Arithmetic reasoning System Using GPT-3 with chain-of-Thought

Group Members: Dian Zhi | Ruobing Yan | Muwen You | Zheyuan Hu

NetID: dz247 | ry201 | my549 | zh216

Agenda

Introduction

Method

GPT-3 Generative Application

• Result and Interpretation



GPT-3 module (LLMs)

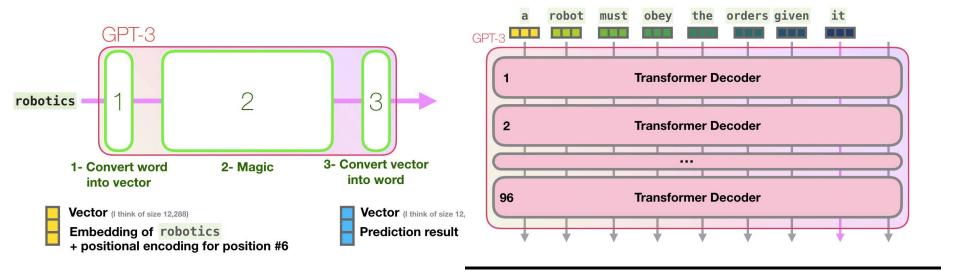


Generative Pre-trained Transformer 3 (GPT-3)

- A neural network machine learning model trained using internet data to generate any type of text.
- The 3rd- generation language prediction model in the GPT-n series created by OpenAI.

GPT-3 module (LLMs)

GPT-3 language model has 175 billion parameters, i.e., values that a neural network is optimizing during the training



Chain-of-Thought (CoT)

- Chain of thought prompting, this method enables models to decompose multi-step problems into intermediate steps.
- With chain of thought prompting, language models of sufficient scale (~100B parameters) can solve complex reasoning problems that are not solvable with standard prompting methods.

Input Q: Roger has 5 tennis balls. He buys 2 more cans of Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now? tennis balls does he have now? A: The answer is 11 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 Q: The cafeteria had 23 apples. If they used 20 to do they have? make lunch and bought 6 more, how many apples do they have? Model Output Model Output A: The cafeteria had 23 apples originally. They used A: The answer is 27. 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. <

Chain-of-Thought (CoT)

Why CoT Works?

Allows models to decompose multi-step problems into intermediate steps.

Why we use CoT?

 Chain-of-thought reasoning can be readily elicited in sufficiently large off-the-shelf language models.

Further explore the ability of CoT, by introducing Auto-CoT.

Input

Chain-of-Thought

Output

Data Source (GSM8K - Grade School Math)

- A dataset of 8.5K high quality linguistically varied grade school math word problems
- Split into 7.5K training problems and 1K test problems.
- Between 2 and 8 steps are required to complete these problems, the solution is reached by conducting a series of simple calculations utilizing the (+ - / *) arithmetic operations.

```
{"question": "James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?", "answer": "He sprints 3*3=<<3*3=9>>9 times\nSo he runs 9*60=<<9*60=540>>540 meters\n#### 540"}
```



Prompting Method

Zero-Shot

Zero-Shot-CoT (with keywords)

Few-Shot: Auto-CoT

Few-Shot: Manual-CoT

Question Cluster

Most Representative

Nearest

Zero-Shot-CoT

- Zero-Shot: Directly give the answer of the question.
- Zero-Shot-CoT: LLMs have shown to be decent zero-shot reasoners whose generated rationales have already reflected the CoT reasoning.

Prompting Method: 0-shot

Prompt: Q: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?

A:

Completion: James runs a total of 540 meters a week.

Prompting Method: 0-shot with keywords

Prompt: Q: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?

A: Let's think step by step.

Completion: James runs three sprints three times a week. Each sprint is 60 meters. So, for each sprint he runs 60 meters, and for three sprints he runs 180 meters. And since he runs three times a week, he runs 180 meters three times. So, the total number of meters he runs each week is 540 meters.

Few-Shot: Manual-CoT

- Few-shot prompting with manual reasoning demonstrations one by one.
- Each demonstration has a question and a reasoning chain, which is composed of a rationale (a series of intermediate reasoning steps) and an expected answer.
- Performance depends on effective hand-drafted demonstration programs.

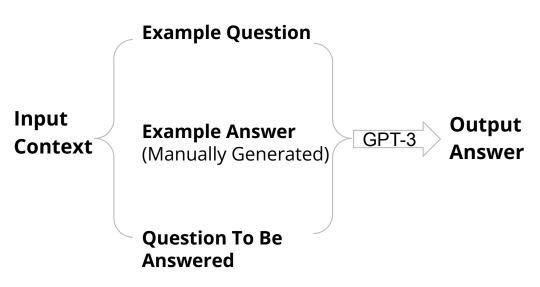
Manual Demos One by One

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today? Question A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.▼ Answer Rationale Q: A pet store had 64 puppies. In one day they sold 28 of them and put the rest into cages with 4 in each cage. How many cages did they use? **Test Question** LLM

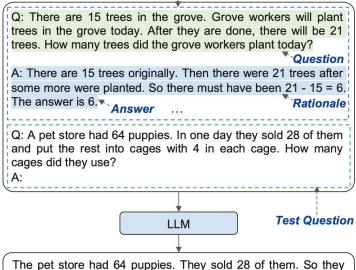
The pet store had 64 puppies. They sold 28 of them. So they had 64 - 28 = 36 puppies left. They put them into cages with 4 in each cage. So they used 36 / 4 = 9 cages. The answer is 9.

(b) Manual-CoT

Few-Shot: Manual-CoT



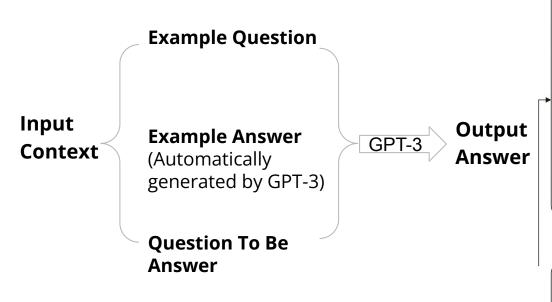
Manual Demos One by One



(b) Manual-CoT

had 64 - 28 = 36 puppies left. They put them into cages with 4 in each cage. So they used 36 / 4 = 9 cages. The answer is 9.

Few-Shot: Auto-CoT



Auto Demos One by One

Q: While shopping for music online, Zoe bought 3 country albums and 5 pop albums. Each album came with a lyric sheet and had 3 songs. How many songs did Zoe buy total?

A: Let's think step by step. Zoe bought 3 country albums. Each album has 3 songs. So she bought 3*3=9 songs from the country albums. Zoe bought 5 pop albums. Each album has 3 songs. So she bought 5*3=15 songs from the pop albums. Zoe bought 9+15=24 songs in total. The answer is 24.

Q: A chef needs to cook 9 potatoes. He has already cooked 7. If each potato takes 3 minutes to cook, how long will it take him to cook the rest?

A: Let's think step by step. The chef has already cooked 7 potatoes. That means it has taken him 7 * 3 minutes to cook those 7 potatoes. That means it will take him 3 more minutes to cook each of the remaining 2 potatoes ...

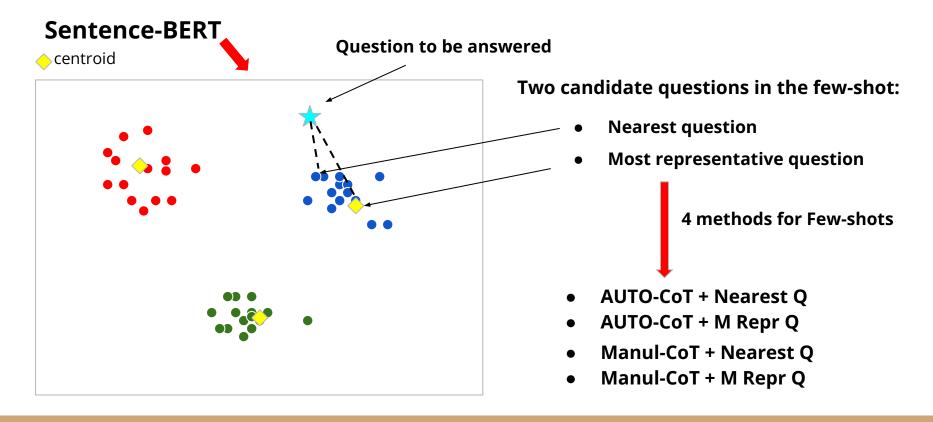
Q: A pet store had 64 puppies. In one day they sold 28 of them and put the rest into cages with 4 in each cage. How many cages did they use?

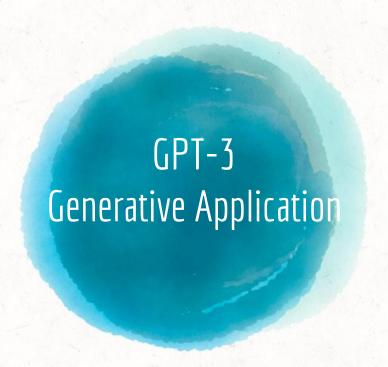
A: Let's think step by step.

Test Question LLM In-Context Reasoning

The pet store had 64 puppies. They sold 28 of them. That means they have 36 puppies left. They put the rest into cages with 4 in each cage. That means they have 9 cages. The answer is 9.

How to decide question in auto-cot & manual-cot?





Math Reasoning Bot

https://codyhuu-auto-cot-qa-app-auto-cot-app-on5fx0.streamlit.app/



Reminder: *GPT-3 is a LANGUAGE MODEL!

*If we want to train a model to get optimal performance in arithmetic reasoning, we would not use GPT-3!

Goal: Explore the limits of GPT-3 (**A LANGUAGE MODEL**) in mathematical reasoning by using different kinds of <u>prompting method</u>

Prompting Method

0-shot

0-shot with keywords

No Info at all

Auto-COT representative question

Auto-COT nearest question

No ground truth answer and COT

Manual-COT representative question

Manual-COT nearest question

Ground truth answer and COT

Qualitative Measure

Correct: The numerical answers from the completion sentences are the same as ground truths'

Accuracy: # correct / # total

Expectations: 1) Auto-COT with nearest questions have higher accuracy compare to Auto-COT with most representative questions

Prompting Method	0-shot	0-shot with keywords	Auto-COT representative question	Auto-COT nearest question	Manual-COT representative question	Manual-COT nearest question
Accuracy	13%	44%	51%	31%	36%	38%

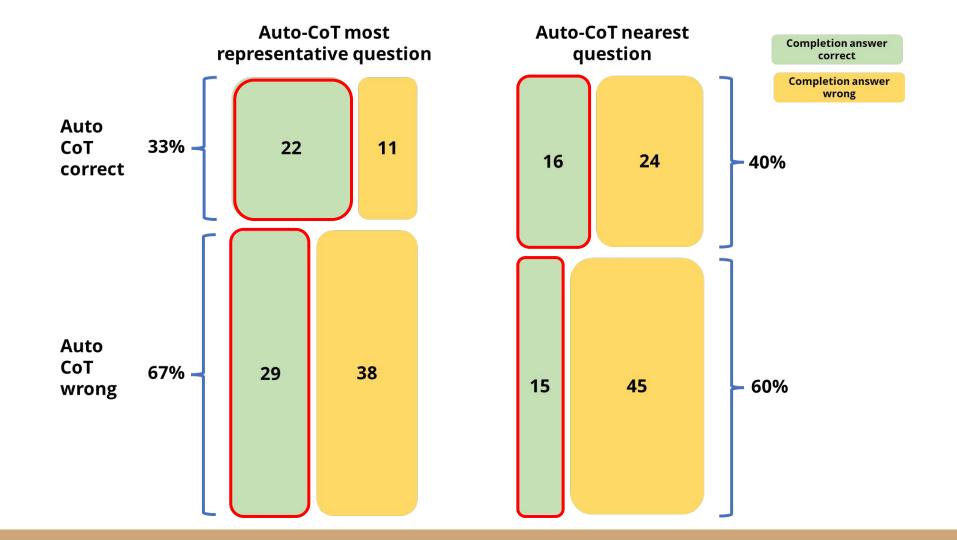
Qualitative Measure

Correct: The numerical answers from the completion sentences are the same as ground truths'

Accuracy: # correct / # total

Expectations: 2) Manual-COT have higher accuracy compare to Auto-COT

Prompting Method	0-shot	0-shot with keywords	Auto-COT representative question	Auto-COT nearest question	Manual-COT representative question	Manual-COT nearest question
Accuracy	13%	44%	51%	31%	36%	38%



Most representative vs Nearest

Extreme Case: The nearest question is the question itself!

Q: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?

1. Generate prompt using 0-shot with keywords:

Q: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?

A: Let's think step by step. For each sprint, James runs 60 meters. He runs 3 sprints three times a week. So, he will run $3 \times 3 = 9$ sprints per week. Since each sprint is 60 meters, he will run a total of $9 \times 60 = 540$ meters per week.

2. Auto CoT nearest question prompt:

Q: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?

A: Let's think step by step. For each sprint, James runs 60 meters. He runs 3 sprints three times a week. So, he will run $3 \times 3 = 9$ sprints per week. Since each sprint is 60 meters, he will run a total of $9 \times 60 = 540$ meters per week.

Q: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?

A: Let's think step by step.

*The completion is most likely to be the same as what 0-shot with keywords generated!

Most representative vs Nearest (Hypothesis 1)

Same (the extreme case of nearest): No additional Information compared to 0-shot with keywords!

Similar example question (nearest) -> few additional information

Representative question: In the same cluster as the new question but not too similar to the new question -> more additional information -> better accuracy

Auto-CoT vs Manual-CoT (Hypothesis 2)

Reminder: *GPT-3 is a LANGUAGE MODEL!

The first priority of GPT-3 is to complete the sentence in a way that is consistent with the previous contexts.

In Auto-CoT, the CoT was generated by GPT-3 itself (extracted from learnt distribution)

E.g. For each sprint, James runs 60 meters. He runs 3 sprints three times a week. So, he will run $3 \times 3 = 9$ sprints per week. Since each sprint is 60 meters, he will run a total of $9 \times 60 = 540$ meters per week.

In Manual-CoT, the CoT was generated by human annotators

E.g. He sprints 3*3=<<3*3=9>>9 times. So he runs 9*60=<<9*60=540>>540 meters. #### 540

Manual CoT format too different from Auto CoT -> model focuses more on the format rather than the logic -> lower accuracy

Conclusion

- Hypothesis 1: Similar example question -> few additional information -> lower accuracy
- Hypothesis 2: Manual CoT format is too different from Auto CoT's ->
 model focuses more on the format rather than the logic -> lower accuracy
- The BEST prompting method (highest accuracy) is Auto CoT with most representative question

Prompting Method	0-shot	0-shot with keywords	Auto-COT representative question	Auto-COT nearest question	Manual-COT representative question	Manual-COT nearest question
Accuracy	13%	44%	51%	31%	36%	38%

Okay I got it... But...

Why does manual-COT with nearest question have higher accuracy than auto-COT with nearest question? Did you just say:

 Manual CoT format is too different from Auto CoT's -> model focuses more on the format rather than the logic -> lower accuracy

Yes... But...

- In Auto-CoT nearest question cases: Similar example question -> few additional information
- In manual-CoT nearest question cases, not only the question is provided to the model but the solution as well! More information provided to the model -> higher accuracy

Prompting Method	0-shot	0-shot with keywords	Auto-COT representative question	Auto-COT nearest question	Manual-COT representative question	Manual-COT nearest question
Accuracy	13%	44%	51%	31%	36%	38%

Thanks For Watching

Work Cited

- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.
- Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). Automatic Chain of Thought Prompting in Large Language Models. arXiv preprint arXiv:2210.03493.
- Cobbe, K., Kosaraju, V., Bavarian, M., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.