LLMs are Zero-Shot Reasoners

Kim Choeun

August 21, 2024

Seoul National University

Outline

1. Introduction
2. Zero-shot Chain of Thought
3. Experiment
4. Conclusion

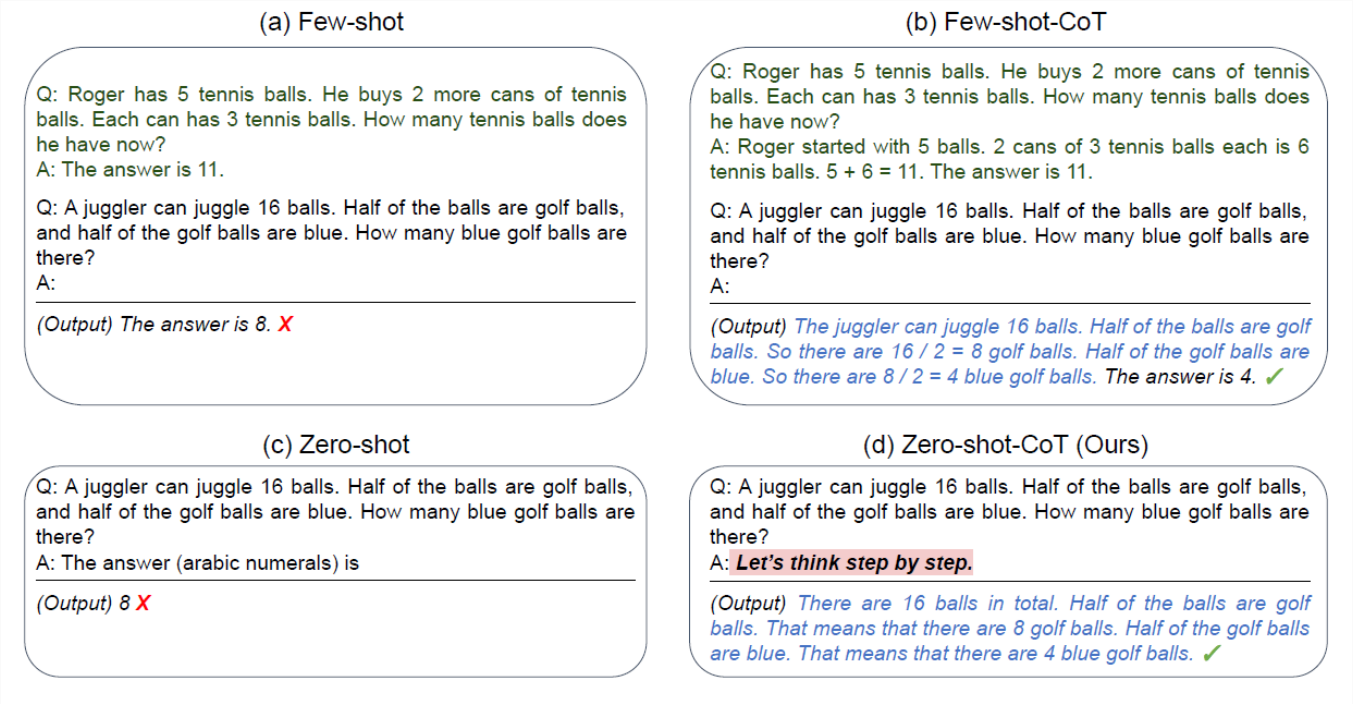
Introduction

LLMs and prompting

- A LM is a model that looks to estimate the probability distribution over text.
- The success of LLMs is often attributed to (in-context) few-shot or zero-shot learning.
- In this paradigm, instead of adapting pre-trained LMs to downstream tasks via objective engineering, downstream tasks are reformulated to look more like those solved during the original LM training **with the help of a textual prompt.**
- For example, if we choose the prompt "English: I missed the bus today. French: ", an LM may be able to fill in the blank with a French translation.

CoT prompting

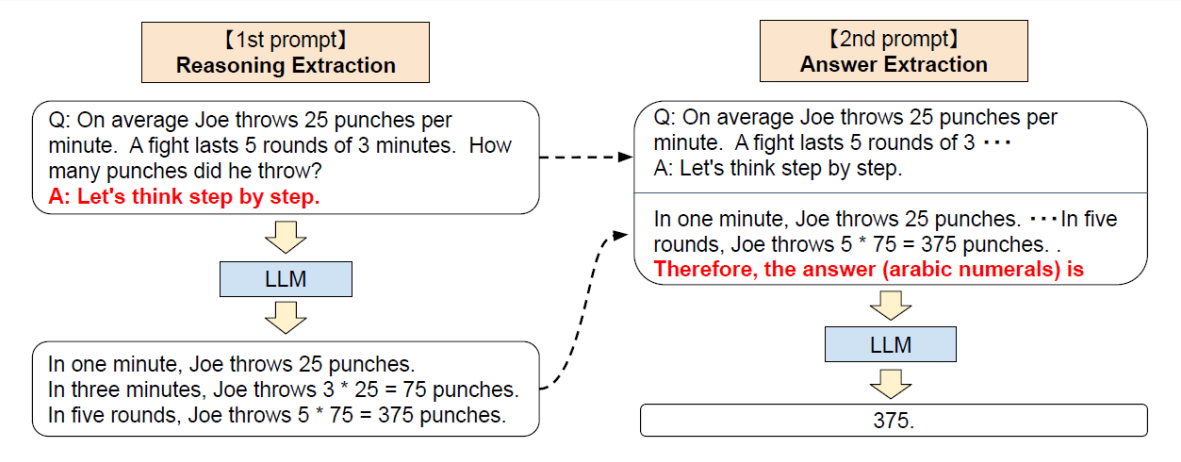
- Multi-step arithmetic and logical reasoning benchmarks have particularly challenged the scaling laws of LLMs.
- Chain of Thought (CoT) prompting proposed a simple solution by modifying the answers in few-shot examples to step-by-step answers.



Zero-shot Chain of Thought

Two stage prompting

- Proposed Zero-shot-CoT by simply adding *Let's think step by step*, or a similar text before the answer.
- Extract step-by-step reasoning.



1. **Reasoning Extraction** `Q:[X].A:[T]` where `[X]` is an input slot and `[T]` is an slot for hand-crafted trigger sentence.
2. **Answer Extraction** `[X']` `[Z]` `[A]` where `[X']` is a 1st prompt, `[Z]` is generated sentence and `[A]` is a trigger sentence to extract answer.

Experiment

Settings

- **Tasks and Datasets** : four categories of reasoning tasks; arithmetic, commonsense, symbolic, and other logical reasoning tasks.
- **Models** : 17 models in total. Instruct GPT3, original GPT3, PaLM, etc.
- **Baselines** : Zero-shot prompting, Few-shot prompting, Few-shot CoT prompting. Used greedy decoding across all the methods.
- **Answer cleansing** : after the model outputs a text by answer extraction, pick up only the part of the answer text that first satisfies the answer format.

Results

Zero-shot-CoT vs. Zero-shot

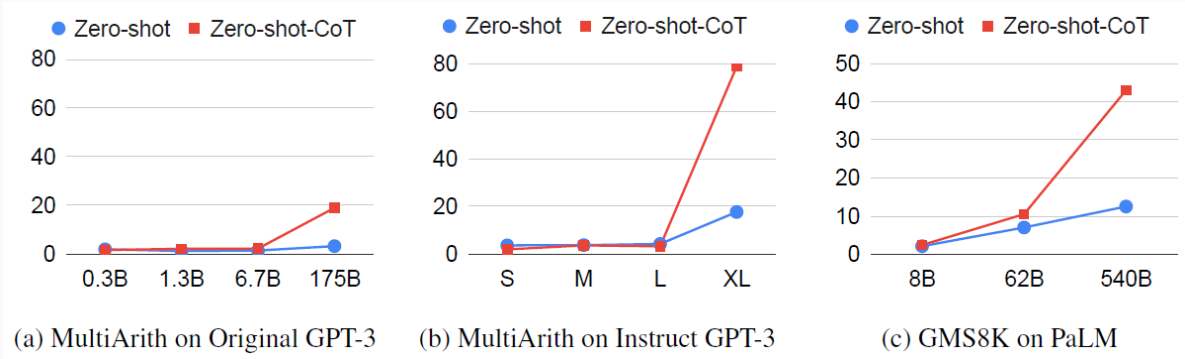
	Arithmetic					
	SingleEq	AddSub	MultiArith	GSM8K	AQUA	SVAMP
zero-shot	74.6/**78.7**	**72.2/77.0**	17.7/22.7	10.4/12.5	22.4/22.4	58.8/58.7
zero-shot-cot	**78.0/78.7**	69.6/74.7	**78.7/79.3**	**40.7/40.5**	**33.5/31.9**	**62.1/63.7**

	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**

- Zero-shot-CoT achieved score gains on arithmetic, symbolic reasoning and other reasoning tasks.
- In commonsense reasoning tasks, zero-shot CoT does not provide performance gains.
- However, many generated chain of thought themselves are surprisingly logically correct or only contains human-understandable mistakes.

Results

	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



(a) MultiArith on Original GPT-3 (b) MultiArith on Instruct GPT-3 (c) GMS8K on PaLM

The performance drastically increases with chain of thought reasoning, as the model size gets bigger, for Original/Instruct GPT-3 and PaLM.

Results

	Zero-shot	Few-shot-CoT †	Zero-shot-CoT	Few-shot-CoT
AQUA-RAT	22.4	31.9	33.5	39.0
MultiArith	17.7	27.0	78.7	88.2

- The table shows the performance of few-shot-CoT when using examples from different datsets: CommonsenseQA to AQUA-RAT and CommonsenseQA to MultiArith.
- The chain of thought examples from different domains but with the same answer format provide substantial performance gain over Zero-shot (to AQUA-RAT).
- For both cases the results are worse than Zero-shot-CoT, affirming the importance of task-specific sample engineering in Few-shot-CoT.

Conclusion

Conclusion

- Proposed a single zero-shot prompt that elicits chain of thought from large language models across a variety of reasoning tasks, in contrast to the few-shot approach that requires hand-crafting few-shot examples per task.
- The method not only is the minimalist and strongest zero-shot baseline for difficult multi-step system-2 reasoning tasks.