Large Language Models are Zero-Shot Reasoners

Takeshi Kojima, et al

The University of Tokyo, Google Research

NeurIPS 2022

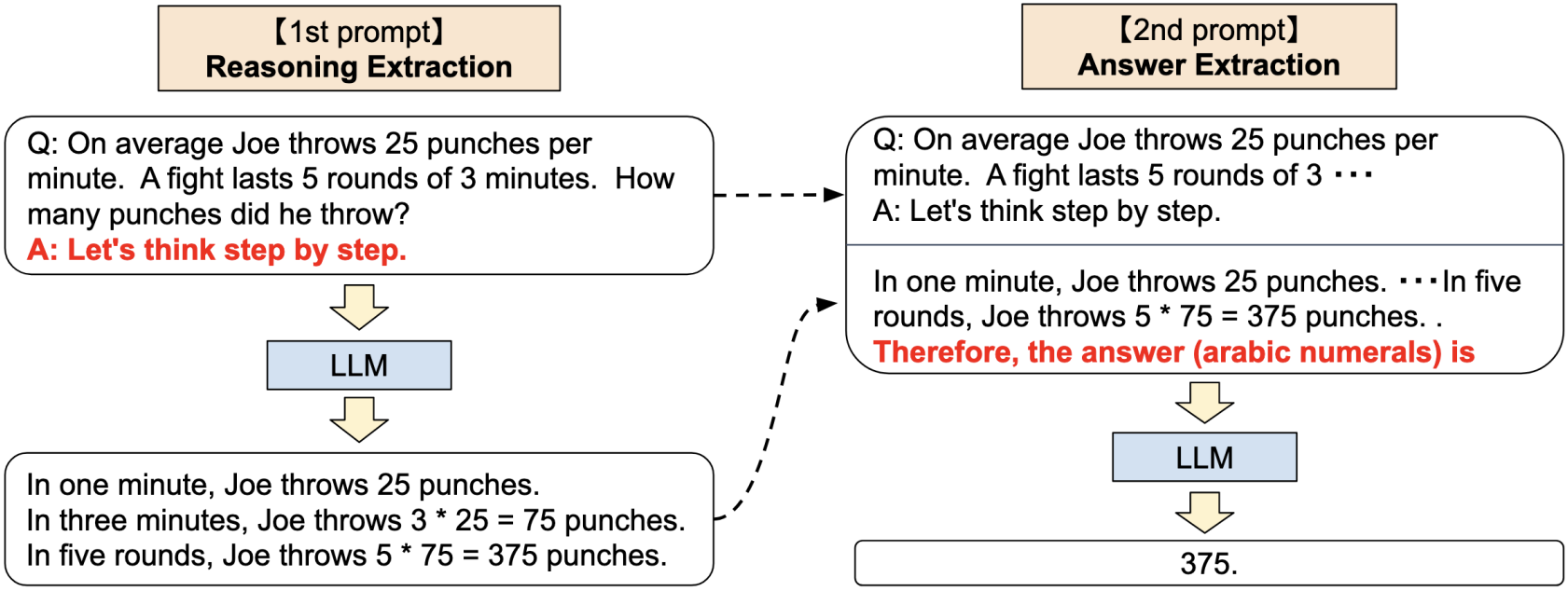
Abstract

- Background
  - Large scale → Prompting → CoT
- Zero-Shot CoT
  - “Let’s think step by step”
- Experiment
  - Zero-Shot, Few-Shot
  - Zero-Shot-CoT, Few-Shot-CoT
  - Zero-Plus-Few-Shot-CoT
  - Self-consistency
- Discussion
  - It is minimalist and elicits CoT

Introduction

- CoT's success
  - Step-by-step example
  - Manually creating example is costly
- Zero-shot CoT
  - Adding a prompt, "Let's think step by step"
  - Simple but strong

Zero-shot Chain of Thought



- 1st prompt: reasoning extraction
  - Input question X, reasoning trigger T
  - Prompting as "Q: [X]. A: [T]"
- 2nd prompt: answer extraction
  - 1st prompt X', Generated Z, answering trigger A
  - Prompting as "[X'] [Z] [A]"

Experiment

- Tasks (Total 12 datasets)
  - Arithmetic reasoning
  - Commonsense reasoning
  - Symbolic reasoning
  - Other reasoning tasks

- Models (Total 17 models)
  - GPT Variations (Instruct-GPT3)
  - PaLM (8B-540B)
  - Other models

Background

- InstructGPT [1]
  - Prompts submitted through the OpenAI API
  - Demonstrations of the desired model behavior
  - Supervised fine-tuning with the dataset
  - 1.3B InstructGPT was preferred to 175B GPT-3

- Self-consistency [2]
  - Different decoding method from greedy-one
  - Get multiple CoTs with temperature-based sampling
  - Aggregate the answers using majority voting

---

[1] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744.

[2] Xuezhi Wang et al. “Self-consistency improves chain of thought reasoning in language models”. In: *arXiv preprint arXiv:2203.11171* (2022).

Zero-shot-CoT vs Zero-shot

	Arithmetic					
	SingleEq	AddSub	MultiArith	GSM8K	AQUA	SVAMP
zero-shot	74.6/**78.7**	**72.2/77.0**	17.7/22.7	10.4/12.5	22.4/22.4	58.8/58.7
zero-shot-cot	**78.0/78.7**	69.6/74.7	**78.7/79.3**	**40.7/40.5**	**33.5/31.9**	**62.1/63.7**
	Common Sense		Other Reasoning Tasks		Symbolic Reasoning	
	Common SenseQA	Strategy QA	Date Understand	Shuffled Objects	Last Letter (4 words)	Coin Flip (4 times)
zero-shot	**68.8/72.6**	12.7/**54.3**	49.3/33.6	31.3/29.7	0.2/-	12.8/53.8
zero-shot-cot	64.6/64.0	**54.8**/52.3	**67.5/61.8**	**52.4/52.9**	**57.6**/-	**91.4/87.8**

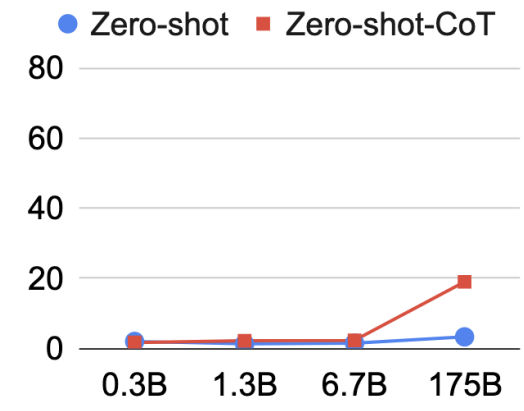
- GPT-3 Results
  - Usually, Zero-shot-CoT outperforms
  - No performance gain in commonsense reasoning
  - PaLM(540B) is expected to solve the problem

Comparison with baselines

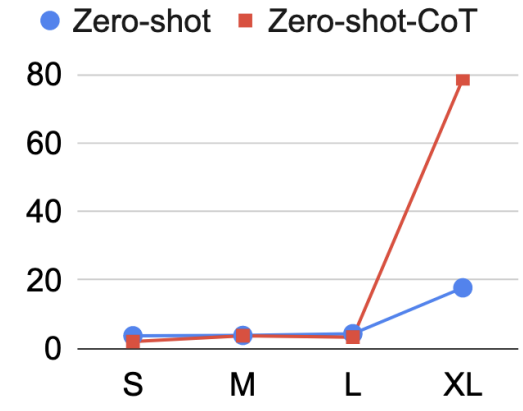
	MultiArith	GSM8K
Zero-Shot	**17.7**	**10.4**
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	**78.7**	**40.7**
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	**92.8**	**51.5**
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
PaLM 540B: Zero-Shot	**25.5**	**12.5**
PaLM 540B: Zero-Shot-CoT	**66.1**	**43.0**
PaLM 540B: Zero-Shot-CoT + self consistency	**89.0**	**70.1**
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4

- Performance increases with CoT

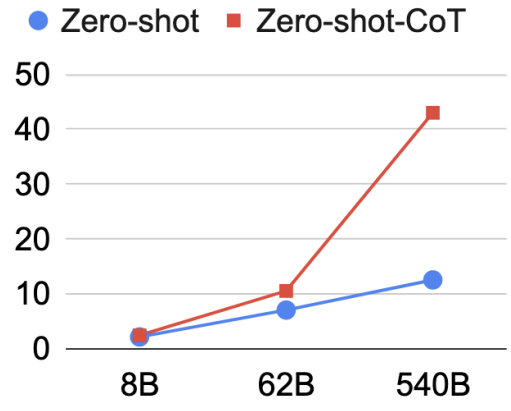
Model scale study



(a) MultiArith on Original GPT-3



(b) MultiArith on Instruct GPT-3



(c) GMS8K on PaLM

- CoT is not effective when model is small
- Performance drastically increases as model gets bigger

Robustness study

No.	Category	Template	Accuracy
1	instructive	Let's think step by step.	**78.7**
2		First, (*1)	77.3
3		Let's think about this logically.	74.5
4		Let's solve this problem by splitting it into steps. (*2)	72.2
5		Let's be realistic and think step by step.	70.8
6		Let's think like a detective step by step.	70.3
7		Let's think	57.5
8		Before we dive into the answer,	55.7
9		The answer is after the proof.	45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		Abrakadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

- Future research: Automatically creating better templates

Conclusion

- Reasoning Ability of LLMs
  - Performance can be increased by 3 ways orthogonally
  - Fine-tuning, Prompting, Step-by-step reasoning
- Limitation
  - LLM amplify biases found in the training data