LLM Prompt Strategies for Commonsense-Reasoning Tasks

Tesa Robič, Tim Dolenc, Matjaž Bevc

May 14, 2024

1 Introduction

In this project we will research, compare and evaluate different prompt strategies for commonsense reasoning tasks used in Large Language Models (LLM). LLMs are really good at understanding and generating text. But when it comes to understanding common sense (the things we know about the world without being told), they need help. So, researchers are trying out different ways to give these models hints or instructions, called prompts. There are many different prompt strategies like Chain Of Thought (CoT), In-Context Learning (ICL), plan-and-solve techniques, Tree Of Thought, Retrieval augmentation (RAG) and more. In the next chapter, we introduce different strategies that we consider using for this project [1].

2 Existing solutions and Realated Work

2.1 Chain-of-Thought (CoT) prompting

Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. CoT prompting involves providing intermediate reasoning steps to guide the model's responses, which can be facilitated through simple prompts such as "Let's think step by step" or through a series of manual demonstrations, each composed of a question and a reasoning chain that leads to an answer [1].

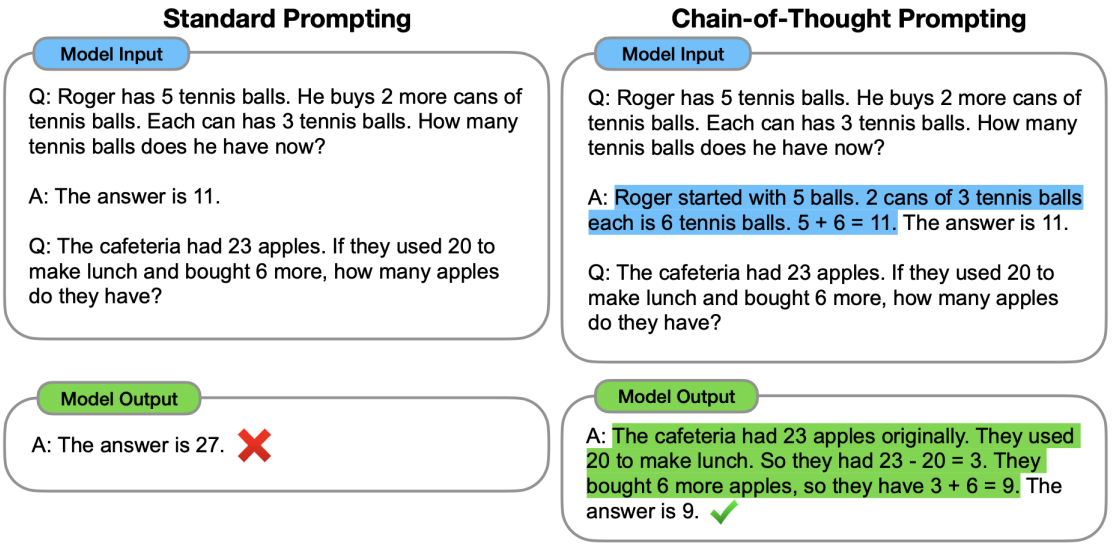

Figure 1: An illustration of standard LLM prompting on the left, and chain-of-thought prompting on the right.

Some benefits of CoT promting:

- **Improved accuracy**: With clear reasoning steps, the LLM is less likely to make mistakes or jump to illogical conclusions. This is especially helpful for tasks with multi-step logic or complex reasoning requirements.
- **Transparency:** CoT prompts make the reasoning process more transparent, allowing us to understand how the LLM arrived at its answer. This is crucial for building trust and identifying potential bias or errors.
- **Better performance on complex tasks:** CoT is particularly effective for tasks that require multi-step reasoning, logical deduction or *common-sense* application. These are areas where past LLMs often struggled.
- **Adaptability**: The technique can be used for various tasks, from solving math problems and interpreting data to summarizing text and even creative writing.
- **Precision-Guided Reasoning:** By providing a clear path, CoT reduces the risk of LLMs stumbling into erroneous conclusions or leaps of illogical faith. Multi-step tasks and convoluted reasoning problems, once impenetrable to LLMs, become navigable landscapes with CoT at the helm.

Limitations of CoT prompting:

- **Manual effort:** Creating effective CoT prompts requires understanding the problem and designing the reasoning steps yourself. This can be time-consuming and complex for intricate tasks.
- **Model size:** CoT seems to be more effective for larger LLMs with stronger reasoning capabilities. Smaller models might struggle to follow the prompts or generate their own reasoning chains.
- **Prompt bias:** Like any other prompting technique, CoT can be susceptible to biased prompts that lead the LLM to incorrect conclusions. Careful design and testing are crucial.
- **Bias Blind Spots:** Just like any prompt, CoT is susceptible to biased information. Careful design and thorough testing are crucial to ensure the LLM doesn't follow a misleading path
- **Crafting the Path:** Building effective CoT prompts requires understanding the problem's intricacies and formulating a clear, logical chain of reasoning. This can be demanding, especially for complex tasks. [6]

2.1.1 Zero Shot CoT

Zero Shot Chain of Thought (Zero shot CoT) prompting is a follow up to CoT prompting, which introduces an incredibly simple zero shot prompt. After each question we append the words "Let's think step by step." or a similar text, to extract step-by-step reasoning. After that LLMs are able to generate a chain of thought that answers the question.[8]

2.2 Plan-and-Solve (PS) prompting

With Zero-shot CoT, there are three pitfalls: calculation errors, missing-reasoning-step errors and semantic understanding errors.

The first two pitfalls can be addressed with Plan-and-Solve prompting (PS and PS+).

Plan-and-Solve prompting consists of two components: devising a plan to divide the task into small subtasks and carrying out the subtasks according to the plan. While Zero-Shot CoT appends the phrase *"Let's think step by step"* to the prompt, PS appends *"Let's first understand theproblem and devise a plan to solve the problem. Then, let's carry out the plan and solve the problem step by step."*

Extending the prompt with additional phrases gets us to PS+. For example, in case of complex arithmetic calculations, these phrases can be added:

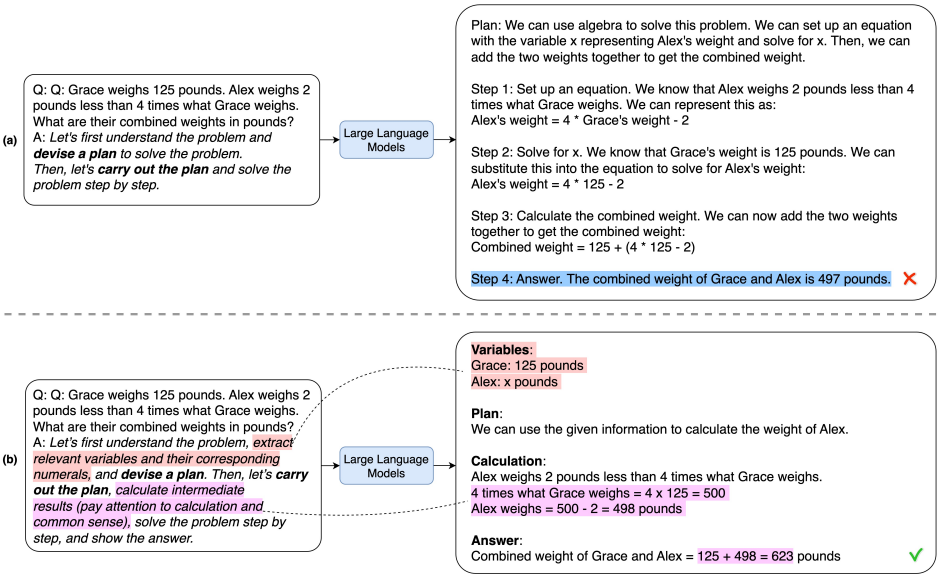

Figure 1: Comparison of PS prompting with PS+ prompting

- *"pay attention to calculation"*
- *"extract relevant variables and their corresponding numerals"*
- *"calculate intermediate results"*

This requests precise calculations from the LLM, tells it to not leave out relevant variables and enhances its ability to generate reasoning steps. [10]

2.2.1 Limitations

There are two main limitations of PS and PS+:

- It takes effort to design the prompt to generate correct reasoning steps.
- PS(+) can address calculation errors and missing reasoning-step errors, but it cannot help against semantic misunderstanding errors. [10]

2.3 Few-Shot Prompting

In Few-Shot Prompting, the model is provided with a small set of examples within the prompt itself. These examples are selected to demonstrate the task or the type of response desired. By analyzing these examples, the LLM learns the context and structure of the task, allowing it to apply this understanding to new, similar problems. Few-Shot Prompting is akin to In-Context Learning (ICL), but instead of providing analogies, it furnishes the prompt with concrete examples followed by the LLM's desired answers.

For instance, consider the following examples:

- Example 1: "Emma left her keys in the freezer. Why? - Emma was carrying a lot of things and absent-mindedly put her keys in the freezer."
- Example 2: "Mike found his phone in the bathroom cabinet. Why? - Mike was multi-tasking, cleaning the bathroom while talking on the phone, and he left it there accidentally."

Now, when the LLM is presented with the target question about John's coffee cup, it has context from these examples on how to approach and solve the problem:

Target Question with Few-Shot Learning Applied:

"John found his coffee cup in the refrigerator. Why was it there?"

Based on the learning from the examples, the LLM might reason in the following manner:

- Similar to Emma, who absent-mindedly left her keys in the freezer while being preoccupied with carrying multiple items, it is plausible that John was also engaged in multitasking.
- Drawing a parallel with Mike's situation, where his attention was divided, leading to him leaving his phone in the bathroom, John might have been similarly distracted.
- Consequently, it is reasonable to infer that John could have placed his coffee cup in the refrigerator during a moment of absent-mindedness or distraction, akin to the scenarios described in the examples.

2.3.1 Auto Few shot

We implemented an automatic few-shot learning approach to enhance the main question-answering process. This method identifies the three most similar questions from the dataset (based on cosine similarity) and incorporates these examples, along with their answers, to provide context for the main question. Here's how it works:

> **Main question** Given the question 'What do people aim to do at work?' and the following choices: A: complete job, B: learn from each other, C: kill animals, D: wear hats, E: talk to each other, which one is correct? Answer only with one of the following A, B, C, D, or E. [**End of main question**]

is expanded with examples in the following way:

> **Introduction** You will see examples and a main question. Please provide the answer to the main question based on these examples. Your response can only include one character: A, B, C, D, or E. [**End of introduction**]
>
> **Example question** Given the question 'A person would join a trade school for finding information related to what?' and the following choices: A: ulcers, B: degree, C: understanding of, D: gaining knowledge, E: happiness, which one is correct? Answer only with one of the following A, B, C, D, or E. [**End of example question**]
> **Answer** D [**End of answer**]
>
> **Example question** Given the question 'Where is known to be a wealth of information?' and the following choices: A: park, B: internet, C: meeting, D: library, E: book, which one is correct? Answer only with one of the following A, B, C, D, or E. [**End of example question**]
> **Answer** D [**End of answer**]
>
> **Example question** Given the question 'What is another name for the color of the fur of a dog with light-colored fur?' and the following choices: A: fair, B: basket, C: dog hair, D: game, E: sun, which one is correct? Answer only with one of the following A, B, C, D, or E. [**End of example question**]
> **Answer** A [**End of answer**]
>
> **Main question** Given the question 'What do people aim to do at work?' and the following choices: A: complete job, B: learn from each other, C: kill animals, D: wear hats, E: talk to each other, which one is correct? Answer only with one of the following A, B, C, D, or E. [**End of main question**]

3 Datasets

3.1 TBA)

3.2 CommonsenseQA

The CommonsenseQA is a dataset for commonsense question answering task. The dataset consists of 12,247 questions with 5 choices each. The dataset was generated by Amazon Mechanical Turk workers in the following process (an example is provided in parentheses):

1. a crowd worker observes a source concept from ConceptNet ("River") and three target concepts ("Waterfall", "Bridge", "Valley") that are all related by the same ConceptNet relation ("AtLocation"),
2. the worker authors three questions, one per target concept, such that only that particular target concept is the answer, while the other two distractor concepts are not, ("Where on a river can you hold a cup upright to catch water on a sunny day?", "Where can I stand on a river to see water falling without getting wet?", "I'm crossing the river, my feet are wet but my body is dry, where am I?")
3. for each question, another worker chooses one additional distractor from Concept Net ("pebble", "stream", "bank"), and the author another distractor ("mountain", "bottom", "island") manually. [7]

4 Initial ideas

Our idea is to test different prompting strategies, described in chapter 2 on two different datasets. One is classification dataset (CommonSenseQA) and the other is X. After gathering all the outputs from different prompting strategies in combination with different datasets we evaluate the model response for each tehnique.

5 Methods

For this research we choose Mistral-7B-Instruct-v0.2 Large Language Model (LLM), which is an instruct fine-tuned version of the Mistral-7B-v0.2. All prevoiusly described strategis were applied in to this model with same question format.

5.1 Mistral 7B

Mistral 7B is 7-billion-parameter language model engineered for superior performance and efficency. Studies show that it outperforms the best open 13B model Llama 2 across all evaluated benchmarks, and the best released 34B model Llama 1 in reasoning, mathematics, and code generation [11].

5.2 Prompt formatting

The instruct version of Mistral 7B model accepts prompts surounded by `[INST]` and `[/INST]` tokens. The very first instruction should begin with a begin of sentence id. The next instructions should not. The assistant generation will be ended by the end-of-sentence token id.

E.g.

```
text = "<s>[INST] What is your favourite condiment? [/INST]"
"Well, I'm quite partial to a good squeeze of fresh lemon juice. It adds
just the right amount of zesty flavour to whatever I'm cooking up in the
kitchen!</s> "
"[INST] Do you have mayonnaise recipes? [/INST]"
```

All outputs for CommonsenseQA dataset were gathered from validation split containing 1221 examples. We used validation set because it is a benchmark dataset and test split correct answers are not publicly available. Output structure contains Question ID, Question, Choices, Correct Answer Key and models output.

Details about seconda dataset will be provided shortly.

6 References

[1] Banghao C., Zhaofeng Z., Nicolas L., Shengxin Z.: Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review; Dostopano na: https://arxiv.org/pdf/2310.14735.pdf [Dostopano Marec 2024]

[2] Zhang Z., Zhang A. , Mu Li , Smola A.: AUTOMATIC CHAIN OF THOUGHT PROMPTING IN LARGE LANGUAGE MODELS; Dostopano na: https://arxiv.org/pdf/2210.03493.pdf [Dostopano Marec 2024]

[3] Zhang Z., Zhang A. , Mu Li , Smola A.: Auto-CoT: Automatic Chain of Thought Prompting in Large Language Models (ICLR 2023); Dostopano na: https://github.com/amazon-science/auto-cot [Dostopano Marec 2024]

[4] Qingxiu D., Lei Li , Damai D. , Ce Z. , Zhiyong W. , Baobao C. , Xu S. , Jingjing X. , Lei L., Zhifang S.: A Survey on In-context Learning; Dostopano na: https://arxiv.org/pdf/2301.00234.pdf [Dostopano Marec 2024]

[5] Medium: Understanding In-context Learning in Large Language Models (like GPT3/GPT-J/GPTNeoX); Dostopano na: https://medium.com/@mlblogging.k/understanding-in-context-learning-in-large [Dostopano Marec 2024]

[6] Pandey p., Medium: Chain of Thought Prompting: Guiding LLMs Step-by-Step; Dostopano na: https://medium.com/@pankaj_pandey/chain-of-thought-prompting-guiding-llms-step-by-step-e6eac32d02d8 [Dostopano Marec 2024]

[7] Talmor et al: CommonsenseQA; Dostopano na: https://paperswithcode.com/dataset/commonsenseqa [Dostopano Marec 2024]

[8] Kojima et al: Large Language Models are Zero-Shot Reasoners; Dostopano na: https://arxiv.org/pdf/2205.11916 [Dostopano Maj 2024]

[10] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, Ee-Peng Lim: Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models; Dostopano na: https://doi.org/10.48550/arXiv.2305.04091 [Dostopano Marec 2024]

[11] Jiang et al: Mistral 7B; Dostopano na: https://arxiv.org/pdf/2310.06825 [Dostopano Maj 2024]