are from the leaderboards of CSQA (Talmor et al., 2019) and StrategyQA (Geva et al., 2021) (single-model only, as of May 5, 2022). Additional results using various sizes of LaMDA, GPT-3, and PaLM are shown in Table 4.

Symbolic Reasoning

- Simple for humans but potentially challenging for language models
- Tasks
  - Last letter concatenation
  - Coin flip

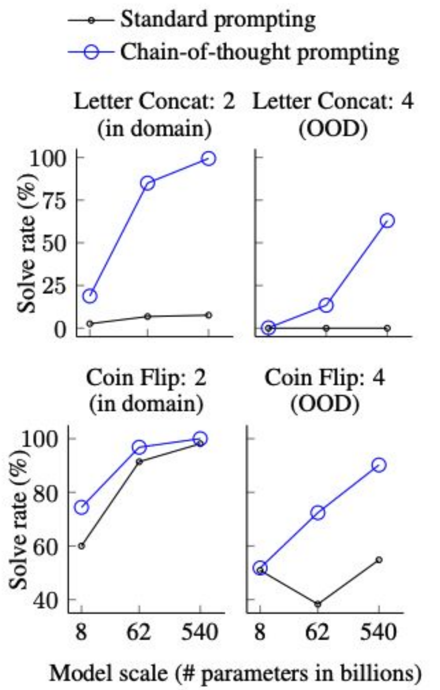


Figure 8: Using chain-of-thought prompting facilitates generalization to longer sequences in two symbolic reasoning tasks.

References

- https://arxiv.org/pdf/2201.11903
- https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf