Abstract. What is the problem, issue and key finding.

Introduction: whats the problem. Why is it important.

Current Research: what are the finding of the current research and summarize

Model development: what model are proposing and why

Analysis: what is the finding

Conclusion

Abstract: LLM has been a heated topic since ChatGPT has showcased impressive conversational capability and other NLP capabilities that involves reasoning and multitasking. It has been a critical topic to understand and train LLM with enhanced problem-solving and reasoning skills. Prompt engineering becomes the most popular approaches to help LLMs fune tune its capability when executing the downstream task without heavy manual labor and has achieved success. This paper presents a literature review in why prompt engineering is a good solution to improve LLMs’ capability and the current best practices adopting the prompt engineering. The paper will also discusses the limitation to the LLMs and provides a future outlook.

Introduction:

Even though we have been researching natural language processing for years, ChatGPT did break a record. It becomes the first app that surpassed 1 billion downloads within a month which tells the enthusiasm that hopes ChatGPT can become the next revolution to everyday life like airplane or internet. ChatGPT, the most recent success of LLMs, much smarter than Alexa and Google Voice, displays its expansive zero-shot capabilities in human level conversation. ChatGPT has already showed capabilities no less compared to humans in several domains (Guo et al., 2023) and in some cases, ChatGPT’s answer is preferred compared to experts because ChatGPT provides the sympathy during medical consultation that a normal doctor would feel but lacks the energy to spill out.

We now refer to models with only over 100B parameters as LLMs which is extremely powerful yet resource draining. All the LLMs are trained by top-tier tech companies. The first open source project Bloom (Workshop et al., 2022) was trained on a super computer for 3.5 months sponsored by the French Government. Since it is so resource consuming to build and train LLMs, finding an efficient yet accurate approach to enhance LLMs capability to solve downstream task becomes very important. This is why prompting comes into place.

First in the traditional ML, we first used well labeled supervised data to fine tune the model. Then we apply reinforcement learning to rank the answers. We then fine tune the model based on the downstream task. This approach is accurate yet requires a lot of careful engineered human labels. Prompting proposes that you just by giving machine good prompt can lead the machine to perform downstream tasks well.

ChatGPT is trained on unsupervised data but with well instructed prompts. Prompting has been used in InstructGPT which shows better performance results than GPT3 even though Instruct has only one tenth of the parameters (Ouyang et al., 2022).

Current Research

What we are trying to do here is to replace or at least reduce the labor of the tuning process while achieves the similar or better tuning performance. Studies have shown that LLMs have achieved success with zero-shot or few shot with no training.

Prompting started with searching for the template which is an auto template learning and manual template engineering performed by humans.

Prompt is the language that humans can interact with machine to help machines understand the requirements and the tasks to execute. It is shown that with zero-shot and few-shot LLMs can perform tasks even with no training (Workshop et al., 2022), zero-shot with examples which can achieve robustness and generalization and chain of thought prompt to improve reasoning capabilities.

There are other approaches that are also prompt free. The mutual information template (Sorensen et al., 2022) detects the amount of mutual information between the entities can greatly affect the accuracy. The PERFECT adapter (Mahabadi et al., 2022) is a prompt-free layer that flexibly helps the LLMs adjust to the downstream task based on the task.

Prompts are good but there are so many ways to prompting. The knowledge aware prompt suggests tuning has more success if the prompts has more relation extracted based on the task (Chen et al., 2022). Prompt patterns are studied in a same way to software patterns to increase the accuracy (White et al., 2023). We even have an automated solution of LLMs as a human level prompt engineer (Zhou et al., 2022).

Why we are prompting? First, scalability. Research has shown that smaller models benefited a lot with more training data but the story is slightly different for LLMs. For LLMs, the more parameters you have, the better it is (Workshop et al., 2022). This also makes sense considering the computational cost and affordability. Second, prompting can increase accuracy which has been tested in multitasking and reasoning tasks (Bang et al., 2023). But we also need to be careful with prompting as researches show that irrelevant prompting can lead to bad results(Mahabadi et al., 2022). Therefore, prompting helps us increase efficiency, accuracy while reducing training time and memory/storage which is critical in training LLMs.

Model development:

The methodology is similar across studies. First, the author will propose a plan for prompt or non-prompt approach. Second, set of reasoning and inference tests are introduced. Third, the model is tested against other state-of-art models or models with even more powerful parameters. Last, the performance are compared and results are analyzed.

Analysis:

Limitations:

Though LLMs already show the many wonderful things it can do, there are still major limitations and risks associated. The main limitations which also associated with all LLMs are the AI hallucinations. Extrinsic hallucination is the main hallucinations that happens to ChatGPT. Extrinsic hallucination means that ChatGPT will make up things that it does not know. Intrinsic, on the other hand, is ChatGPT will say untruthful facts against its own information (Bang et al., 2023).

The major risks we have about LLMs are the unknown part of AI. LLMs are like we put all the information on the internet into a blackbox and then we start to teach it logic and reasoning of a six year-old. Once the model has some basic understanding of inference, summarization capabilities we already start imaging it can automate half of the workforce. The machine never gets to learn about the boundaries in this entire process. Once it learns and understands to a certain level, the machine can start prompt itself. Since no LLMs is connected to the internet and every model is trained locally, the model’s accessibility is limited. This does not remove the possibility of the machine developing rudimentary intelligence. Right now, the machine is said to have “lazy reasoning” (Bang et al., 2023) which can improve a lot by prompting step-by-step. The training process is quite similar to training teenagers and college students with basic problem-solving skills. There are also some good applications out there, for examples using LLMs to red team other LLMs to detect violations of privacy data and inappropriate use of offensive language. Or the application to deploy prompt chain which executes several prompts to achieve complex situations in real life. We still need to be careful with what AI is empowered with and capable of.

Conclusions: