FRI Natural language processing course 2024
Project report

LLM Prompt Strategies for Commonsense-Reasoning Tasks

Žan Počkar, Amer Mujagić, and Ivan Nikolov

Abstract

In order to systematically evaluate and compare the efficiency of different prompt strategies, we developed a flexible testing framework capable of accommodating multiple prompting techniques and LLMs. We utilized the DSPy framework to reshape prompts from a selected dataset to align with the Chain-of-Thought strategy. We used Llama3 to generate predictions based on Chain-of-Thought prompts and zero-shot prompts on 100 examples from the Physical IQa dataset.

Keywords
Prompt Strategies, Prompt Evaluation

Advisors: Aleš Žagar

Introduction

Large Language Models (LLMs) have shown remarkable progress in various tasks when supplied with extensive training material. However, improving their commonsense reasoning capabilities remains a considerable challenge. Commonsense reasoning, which involves making judgments based on everyday knowledge and experiences, is a vital aspect of human intelligence. Incorporating this ability into LLMs is a complex task.

In this paper, we compare two prompting strategies Chain of Thought and Zero-Shot. Our evaluation framework, designed to accommodate multiple models and strategies in the future, uses a subset of the Physical IQa benchmark for testing. The primary objective of this paper is to serve as a proof of concept for our evaluation framework and to provide a blueprint for comparing the effectiveness of different prompt strategies on various LLMs.

Problem

Commonsense reasoning in Large Language Models (LLMs) is a complex problem due to the implicit and context-dependent nature of commonsense knowledge. For instance, consider the sentence “The ice cream was too hot to eat.” Humans intuitively understand this sentence is likely incorrect because ice cream is typically cold. However, an LLM might not flag this as an anomaly unless it has been explicitly trained on similar examples or has learned to associate certain objects with their typical properties.

Another challenge is the representation and utilization of commonsense knowledge in LLMs. LLMs learn from the patterns in the data they are trained on, so the size and quality of the training dataset significantly influence their performance. However, encoding commonsense knowledge into a machine-readable format that an LLM can learn from is not straightforward. For example, understanding that “people usually sleep at night” requires the model to learn not just the concepts of “people,” “sleep,” and “night,” but also the typical association between these concepts.

Lastly, the evaluation of commonsense reasoning in LLMs is a significant challenge. Traditional evaluation metrics may not adequately capture the nuances of commonsense reasoning. For instance, if an LLM generates the sentence “The man opened his umbrella because it was raining,” it demonstrates an understanding of the causal relationship between rain and using an umbrella. However, quantifying this understanding and comparing it across different models or tasks is not trivial, necessitating the development of more sophisticated evaluation metrics and benchmarks.

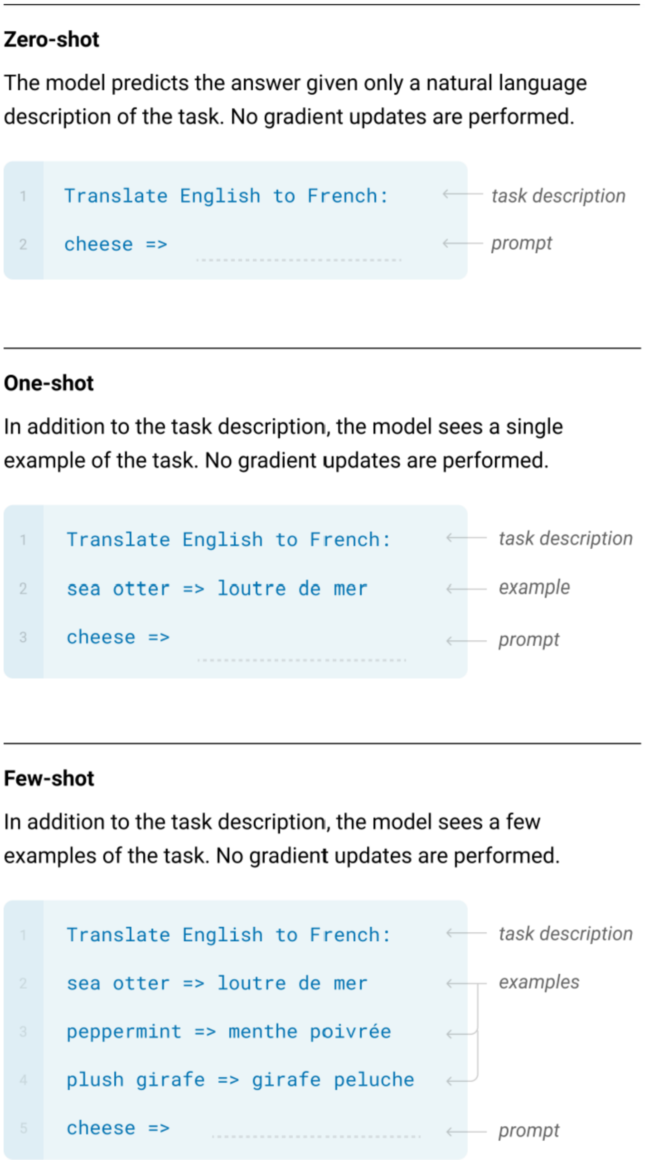
Our approach to this implementation starts with taking the prompts from a given dataset and using an LLM to reshape them into a given strategy. Unfortunately not many datasets are available that compare different strategies and most are used to compare different LLMs. After reshaping the prompts we run them through the model and compare the results we get to the correct results from the dataset.

Prompting Techniques

Prompts techniques are the proposed input structures that guide LLMs to produce outputs with the desired form and content quality. That is, to simplify, we can say that we use different prompting techniques to ”program” LLM inputs by following specific patterns in order to obtain an output. [1]

Zero shot prompting

Zero-shot prompting [2] is a prompting technique in which the prompt is made up only of natural language instruction describing the task. Meaning that no additional demonstrations are allowed. An example of zero-shot, one-shot and few-shot learning, with each ”shot” being an additional demonstration, can be seen in Figure 1



Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

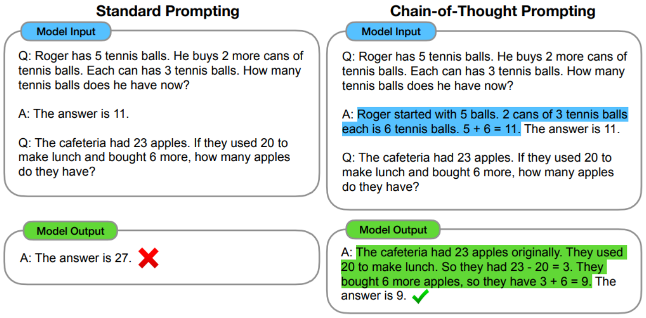
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

Figure 1. Examples of zero-shot, one-shot and few-shot prompts

Chain of thought

Chain of Thought (CoT) [3] prompting involves adding a series of intermediate reasoning steps to the intended task. It has been shown to improve performance in more complex prompts. These intermediate reasoning steps can be thought of as giving an answer to a similar problem but with detailed steps on how to get to that answer as shown in Figure 1. This is similar to few-shot prompting, which also provides examples of answers to similar questions, however few-shot prompts include less detail that is they usually provide only questions and answers, no intermediate reasoning steps in between. [3] An example of Chain of Thought prompting can be seen in Figure 2



Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

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 do they have?

A: The cafeteria had 23 apples originally. They used 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. ✓

Figure 2. Example of a Chain of Thought prompt versus a regular prompt

Datasets

Because commonsense reasoning is a complex topic and encompasses a wide collection of different tasks, it is difficult to generate a dataset that will evaluate the model’s generalization abilities.

The Winograd schema challenge [4] was the standard for testing the commonsense reasoning, however from 2019 it is considered defeated since numerous transformer-based models achieved over 90% accuracy. [5] For that reason, we will use more recent and challenging datasets.

On the one side, we have different SOTA methods [3] evaluated on task specific datasets like GSM8K [6]. This dataset focuses on linguistically diverse grade school math problems created by human problem writers.

On the other side, there are benchmarks that combine multiple commonsense datasets. For this project we will use Rainbow which is a suite of commonsense benchmarks that contain multiple choice question-answering datasets. Rainbow contains the following datasets

- αNLI [7] - tests abductive reasoning in narratives. Models need to find the best explanation among the presented options connecting a beginning and ending;

- CosmosQA [8] — asks commonsense reading comprehensions about narratives in everyday situations

- HellaSWAG [9] — models need to find the most plausable ending to a short content.
- PIQA [10] — question answering benchmark for commonsense reasoning.
- SocialIQa [11] — commonsense reasoning about social situations and interactions
- WinoGrande [12] — large-scale collection of Winograd schema-inspired problems that test reasoning about social and physical interactions.

Other possibilities for testing commonsense reasoning are the GLUE [13] and SuperGLUE [14] datasets, which are benchmarks for general language understanding systems.

1. Results

1.1 Setup

Our plan was to test our implementation on two promoting strategies. We used the DSPy framework [15] to reshape prompts into CoT prompts, and used the zero-shot prompts as a baseline. We generated predictions on a subset of 100 examples from the Physical IQa dataset using the Llama3-8B-instruct model.

1.2 First results

Initial results for the chain-of-thought prompting technique are promising. In a benchmark test where we provided the model with just the question, the model achieved approximately 55% accuracy, which is only marginally better than guessing in datasets where the model is given a question and two possible answers. However, upon implementing the chain-of-thought technique, we observed a significant increase in accuracy, reaching over 70%. This 15% improvement is a notable achievement for our first and rudimentary implementation of the COT technique.

Evaluation

We are currently performing evaluation of the models answers with two different methods. In the first method where the model has to select between two choices and then give his answer we use a simple comparison algorithm. If the model gives multiple answers, we count up the choices and use his most often given answer as the answer we use for comparison. The comparison algorithm is then a simple checker which has the list of correct answers. The checker compares the question id and content from the correct list with those give by the model. And then prints the results.

The other comparison method is a much more involved process for when the model gives a descriptive answer. In this method, we are using an uncased BERT model to generate the embedding and calculate the cosine similarity between the correct answer and the given one. As an additional layer of grading we are comparing cosine similarity with Word2Vec model to better determine the meaning of the sentence these two scores are then combine and if the answer pass the threshold of correctness we are considering the answer as correct. This model also supports comparison of different techniques, and not just direct comparison between the result and the technique. This way we can determine which technique is closer to the right answer even if this particular answer is incorrect.

2. Discussion

In spite of the promising results, we encountered a challenge related to increased run times when employing the chain-of-thought technique, even on an HPC cluster. We attribute this issue mainly to the current implementation's inefficiency. Addressing this inefficiency will be a key focus in refining the chain-of-thought approach for real-world applications. As our main goal is to compare multiple prompt strategies, our next steps will be implementing additional strategies and additional models and creating a detailed analysis of the final results.

References

[1] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.

[4] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

[5] Hector J. Levesque. On our best behaviour. *Artificial Intelligence*, 212:27–35, 2014.

[6] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Yih, and Yejin Choi. Abductive commonsense reasoning. *ArXiv*, abs/1908.05739, 2019.

[8] Zhixing Tian, Yuanzhe Zhang, Kang Liu, Jun Zhao, Yantao Jia, and Zhicheng Sheng. Scene restoring for narrative machine reading comprehension. 2020.

[9] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? 2020.

[10] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. 2020.

[11] Gengyu Wang, Xiaochen Hou, Diyi Yang, Kathleen McKeown, and Jing Huang. Semantic categorization of social knowledge for commonsense question answering. 2021.

[12] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. 2020.

[13] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. 2018.

[14] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. volume 32, 2019.

[15] *DSPy Documentation*, 2024.