FRI Natural language processing course 2024
Project report

University of Ljubljana
Faculty of Computer and
Information Science

LLM Prompt Strategies for Commonsense-Reasoning Tasks

Matic Pristavnik Vrešnjak, Mitja Kocjančič, and Songeun Hong

Abstract

The surge in the popularity of Large Language Models (LLMs) such as chatGPT, PaLM, and Gemini has led to their widespread adoption in both personal and commercial domains. Many of these cutting-edge models rely on the transformer architecture. With the increasing use of LLMs, there is a growing need to devise prompts that facilitate the generation of relevant and informative responses, particularly for tasks necessitating commonsense reasoning. Such tasks draw upon everyday knowledge for resolution. Consequently, various prompt strategies have emerged to enhance model performance on such tasks. In response to the expanding array of prompt strategies, this paper offers a comprehensive comparison of each approach, aiming to shed light on their effectiveness and applicability in enhancing LLM performance.

Keywords
Large Language Models (LLMs), Prompt Strategies, Comparative Analysis

Advisors: Slavko Žitnik

Introduction

Recent years have seen a rise in the popularity of Large language models (LLMs). LLMs such as chatGPT, PaLM, Gemini, and others have already been adopted for personal and commercial use. Many of the current state-of-the-art models are based on the transformer architecture[1]. Because of the wide adoption of LLMs, it has become more important to provide prompts that encourage the model to generate relevant and informative responses to tasks that require commonsense reasoning. Tasks that require commonsense reasoning are those that require everyday knowledge to solve. This has led to the development of several strategies for providing prompts that allow models to perform better on such tasks. Due to the growing number of prompt strategies, we present a comprehensive comparison of each strategy.

Related work

One strategy for providing prompts is Chain of thought (CoT) [2]. The core idea of CoT is that we give the model a demonstration of how to come to the solution or ask it for a step-by-step response. CoT has been proven to improve the performance of LLMs on commonsense reasoning and mathematical datasets. One limitation mentioned in the original CoT paper was that CoT only benefits large models like chatGPT and not smaller models. Frameworks like DSPy[3] have also integrated functionality that enables the implementation of custom Chains of thought. Strategies with similar approaches are In-context Learning[4] and Plan-and-Solve Techniques[5].

Another example of prompt strategies is Promptbreeder[6]. In contrast to methods like CoT, which are based on hand-crafted prompts, Promptbreeder can generate optimized prompts. Promptbreeder relies on LLM and genetic algorithms to mutate a set of prompts, which are then given to the LLM. Strategies that use generated prompts have been shown to perform better than the prompt strategies that are based on hand-crafted prompts.

The aforementioned strategies are more complex and require additional text to be added to the instruction. Simpler strategies are more oriented towards creating prompts that are clear and precise[7]. This includes forming unambiguous and specific prompts, allowing the model to give better and more detailed answers. Research has also shown that models can provide better answers when given emotional prompts[8].

Previously mentioned strategies have been tested on various LLMs. LLMs that are most commonly used for testing are chatGPT[9], PaLM[10], Gemini[11] and other models. Most researchers prefer to use models with a large number of parameters because larger LLMs benefit more from prompt engineering. This has led to very little research being done on smaller models.

Because of the growing number of prompting strategies, many datasets have been used to measure their performance. Examples of such datasets are the Winograd Schema Challenge[12], Textbook Question Answering[13], SocialIQA[14], and many other datasets. Many of the mentioned datasets were made to challenge LLM on commonsense sense reasoning tasks.

Objectives

In this project, our specific objectives are as follows:

1. **Comparison of Various Prompt Strategies**:
   - We aim to compare the performance of different prompt strategies through carefully designed experiments under specific conditions. The goal is to assess their effectiveness in various commonsense reasoning tasks quantitatively.
2. **Analysis of Commonsense Reasoning Processes**:
   - Through meticulous analysis, we seek to understand the impact of each prompt strategy on the reasoning processes of the language models. This involves identifying how each strategy influences the decision-making process of the models, with the aim of enhancing their inference capabilities.
3. **Derivation of Optimal Prompt Strategies**:
   - By analyzing the experimental results, we intend to derive optimal prompt strategies and propose the most effective approaches for commonsense reasoning tasks. Our objective is to maximize the performance of the models and enhance their applicability in real-world scenarios.
4. **Evaluation of Generalization Potential**:
   - We aim to evaluate the potential of generalizing the findings of this project to different contexts and tasks. This assessment will help determine the broader applicability of our research across various domains.

Initial Idea

1. **Dataset Selection**:
   - **Winograd Schema Challenge (WSC)**: WSC encompasses a spectrum of scenarios mimicking real-world situations, rendering it apt for assessing a model's common-sense reasoning prowess. We shall utilize WSC to scrutinize and juxtapose the common-sense reasoning proficiencies of various models.
   - **Textbook Question Answering (TQA)**: TQA evaluates the adeptness to grasp textbook content and respond to associated queries. Employing this dataset will enable us to gauge whether a model can assimilate and apply the requisite common-sense knowledge to address real-world predicaments.
   - **SocialIQA**: SocialIQA scrutinizes the acumen in social contexts, assessing common-sense communication aptitudes. This dataset will evaluate a model's capacity to discern common-sense decisions across diverse social scenarios.
2. **Prompt Strategies**:
   - **Chain of Thought (CoT)**: The CoT strategy steers the model through sequential problem-solving processes or solicits step-by-step responses to enrich the model's common-sense reasoning acuity.
   - **In-context Learning**: This strategy facilitates the model in assimilating novel information within a given context, empowering it to conduct superior reasoning by leveraging previously acquired knowledge.
   - **Plan-and-Solve Techniques**: This strategy furnishes explicit blueprints and directs the model to resolve problems methodically. This is anticipated to elucidate the model's reasoning trajectory and refine its logical reasoning skills.
3. **Experiment Design and Analysis Plan**:
   - Each experiment pertaining to both datasets and prompt strategies shall be meticulously delineated as follows:
     (a) **Dataset Selection and Preprocessing**: We shall leverage sentiment inference within three datasets. Each dataset will be segregated into training, validation, and test subsets, with judicious sampling techniques applied to mitigate data imbalances.
     (b) **Implementation of Prompt Strategies**: Three prominent prompt strategies shall be implemented. Moreover, we shall tailor or extend these strategies for fine-tuning and devising novel approaches.
     (c) **Experiment Design**: Experiments for each prompt strategy will adhere to a uniform structure. Each strategy will undergo training utilizing identical model architecture and hyperparameter settings. Experiments will be conducted under reproducible conditions to ensure replicability.
     (d) **Performance Measurement**: We shall gauge the model's training and inference durations for each strategy. Performance metrics will encompass accuracy, precision, recall, and F1 scores across each dataset.
     (e) **Experiment Execution and Analysis**: Experiments for each strategy will incorporate measures to ensure stability, such as cross-validation or bootstrapping. The resultant

outcomes will be juxtaposed to dissect the performance of each strategy, with particular emphasis on discerning performance disparities in prevalent sentiment inference tasks.

(f) **Result Interpretation and Reporting**: Experiment findings will be expounded to discern the strengths and weaknesses of each strategy. Drawing upon this analysis, a conclusive report will be compiled, advocating the optimal prompt strategy and outlining avenues for future research.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[3] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[4] Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey, 2024.

[5] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, 2023.

[6] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution, 2023.

[7] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review, 2023.

[8] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli, 2023.

[9] OpenAI et al. Gpt-4 technical report, 2024.

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[11] Gemini Team et al. Gemini: A family of highly capable multimodal models, 2023.

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

[13] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017.

[14] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019.