

# LLM's Performance regarding the number of Reasoning Steps in CoT

Original Paper : The Impact of Reasoning Step Length on Large Language Models (ACL 2024)

---

Presented by

Yoojun Lee  
Jewoo Park  
Lowren Wu

# CONTENTS

---

01 Paper Summary

---

02 Experiments

---

04 Result Analysis

---

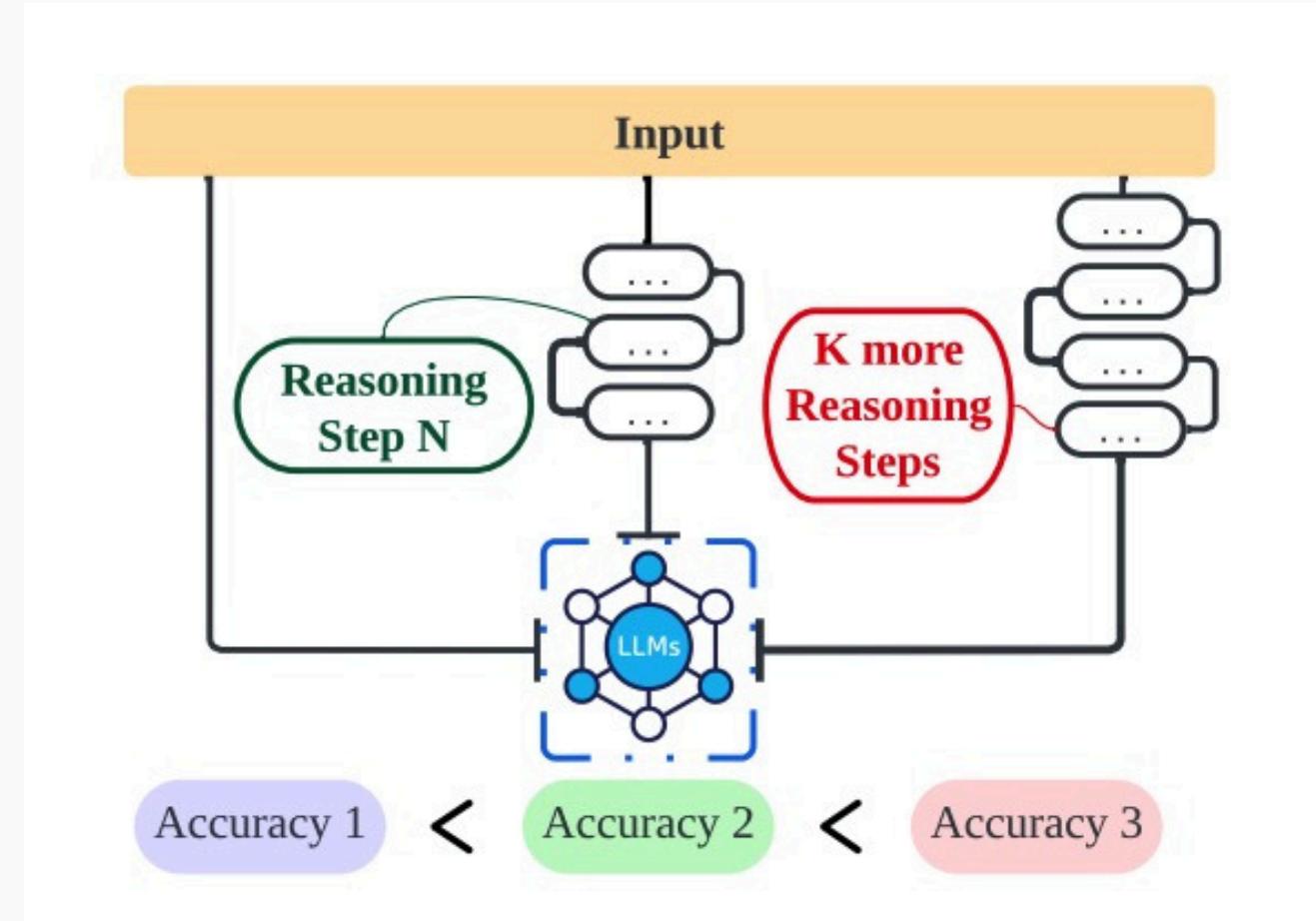
05 QnA

---

01

# PAPER SUMMARY

# INTRODUCTION



- CoT(Chain of Thought) prompting shows remarkable performance in LLM reasoning.
- However, nothing was done regarding **WHAT MAKES CoT PERFORM WELL.**
- Motivation of this work is to **NECESSITY TO UNDERSTAND CoT's INTERNAL MECHANISM**

**Hypothesis : The number of reasoning steps is the main factor that determine the performance of CoT prompting**

# INTRODUCTION

## Contribution

- In both Zero Shot CoT and Few Shot CoT, the accuracy of LLM increases as the number of reasoning steps increases.
- If the reasoning steps decreases, the accuracy gets worse linearly.
- Even if the wrong information is included in the reasoning steps, still the performance is improved.

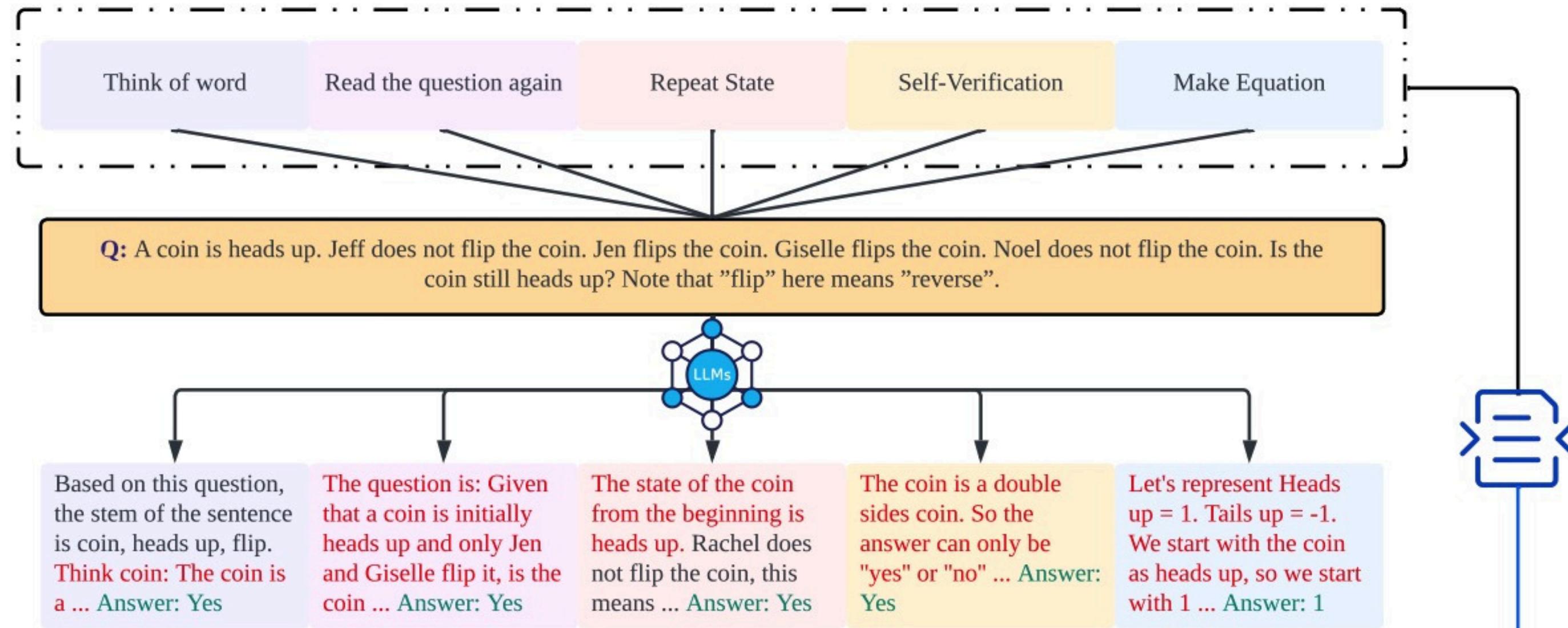
# METHODS

- **Contributions**
  - Provide question and choices
- **Zero Shot**
  - “Let’s Think Step by Step” / “Let’s Think Step by Step, and should think more step”
- **Few Shot (Auto CoT/Manual CoT)**
  - Provide reasoning example. To extend reasoning steps, include more steps in the example.
  - Strictly refrain from adding new information in the example.
  - The examples are imitating human’s reasoning process.

ME

- No Co
- Zero
- “
- Few
- P
- St
- Tl

### Reasoning Steps Expansion of CoT Rationales



### Reasoning Steps Compression of CoT Rationales

# METHODS

## Few Shot Example

Table 1: Case of Think About The Word, select a few words and ask the model to think about the meaning in text.

Think About The Word
<b>Prompt:</b> Q: Could someone in Tokyo take a taxi to the Metropolitan Museum of Art? A: Let's think step by step. The stem of the sentence is Tokyo, take a taxi, Metropolitan Museum of Art. <b>Think about Tokyo...</b> <b>Think about taking a taxi...</b> <b>Think about Metropolitan Museum of Art...</b> Inference: Tokyo is in Japan and the Metropolitan Museum of Art is in New York. The two places are separated by the sea, so you can't take a taxi there. The answer is no.
Q: Could a chipmunk fit 100 chocolate chips in his mouth?
<b>Answer:</b> A: Based on this question, the stem of the sentence is chipmunk, fit, 100 chocolate chips, mouth. <b>Think about chipmunk...</b> <b>Think about fit...</b> <b>Think about chocolate chips...</b> Inference: Chipmunks are small animals with relatively small mouths, so it's unlikely that a chipmunk would be able to fit 100 chocolate chips in its mouth at once. The answer is no.
<b>Prediction:</b> no
<b>Ground Truth:</b> no

# EXPERIMENTAL SETTINGS (PAPER'S)

**Dataset :** MultiArith, GSM8K, AQuA, SingleEq, SVAMP, Letter, Coin, StrategyQA

**Models :** text-davinci-002, GPT-3.5, GPT-4

**Prompting :** Zero Shot, Zero Shot CoT, Manual CoT, Auto CoT

**Metric :** Accuracy

# RESULTS

Table 2: Comparison of accuracy of our method with four baselines on eight datasets

Models	Arithmetic					Commonsense		Symbolic	
	MultiArith	GSM8K	AQuA	SingleEq	SVAMP	Strategy	Letter	Coin	
Zero-Shot	40±1.0	30.4±1.7	29.9±1.8	82.7±1.3	56±1.0	59±1.0	43±1.0	79.8±1.2	
Zero-Shot-CoT	91.5±1.2	64.1±1.1	55.6±1.3	87.43±0.25	79.99±1.41	58.34±1.56	69±1.0	93±1.0	
Manual-CoT	93.5±0.1	64.7±1.5	55±1.0	92.1±0.2	82.3±0.3	65.3±1.4	75±0.0	92.7±0.1	
Auto-CoT	94±0.0	65.8±0.9	65±0.0	92±0.0	81.9±0.3	65.3±0.5	73.5±0.2	93±0.0	
Must Think More Step (Zero-Shot-CoT)	95.2±0.3	76.1±0.1	62.11±0.24	87.43±0.16	79.99±0.18	72.6±0.21	69±0.0	97±0.0	
Add Reasoning Step (Manual-CoT)	97±0.0	70.1±0.3	62.5±0.5	88.97±0.27	85.2±0.2	68.86±0.27	77.8±0.4	97.3±0.1	
Add Reasoning Step (Auto-CoT)	97.2±0.1	78.8±0.2	64.03±0.36	92.71±0.14	83.7±0.2	70.26±0.19	71.2 ±0.5	99±0.0	

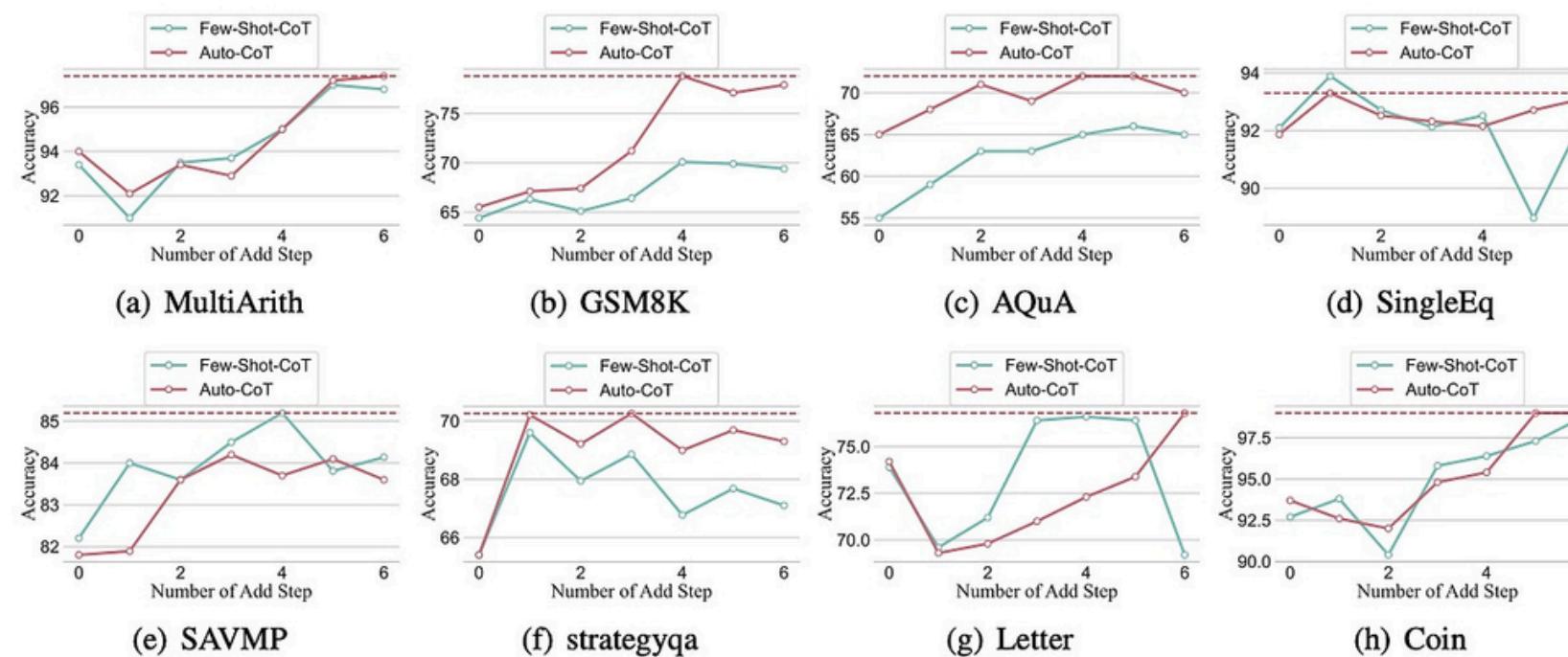
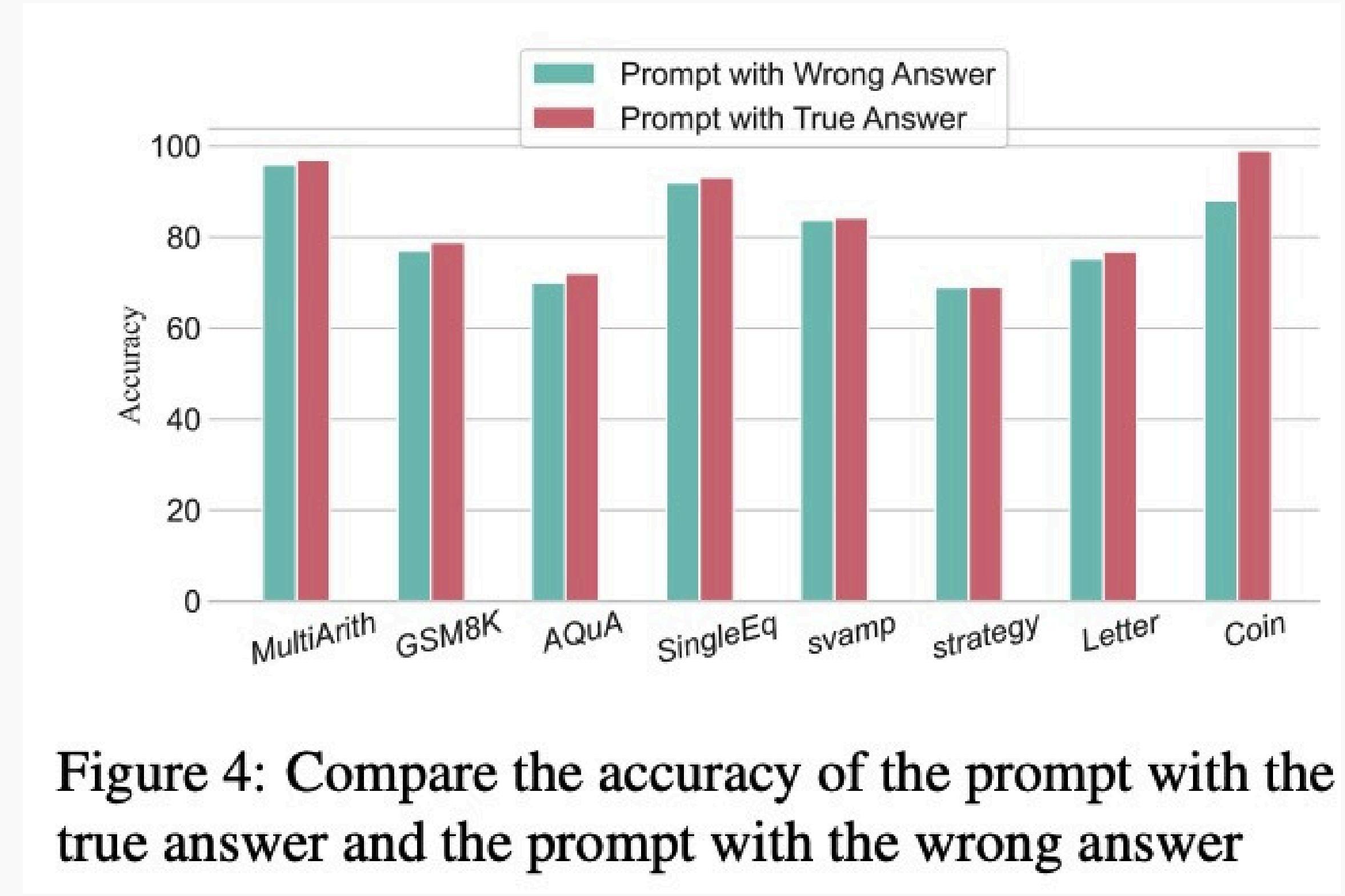


Figure 3: Linear Relationship Between Step Quantity and Accuracy

# RESULTS



# RESULTS

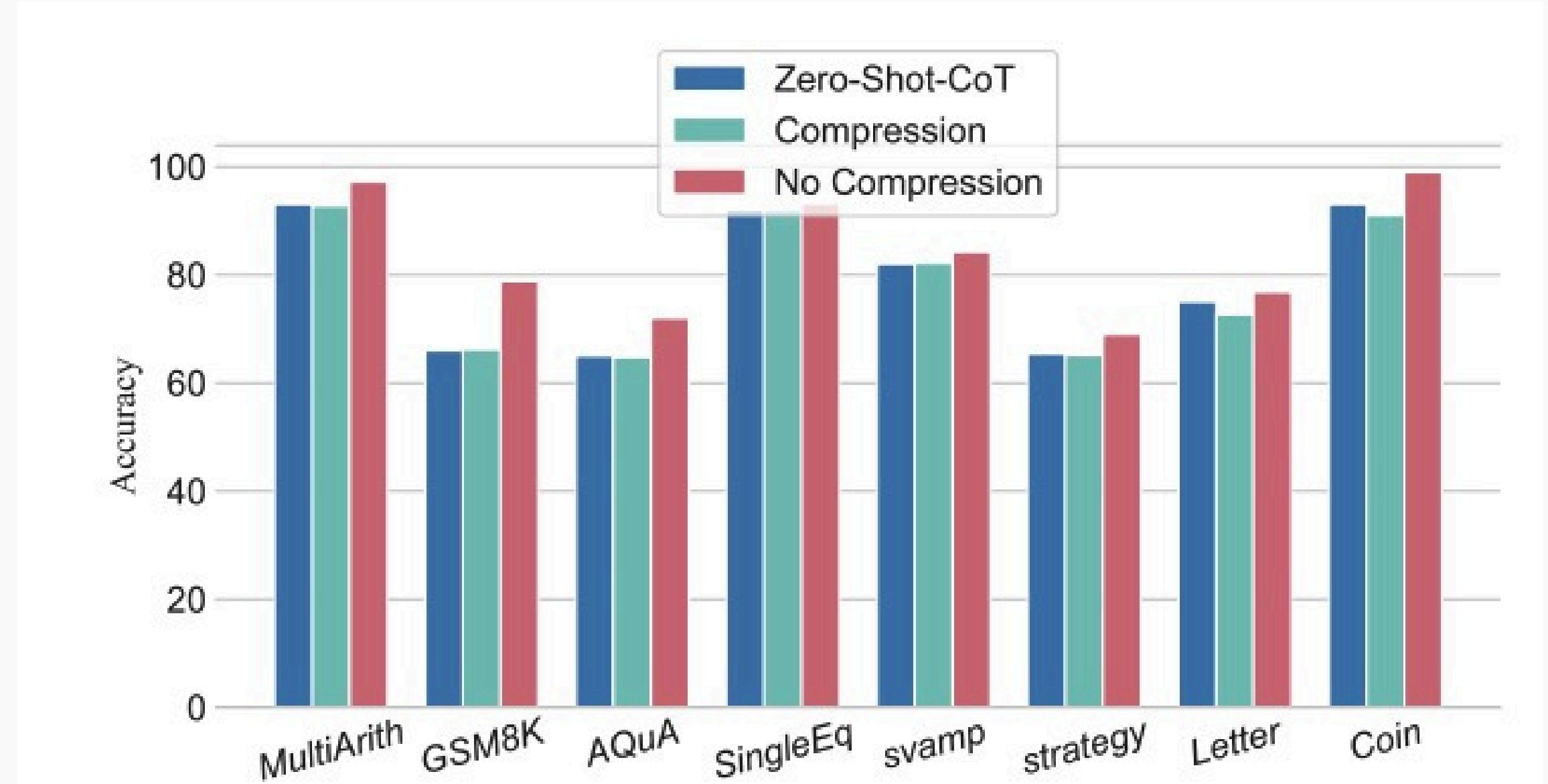
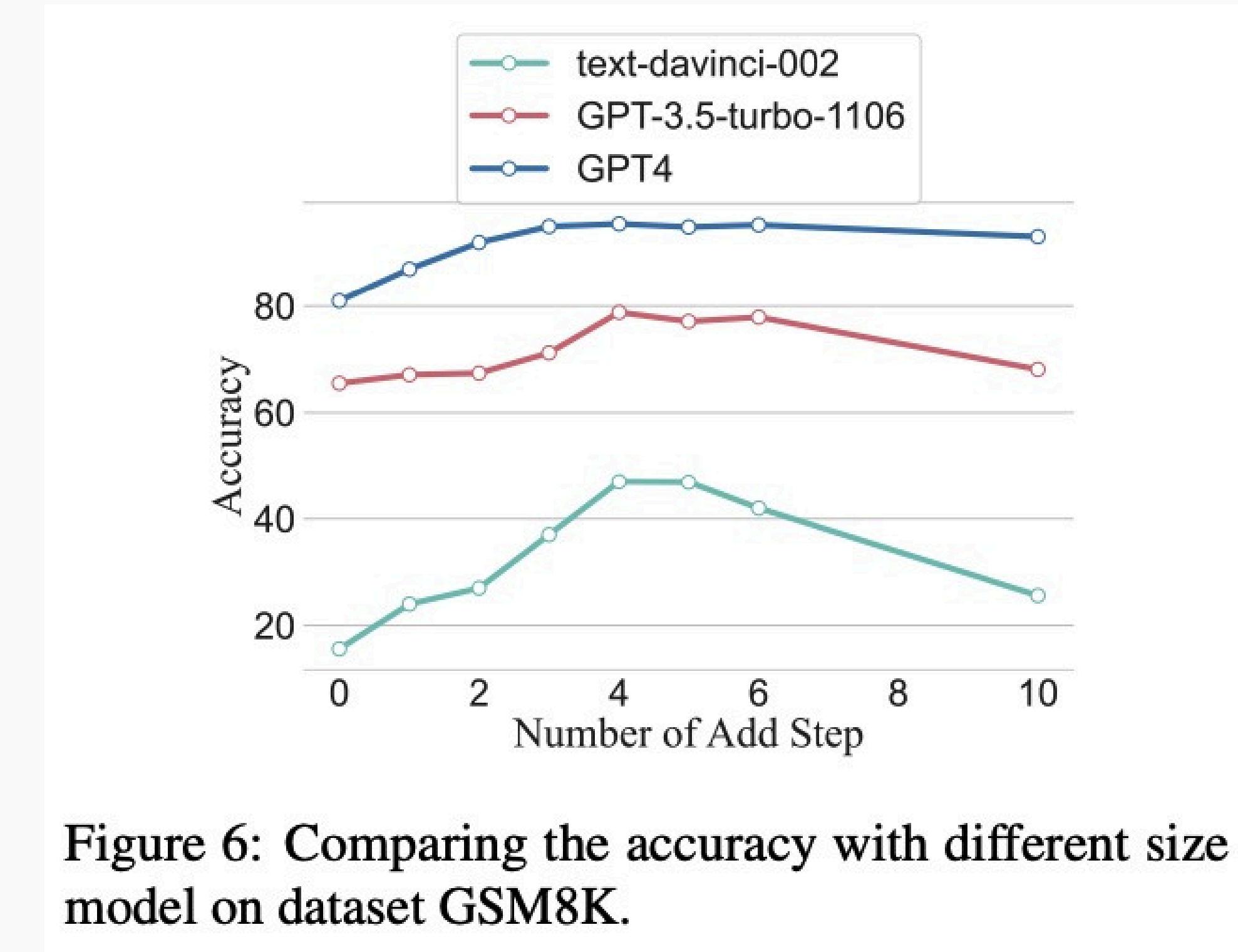


Figure 5: Compare the accuracy of the prompt with Compression and the prompt with No Compression

# RESULTS



# RESULTS

Table 5: Accuracy comparison of models on different datasets

Models	MultiArith	GSM8K
Zero-Shot	40±1.0	30.4±1.7
Zero-Shot-CoT	91.5±1.2	64.1±1.1
Manual-CoT	93.5±0.1	64.7±1.5
Auto-CoT	94±0.0	65.8±0.9
Changing Question (Manual-CoT)	92.9±0.1	62.1±1.7
Changing Question (Auto-CoT)	92.5±0.1	63.5±1.0
Add Reasoning Step (Manual-CoT)	97±0.0	70.1±0.3
Add Reasoning Step (Auto-CoT)	97.2±0.1	78.8±0.2
Add Reasoning Step and Changing Question (Manual-CoT)	96.6±0.1	69.6±0.2
Add Reasoning Step and Changing Question (Auto-CoT)	95.7±0.1	75.2±0.2

Table 4: Making deliberate alterations to sample questions

Original Prompt
<b>Q:</b> Wendy uploaded 45 pictures to Facebook. She put 27 pics into one album and put the rest into 9 different albums. How many pictures were in each album?
<b>Rationale:</b> “Let’s think step by step. First, Wendy uploaded 45 pictures in total. Second, Wendy put 27 pictures into one album. That means that Wendy put the remaining 18 pictures into 9 different albums. Each album would have 2 pictures.”
<b>Pred_ans:</b> “2”
<b>Gold_ans:</b> “2”
Making deliberate alterations
<b>Q:</b> Wendy uploaded 66 pictures to Facebook. She put 89 pics into one album and put the rest into 7 different albums. How many pictures were in each album?
<b>Rationale:</b> “Let’s think step by step. First, Wendy uploaded 54 pictures in total. Second, Wendy put 27 pictures into one album. That means that Wendy put the remaining 12 pictures into 6 different albums. Each album would have 7 pictures.”
<b>Pred_ans:</b> “2”
<b>Gold_ans:</b> “2”

# 02

# EXPERIMENTS

# EXPERIMENTAL SETTINGS

- Models : Mistral 7B, Qwen2 7B, Vicuna 13B
- Configuration
  - Quantization: 4-bit
  - Temperature: 0.0 (greedy decoding for reproducibility)
  - Max tokens: 512
  - Hardware: Google Colab A100 GPU
- Experiments : None CoT, Few Shot CoT, Zero Shot CoT

# BENCHMARKS

- **AQuA (Algebraic)**
  - 200 samples, Multiple-choice algebra questions
- **GSM8K (Math Reasoning)**
  - 200 samples, Elementary school math problems
- **MultiArith (Arithmetic)**
  - 180 samples, Basic arithmetic word problems
- **SVAMP (Math Variations)**
  - 200 samples, Math problem variations
- Each dataset was 30% sampled due to computational cost.
- Letter, Coin datasets weren't inferenced by the model.
- SingleEq, StrategyQA were not accessible

# ZERO SHOT PROMPT DESIGNS

**# Baseline: No reasoning**

**Q: [question]**

**A:**

**# Standard CoT: Basic step-by-step**

**Q: [question]**

**A: Let's think step by step.**

**# Extended CoT: Explicit instruction for more steps**

**Q: [question]**

**A: Let's think step by step, you must think more steps.**

# FEW SHOT PROMPT DESIGNS

## # Standard (2-3 reasoning steps)

Q: Janet has 10 apples. She gives 3 to her friend. How many apples does Janet have now?

A: Janet starts with 10 apples. She gives away 3 apples. So  $10 - 3 = 7$ . The answer is 7.

Q: [actual question]

A:

## # Extended (5-6 reasoning steps)

Q: Janet has 10 apples. She gives 3 to her friend. How many apples does Janet have now?

A: The question is: How many apples does Janet have now?

Janet starts with 10 apples.

...

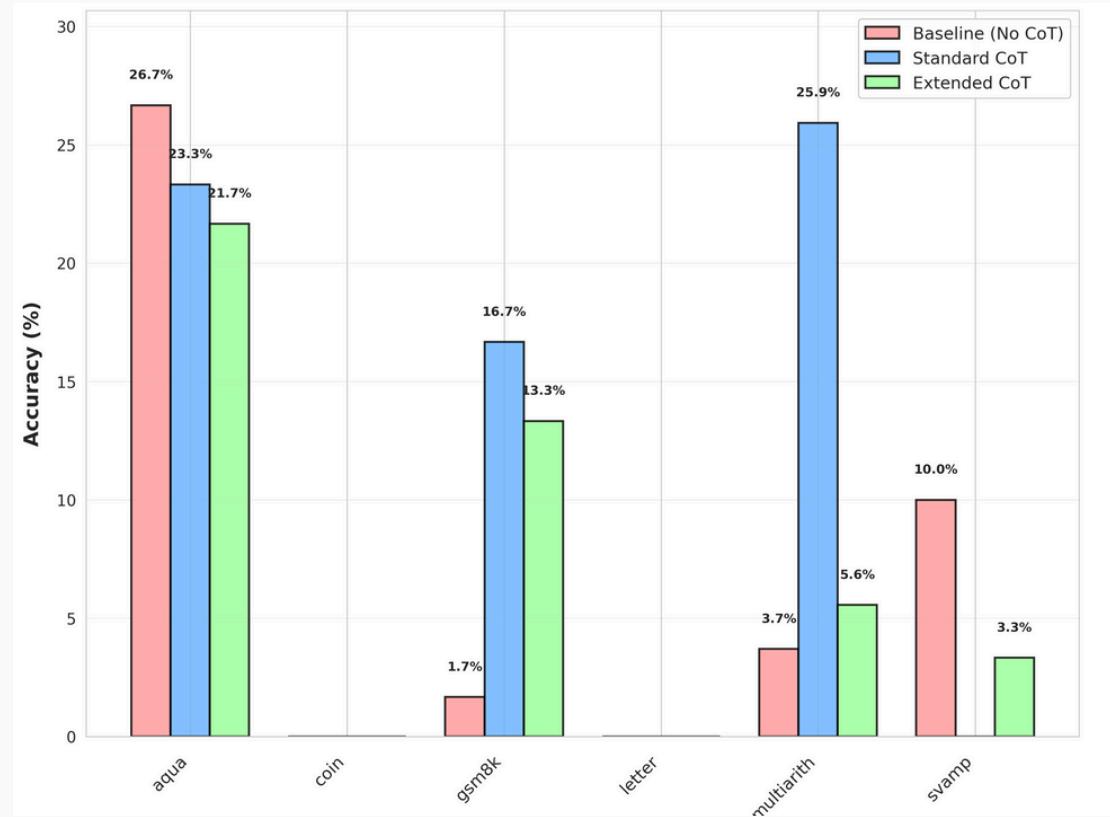
Self-verify:  $10 - 3 = 7$ . This is correct.

The answer is 7.

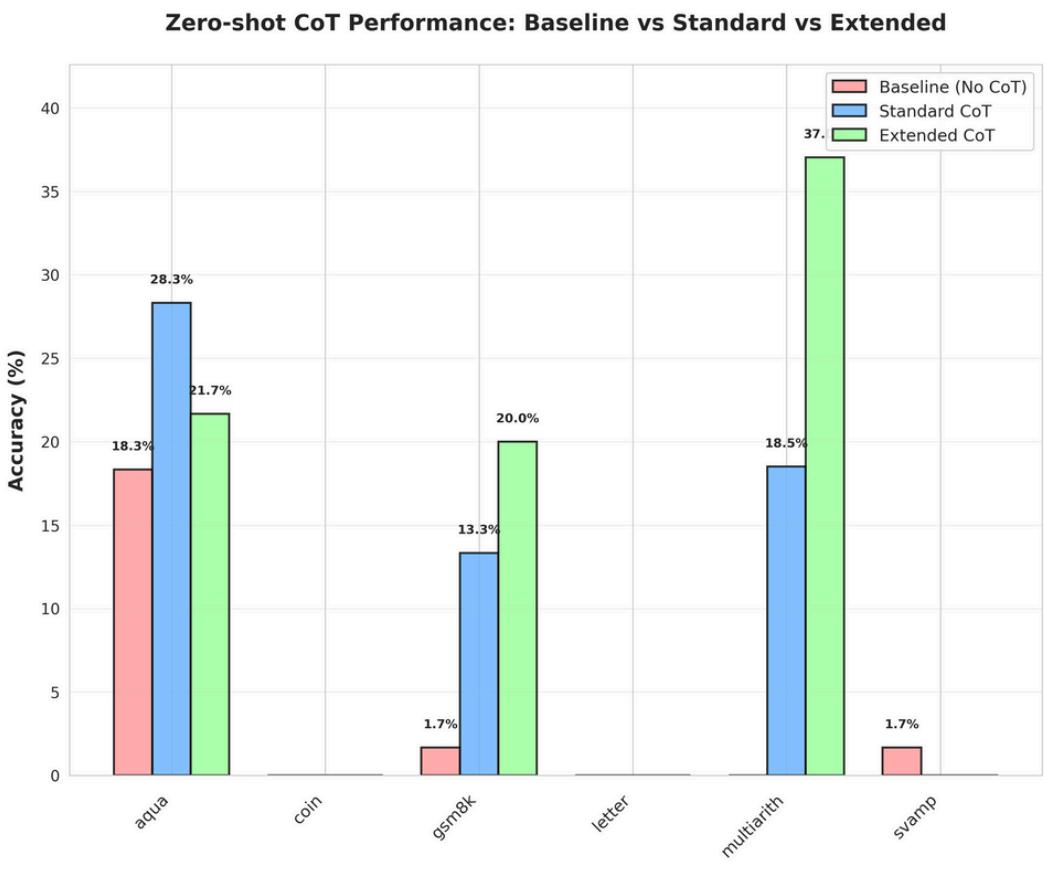
Q: [actual question]

A:

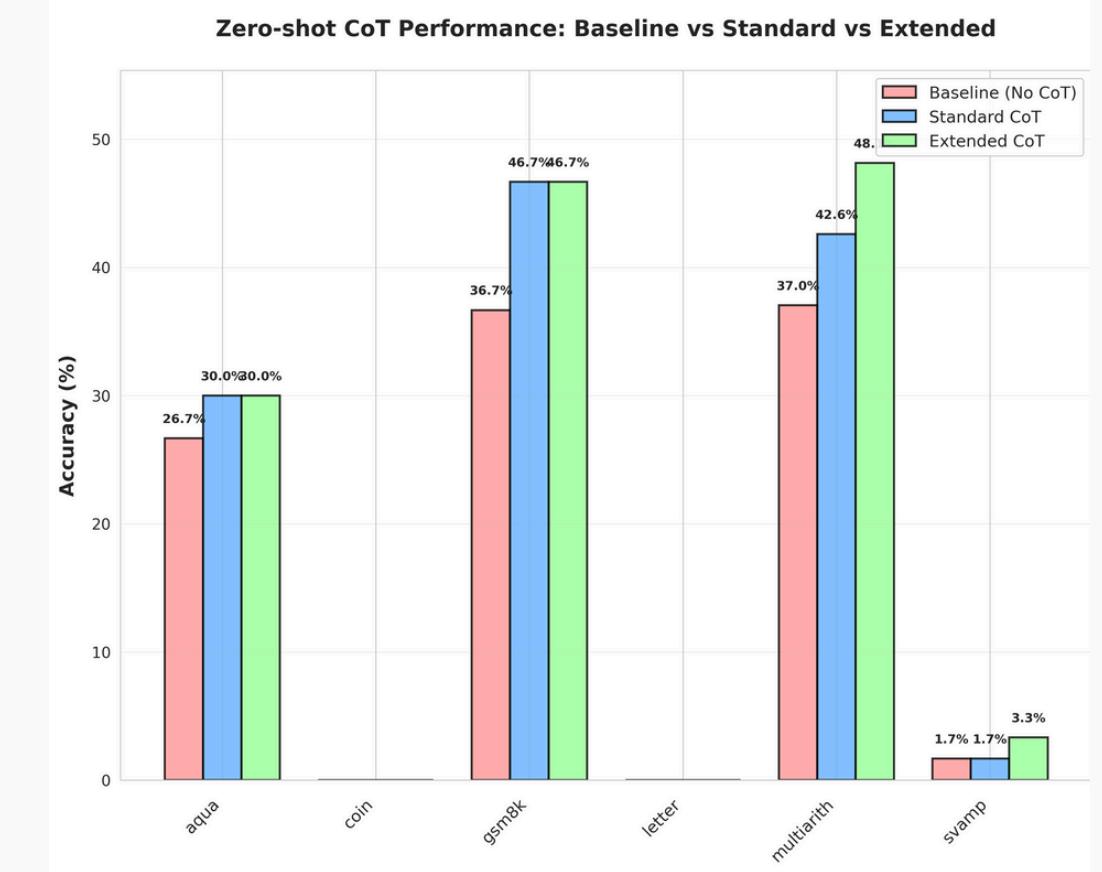
# RESULTS - ZERO SHOT



MISTRAL

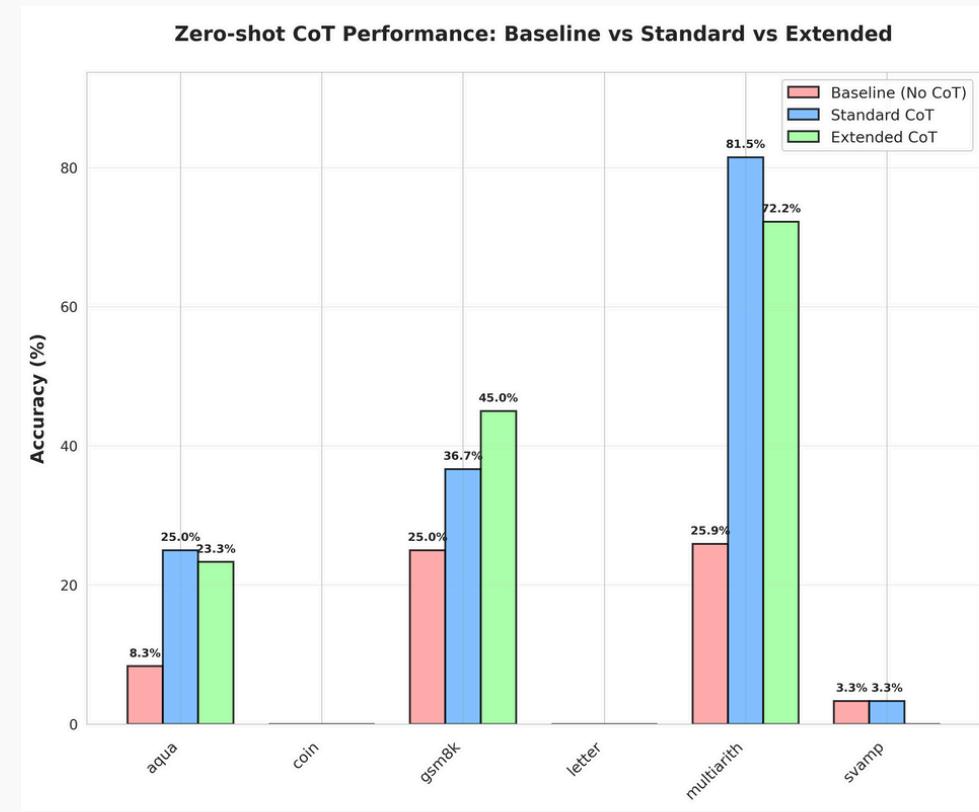


VICUNA

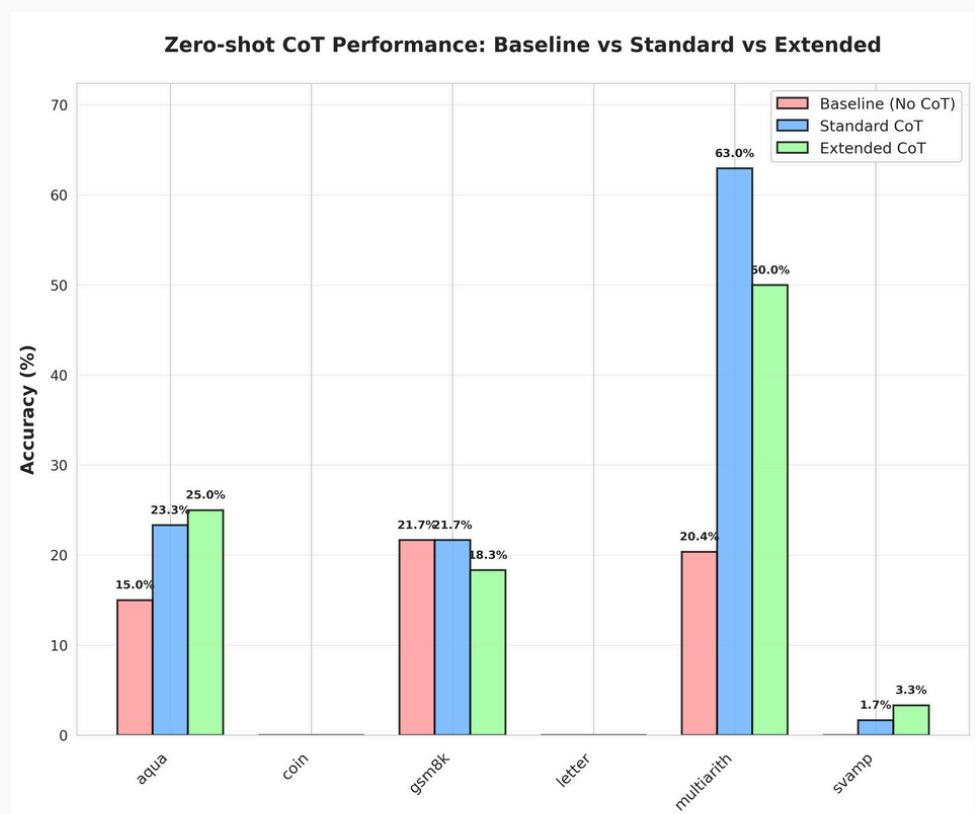


QWEN2

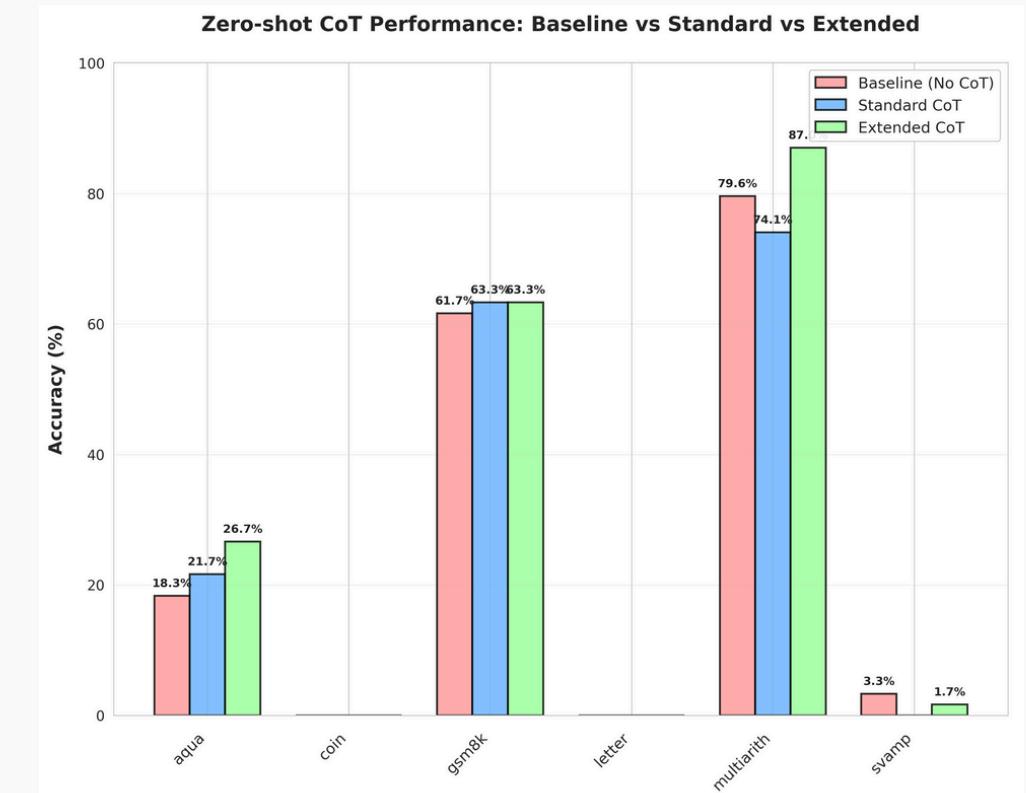
# RESULTS - FEW SHOT



MISTRAL



VICUNA



QWEN2

03

# RESULT ANALYSIS

# COMPARISON WITH ORIGINAL PAPER

- In ‘Limitation’, researchers mentioned that they only experimented with GPT 3.5 version. So we tried 3 (relatively small) open source models.
- **UNLIKE** the paper, longer reasoning steps **doesn’t always guarantee** better performance in small models.
- Sometimes **STANDARD** or **BASELINE** excels extended prompt

# COMPARISON WITH ORIGINAL PAPER

- **Extended CoT** works best for **deep multi-step reasoning**, but sometimes degrades performance in small models due to parameter scale.
- **Standard CoT** shows the **most stable and consistently strong** performance across tasks.
- **Baseline** prompting is only effective for **simple or shallow reasoning tasks**.

# LIMITATIONS

- We discovered that in small open source model, long CoT reasoning does not always good.
- However, our experiment still couldn't find out what is optimal CoT prompting in specific tasks or model type. Still, there are deviation between datasets and models.
- To find out robust CoT prompting steps would be next stage of this study.
  - e.g. N steps of reasoning shows robust performance in Video Reasoning / Mathematical tasks / ...

# Q&A

Our Works can be found at this [github.repository](#)