CS598TZ: Autonomous Prompting

Determining best prompting strategy with respect to user query

I) Group members

- Ishaan Singh : is14@illinois.edu
- Henry Yi : weigang2@illinois.edu
- Saharsh Barve : ssbarve2@illinois.edu
- Veda Kailasam : vedak2@illinois.edu

II) Project Description & Goals

1. Description

Recent advancements in large language models (LLMs), have demonstrated that enhancing the quality of input prompts can significantly boost model performance across a range of tasks. However, determining the most effective prompting strategy for each task remains a critical challenge, often involving manual processes that hinder scalability.

This project proposes to build on top of existing strategies by developing an autonomous system that can balance prompt generality and specificity while efficiently generating diverse prompts tailored to different input types. By automating the generation and selection of high-quality prompts, we aim to minimize manual intervention and enhance the overall utility of LLMs for diverse applications.

2. Related Work

After several years after large language models' emergence, the ability of LLM could be amenable to many different tasks by simply tweaking the prompts. As Brown (2020) , indicates, as the scale of the LLM reaches a certain level, by providing some examples in the prompt, the LLM could better handle the new requests in the same task. Later, Wei (2022), showed asking the model to produce

the intermediate reasoning steps could vastly improve the model's arithmetic, commonsense and symbolic reasoning capabilities, which were really hard to improve even if you feed more data and further increase the scale of the model. This prompting technique is called Chain of Thought (CoT). However, even with this technique, LLM could still produce wrong reasoning paths and reach wrong answers. Several techniques were invented to improve the consistency of the reasoning process of the LLM, such as Self-Consistency (Wang, X., et al, 2022), Tree of Thought (Yao, S., et al, 2024), Complexity-Based CoT-SC (Fu, Y., et al, 2022). In the meantime, several other prompting techniques were proposed to improve the arithmetic, and reasoning capability by incorporating other tools, such as SymbCoT (Xu, J., et al, 2024) aiming for improve the model's symbolic reasoning capability by incorporating a logical form translator and PAL (Gao, L., et al, 2023) aiming for improving the accuracy of the answers produced by LLM by delegating the calculation and solution process of the LLM to a third-party runtime, such as a Python interpreter.

3. Goals:

Current approaches either rely on a fixed prompt for all tasks, which lacks flexibility, or dynamically generate prompts for each input, which is computationally expensive and prone to errors in reasoning steps. The primary goal of this project is to improve the efficiency and effectiveness of prompt generation for large language models (LLMs) by automating the process and addressing current limitations. We aim to explore and evaluate existing prompting methods, including Zero-Shot, Few-Shot, Chain-of-Thought (CoT), and Automatic Chain-of-Thought (Auto-CoT) techniques, to gain insights into their strengths and weaknesses across diverse tasks.

Approach

Our approach involves conducting a series of experiments where we modify key components of the prompt optimization process. These modifications include, but may not be limited to, altering the **evaluator** to improve prompt relevance

scoring, refining the **ranking function** to better prioritize high-quality prompts, and employing **tree search techniques** at the prompt sampling level to enhance prompt diversity and reduce computational overhead. We also aim to benchmark an ensemble of changes to see if the effects are compounded and whether a universal ensemble prompting strategy might provide better results than dynamic prompting.

Benchmarks

We will benchmark our results using standard evaluation metrics from the benchmarks, such as **GSM8K**, **MultiArith**, and **AQuA**. These benchmarks focus on tasks like arithmetic reasoning, commonsense reasoning, and multi-step problem solving, providing a comprehensive evaluation of the proposed approach.

Implementation

We plan to conduct our evaluations using small, open-source models (within the range of LLaMa 3.2 3B to LLaMa 3 8B). We will use existing benchmarks and randomly sample n rows from the datasets to benchmark each system against (within a reasonable range of trials to limit overspending on compute). Using an open-source model allows us to access richer information like log probabilities of tokens, logit distributions and more as well as use MCTS on sampling. Tree search techniques like MCTS in dynamic prompting systems enable intelligent exploration of diverse, effective prompts while adapting to real-time inputs. This approach optimizes computational resources by strategically sampling prompts, learning successful patterns, and avoiding exhaustive searches.

III) Resources

- **Datasets: GSM8K, MultiArith, AQuA**
- **Compute**: A100 (either via NCSA or Runpod)

IV) References:

- Automatic Chain of thought Prompting Strategy in Large Language Models: https://arxiv.org/pdf/2210.03493
  https://github.com/amazon-science/auto-cot
- Automatic Prompt Selection for Large Language Models: https://arxiv.org/pdf/2404.02717v1
- https://medium.com/the-modern-scientist/best-prompt-techniques-for-best-llm-responses-24d2ff4f6bca
- Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.
- Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2022, October). Complexity-based prompting for multi-step reasoning. In The Eleventh International Conference on Learning Representations.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36.
- Xu, J., Fei, H., Pan, L., Liu, Q., Lee, M. L., & Hsu, W. (2024). Faithful Logical Reasoning via Symbolic Chain-of-Thought. arXiv preprint arXiv:2405.18357.
- Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., ... & Neubig, G. (2023, July). Pal: Program-aided language models. In International Conference on Machine Learning (pp. 10764-10799). PMLR.