Course Project Proposal for CSE561 Fall 2024

Zheyuan Wu
Department of Mathematics
Washington University in St.Louis
1 Brookings Dr, St. Louis, MO 63130
w.zheyuan@wustl.edu

Abstract

Project choice: Designing a novel algorithm to do inference on large language models (white box models such as LLaMA2 models, or black box models such as GPT-4, CLAUDE, etc.) to solve some type of complex problems, and analyze their limitations.

1 Main goal of project

To investigate how to improve the memory of a large language model (LLM) such as GPT-4o without relying on additional pre-training.

One significant constraint faced by LLMs is the context window size. This can lead to situations where the model forgets previous steps when working through complex problems, particularly in scenarios that require Chain-of-Thought (CoT) reasoning[7].

In this final project, the focus is on studying how LLMs store and manage "knowledge" and "references". Knowledge in this context refers to the information and insights the model has learned and can use to generate responses. It is derived from the vast amounts of text data that the model has been exposed to during training. References, on the other hand, involve how the model cites or retrieves specific pieces of information relevant to the problem at hand. The challenge is to enhance the model's ability to effectively retain and utilize this knowledge throughout the problem-solving process.

By addressing these issues, the project aims to explore methods to improve memory management and enhance the model's performance in handling complex queries. This may involve developing new strategies for context management, incorporating external memory systems, or optimizing the model's ability to retain and reference critical information over longer interactions.

2 Importance of project

The memory problem of LLM has persisted for a long time. Some recent research finds that modifying where important information is placed within the language model's input context—such as moving the section that answers a question—creates a U-shaped performance pattern. The model performs better when this information is at the very start (primacy effect) or at the end (recency effect) of the input context, but its performance drops notably when the information is situated in the middle of the context. [6]

However, in real life, we need to study and gather tons of information when solving problems, an agent must be aware of many aspects of a question before making correct and consistent decisions. Increasing the memory size or other methods that help LLMs to gain information in the large corpus is essential to make the models solve problems like a human expert.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

3 Tentative plan

3.1 Time-line

3.1.1 September

Investigate possible approaches and articles related to LLM memories and propose a Transformer architecture that might increase the performance of small context window models to solve problems requiring large corpus data collection.

3.1.2 October

Test on different architectures and collect data.

3.1.3 November

Compose the paper and analyze the data

3.2 Tentative approaches

There are various promising approaches to address and provide insights into solving the memory problem in large language models (LLMs). Each method offers unique strategies for extending the model's ability to manage and utilize extensive context, which is crucial for improving performance on complex tasks.

Firstly, Space Mamba has shown significant potential in addressing computational inefficiencies associated with transformers and LLMs when dealing with longer sequences. This model improves processing efficiency for small to medium NLP tasks by optimizing how information is managed across extended contexts [4]. By leveraging Space Mamba, we might be able to overcome some of the inherent limitations of the traditional transformer architecture, potentially enhancing the LLM's ability to handle larger amounts of information.

Theoretically, exploring how LLMs acquire and utilize complex skills is another critical avenue of research. Understanding the mechanisms behind skill development and the types of information that facilitate skill acquisition can provide valuable insights into how to extend LLM memory. This research aims to identify methods and findings that can be applied to enhance the model's problem-solving capabilities [1]. By incorporating these theoretical insights, we can develop more effective techniques for managing and recalling information within the context window.

Another relevant approach is detailed in a paper discussing the certainty of LLMs in problem-solving [5]. This research focuses on how models express and handle uncertainty, which can be instrumental in determining when to terminate the search or prompting process. Understanding and incorporating measures of certainty can help optimize when and how the model utilizes its context, potentially leading to more accurate and efficient problem-solving.

Additionally, the LongNet architecture presents an innovative solution for scaling sequence lengths to over 1 billion tokens without compromising performance on shorter sequences [3]. This Transformer variant demonstrates a significant leap in handling extended contexts, making it a promising candidate for extending the memory capacity of LLMs and improving their ability to manage large-scale data.

Furthermore, exploring memorizing transformers and their approaches could provide additional strategies for extending Transformer architectures [8]. These methods focus on enhancing the model's ability to retain and recall information across longer contexts, potentially offering practical solutions for memory limitations.

Lastly, integrating concepts such as the Graph of Thoughts and other search methods might provide valuable support for solving problems with a limited context window [2]. By employing these techniques, we can enhance the model's ability to organize and retrieve relevant data, facilitating better problem-solving even with constrained memory resources.

By investigating and applying these diverse approaches, we can develop more robust methods for improving LLM memory, leading to enhanced performance across a range of complex tasks and applications.

3.3 Experiments to be conducted

I will try to use different frameworks to add additional memory or networks to facilitate problem-solving for GPT models in solving a problem that requires large text generalization and understanding. A long paragraph exceeding the context window size of GPT-4o will be used to test the new architectures for the Transformer models.

3.4 Data collection

I'm currently looking for a large text database available on Hugglingface and try to sample them to generate a dataset for testing memory of LLM.

GPT-generated articles may also be used when we test the long-text reading ability of the model under our methods. We may provide an outline to a long story exceeding the context window size and use least-to-most prompting [9] to generate subsections for the article and feed the LLM in various architectures to test their ability to solve these problems.

References

[1] Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models, 2023.

[2] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, March 2024.

[3] Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens, 2023.

[4] Emadeldeen Hamdan, Hongyi Pan, and Ahmet Enis Cetin. Sparse mamba: Reinforcing controllability in structural state space models, 2024.

[5] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022.

[6] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

[7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

[8] Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers, 2022.

[9] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023.