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

Jason Wei, et al

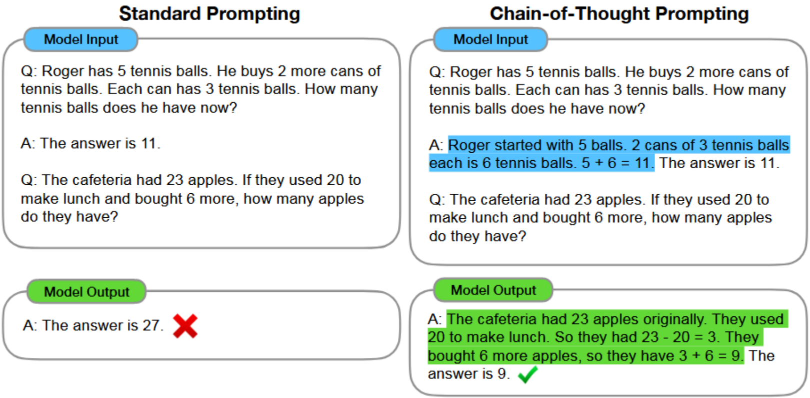
Google Research, Brain Team

NeurIPS 2022

Abstract

- *Chain-of-Thought Prompting*
  - Reasoning ability emerges naturally in large models
- Reasoning Experiments
  - Arithmetic / Commonsense / Symbolic
  - GPT-3, LaMDA, PaLM, etc.
  - For each model, variate parameter size
  - Compare standard prompting, CoT, and supervised SOTA
- Results
  - As model becomes larger, CoT outperforms SOTA
  - Limitations: Mimicking, Manual exemplar cost

Abstract



- Reasoning abilities emerge naturally via Chain-of-Thought (CoT)
- Experiments on arithmetic, commonsense, and symbolic reasoning tasks

Introduction

- Few-shot prompting [1]
  - We could perform unsupervised learning
  - However, it was poor at 'Reasoning'

- *Chain-of-thought*
  - [Input, Chain of thought, Output] is given
  - Decomposes problem into subproblems
  - Provides interpretable window for debugging
  - Potentially applicable for any tasks

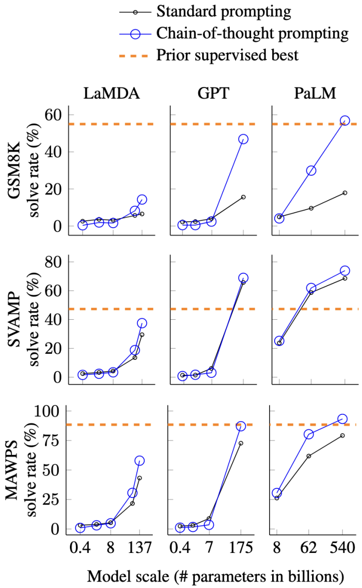
---

[1] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

Arithmetic Reasoning - Result

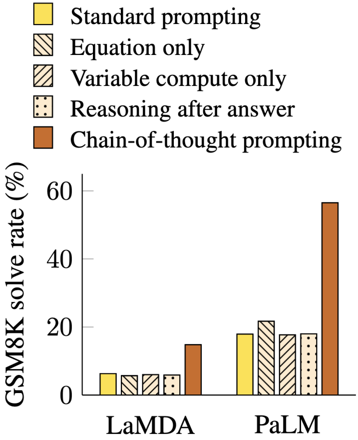
- Benchmarks
  - GSM8K
  - SVAMP
  - ASDiv
  - AQuA
  - MAWPS

- CoT is efficient when
  - model is larger
  - problems are complicated



Arithmetic Reasoning - Ablation

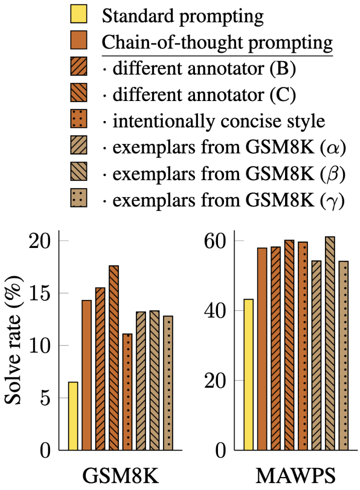
- Equation only
  - Only helpful for short problems
- Variable compute only
  - Length of CoT is given
  - No improvement
- Reasoning after answer
  - No improvement



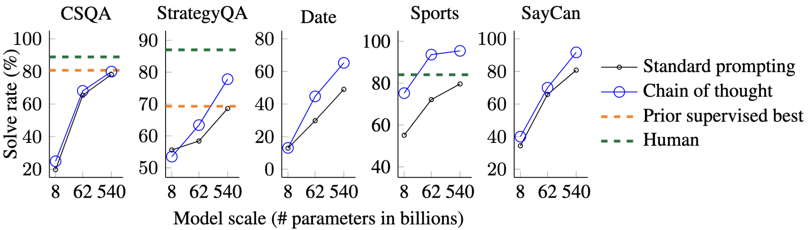
Arithmetic Reasoning - Robustness

- Sensitivity to exemplars
  - Different exemplar annotators
  - GSM8K Training exemplars

- Result
  - All CoT outperformed standard
  - CoT is robust to linguistic style



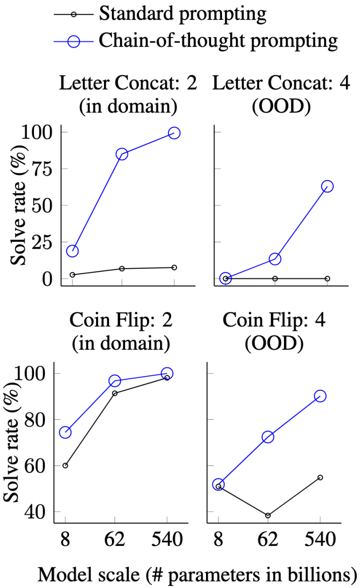
Commonsense Reasoning



- Benchmarks
  - CSQA, StrategyQA, Date, Sports, SayCan
- Experimental setup
  - Same as prior section

Symbolic Reasoning

- Last letter concatenation
  - "Amy Brown" → yn
  - Challenging than first letter
- Coin flip
  - People flip or don't flip the coin
  - Asks the model to answer whether a coin is still heads up
- Results
  - Performed well even in OOD
  - Length generalization by CoT



Conclusion

- Chain-of-Thought performance
  - Standard prompting has a flat scaling curve
  - CoT has dramatically increasing scaling curves
- It raises more questions
  - How much improvement with a further increase in model scale?
  - What other prompting methods would be there?
- Limitation
  - Open question: Is it actually "Reasoning"?
  - Cost of manually augmenting exemplars
  - There is no guarantee of correct reasoning paths
  - CoT appears only at large models