LLM Memory Evaluation and Enhancement Research Formal Proposal

Ryunosuke Saito
Department of Computer Science
Johns Hopkins University
rsaito1@jh.edu

Owen Bianchi
Department of Computer Science
Johns Hopkins University
obianch1@jh.edu

Samer Aslan
Department of Computer Science
Johns Hopkins University
saslan1@jh.edu

Abstract

Large Language Models (LLMs), such as GPT-4 and ChatGPT, demonstrate impressive language generation capabilities but encounter challenges related to memory. Notably, these models exhibit undesirable characteristics, including forgetfulness and hallucination, which can compromise their overall performance and reliability in various natural language processing tasks. Forgetfulness pertains to the model's tendency to inadequately retain relevant information from the beginning of a text or conversation, while hallucination refers to generating outputs containing fictional details not present in the input data. These issues significantly impact LLMs' performance, reliability, and trustworthiness in real-world applications, particularly during interactive conversations with users. Despite this, the domain lacks a well-defined memory evaluation standard. In response, our research focuses on addressing LLM memory performance, aiming to define an evaluation framework and enhance memory capabilities, specifically emphasizing conversational dialogue. By advancing memory management in LLMs, we seek to develop more robust and dependable models capable of delivering coherent, accurate, and trustworthy outputs in various language processing scenarios.

1 Introduction

The development of LLMs, such as ChatGPT and GPT-4, has brought remarkable advancements to the field of Natural Language Processing (NLP), empowering machines with human-like language generation abilities. However, despite these impressive capabilities, LLMs face inherent challenges associated with their memory capacity, specifically manifested in the form of forgetfulness. This characteristic limits their ability to effectively retain critical information from earlier segments of long conversations, resulting in the loss of context and personalized responses. Additionally, LLMs demonstrate the phenomenon of hallucination, generating imaginative or fictional content that lacks grounding in factual or contextually appropriate data. These limitations considerably affect their performance, reliability, and trustworthiness in practical applications, especially during interactive dialogues with users. Efforts to improve LLM memory have led to research focusing on expanding the context window as a potential solution; however, empirical evidence has indicated its ineffectiveness. [3] Despite ongoing research, a lack of well-defined memory evaluation standards and standardized testing datasets hampers objective quantification of memory performance in LLMs, resulting in an overreliance on anecdotal reports and subjective assessments. To address these challenges, our research endeavors to enhance LLM memory performance by establishing a comprehensive evaluation framework. We place a specific emphasis on conversational dialogue, as it is a pivotal context for assessing memory capabilities. Our proposal includes the selection of two key evaluation metrics, accompanied by a rationale for their choice. Moreover, we outline our proposed methodology for dataset construction, taking into account the unique requirements of memory evaluation. Establishing a robust evaluation standard and dataset would enable objective quantification of LLM memory performance, which represents a critical aspect in advancing the capabilities of these powerful artificial intelligence models. By augmenting their memory capacity, LLMs can better store and recall relevant information from earlier parts of a conversation beyond token limits, facilitating more effective communication and closely emulating human memory.

2 Relevant Literature

Amidst the recent boom in NLP research centered around LLMs, there exist some literature focused on the memory capabilites of these models. There are those that hone in on the evaluation of memory to some degree and those that are more focused on the enhancement of memory. Our proposed research adds a unique and more comprehensive angle at the evaluation of memory for LLMs along with a novel model for enhanced performance.

2.1 Memory Evaluation

As aforementioned, one of the most popular proposed methodologies for improving the memory performance of LLMs is increasing the context window. However, the study titled ”Lost in the Middle: How Language Models Use Long Contexts” [**longcontexts**] demonstrates that the effective utilization of extended context in LLMs remains unclear. The study analyzes language model performance in tasks requiring the identification of relevant information within the input context. Results show that performance is strongest when relevant information is at the context's beginning or end, degrading significantly when it lies in the middle. These findings highlight the limitations and challenges that arise when language models process long input contexts. Despite their ability to handle lengthy context, there is still a lack of knowledge regarding how effectively they utilize this extended information. Therefore, an alternative approach to enhancing memory is needed rather than merely increasing the context window.

2.2 Memory Enhancement

The paper titled ”Unleashing Infinite-Length Input Capacity for Large-scale Language Models with Self-Controlled Memory System” [2] presents a viable solution to address the limitation in LLMs ability. The proposed ”Self-Controlled Memory (SCM)” system aims to overcome the challenge of processing lengthy inputs faced by large-scale LLMs. SCM empowers LLMs to handle ultra-long texts without necessitating modification or fine-tuning. Comprising the language model agent, memory stream, and memory controller, SCM exhibits successful results in achieving multi-turn dialogue capabilities comparable to ChatGPT, while also outperforming it in tasks such as ultra-long document summarization and long-term conversations. This paper serves as a source of inspiration for our proposed model, as it presents valuable ideas that we intend to refine and enhance further. Particularly noteworthy is the SCM system's ability to seamlessly integrate with any other LLM, effectively acting as an upgrade to existing models.

3 Methodology

3.1 Evaluation Metrics

Memory is a multifaceted concept with diverse applications. In our research, we have diligently explored various potential applications of memory and have developed quantifiable metrics to assess performance in these contexts. To define the framework for evaluating the memory capabilities of LLMs, we pondered a fundamental question: What attributes of memory are considered desirable? From a comprehensive set of answers, we identified two key applications that best encompass memory's essence in evaluating LLMs, focusing on their ability to avoid forgetfulness and hallucination.

The first evaluation metric, termed ”Important Points Selection,” assesses how effectively a model can select (X) essential key points from a larger pool (Y) of potentially important dialogue points. This metric delves into the model's proficiency in recalling crucial facts from the context, ensuring that significant information is retained while avoiding the fabrication of false details. Specifically, we will measure the LLMs Retrieval Rate, indicating the percentage of the X key points successfully retrieved; the Top-Z Retrieval Rate, which evaluates the ranking of the most important facts retrieved; and the Hallucination Rate, representing the percentage of retrieved key points that are inaccurately fabricated.

The second evaluation metric, titled ”Question-Answering Correctness,” examines how well a model can provide accurate answers to pertinent questions based on the dialogue context. In addition to the model's capacity to form memories, it is equally essential to ascertain its ability to appropriately utilize these memories in response to specific prompts. The evaluation will include a set of multiple-choice questions about the dialogue context, designed to test the model's ability to recall relevant memories and provide correct answers.

These proposed evaluation metrics aim to furnish a much-needed and comprehensive framework for assessing LLM memory capabilities. By focusing on key applications of memory in avoiding forgetfulness and hallucination, we believe these metrics will significantly contribute to the understanding and improvement of LLM memory performance.

3.2 Dataset Creation & Evaluation

In order to rigorously evaluate the memory capabilities of LLMs, we will construct a dataset incorporating specific features, including long conversational dialogues, ranked key facts pertaining to each dialogue, and multiple choice questions associated with each dialogue.

Current dialogue datasets suffer from a drawback of being relatively short and having a limited number of turns per person, which does not meet the requirements for evaluating memory effectively. To address this challenge, we explored different dataset generation options, including generating a dataset from scratch—either human-written or LLM-generated. However, human-written datasets are not favorable due to their high cost, challenges in quality control, and real-time customization. Similarly, using an LLM to generate the dataset raises concerns about evaluating LLMs using data from another LLM.

After careful consideration, we arrived at a solution that involves utilizing an LLM to augment a high-quality human-written dataset, thus offering a favorable balance. In this approach, we achieve the advantages of both human-written and LM-generated datasets.

During the dataset selection process, we considered several options, including DailyDialog [1], PersonaChat [5], and Multi-Session Chat [4]. Preliminary evaluations indicated that Multi-Session Chat (MSC) stands out as the most suitable choice. MSC is a human-written dialogue dataset, extending PersonaChat, and comprises relatively long dialogues. Each sample in MSC represents a conversation where users are provided with prompts based on specific personality traits to play designated characters. The original conversations are divided into five separate sessions, simulating time pauses. For our evaluation purposes, we will combine these sessions to create one continuous, natural, and high-quality long conversational dialogue. This dataset augmentation approach ensures an ideal setting for evaluating memory in LLMs.

For evaluating the selection of important points, we will task an LLM with parsing the dialogue into X tokens and generating key facts about the context. Additionally, the original human-generated personality traits from MSC will provide a separate means of evaluating important points selection.

For the question-answering evaluation, we will prompt a Question-Answering Language Model (QA LM) to create relevant multiple-choice questions about the context. Particular attention will be paid to employing diverse prompting methods and manual data quality checks.

Once our memory-evaluation dataset is ready, we will proceed with evaluating the memory abilities of current state-of-the-art models. This evaluation will involve presenting the dialogue as context and asking the model to identify relevant key points, determine which personality traits pertain to the users, and respond to the multiple-choice questions. The dataset, along with these evaluation tasks, will enable us to comprehensively assess the memory performance of LLMs and advance the understanding of their capabilities.

3.3 Enhanced Memory Model

At this point, the goal is to create a new model that will increase an LLMs performance on these metrics. An idea that we had was to focus on the use of summarization models to condense prior chat contexts, essentially creating a meaningful, compact memory for the LLM. Another idea that we are brainstorming now is the formation of memories through a question/answer pipeline in which we form memories by answering key questions about the dialogue throughout the conversation. These memories get stored separately, and the model will call upon the memories when needed in order to respond to a user. The architecture of the storing and usage of memories will be inspired by the paper mentioned above, ”Unleashing Infinite-Length Input Capacity for Large-scale Language Models with Self-Controlled Memory System” [2].

4 Expected Outcomes, Timeline, and Budget

The primary expected outcome of this research is the establishment of a well-defined evaluation standard for LLM memory. This will be achieved through the formulation of two precise evaluation metrics and the development of an accompanying dataset, enabling comprehensive assessment of LLMs. This is the first big milestone we hope to achieve before the start of the semester. Estimated cost here depends on the size of the dataset we decide to create and cost of the LLM we chose to use, the aim is to keep it under $100.

The second expected outcome is the creation of an improved memory model that surpasses the performance of current state-of-the-art LLMs according to the identified metrics. This enhancement will address the existing limitations of LLM memory, paving the way for more effective and reliable language models. This is the second big milestone, which we aim to achieve by the end of October. The cost here depends on any datasets we create as well as any computational costs of training models.

Finally, the research aims to contribute to the field of NLP by producing a formal research paper and submitting it for publication. By disseminating the findings, this research intends to advance the knowledge and understanding of LLM memory, serving as a foundation for future studies to build upon and foster further progress in the NLP domain. This is the final milestone which we aim to achieve by the end of the semester and have submitted for publication.

References

[1] Yanran Li et al. *DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset*. 2017. arXiv: 1710.03957 [cs.CL].

[2] Xinnian Liang et al. *Unleashing Infinite-Length Input Capacity for Large-scale Language Models with Self-Controlled Memory System*. 2023. arXiv: 2304.13343 [cs.CL].

[3] Nelson F. Liu et al. *Lost in the Middle: How Language Models Use Long Contexts*. 2023. arXiv: 2307.03172 [cs.CL].

[4] Jing Xu, Arthur Szlam, and Jason Weston. *Beyond Goldfish Memory: Long-Term Open-Domain Conversation*. 2021. arXiv: 2107.07567 [cs.CL].

[5] Saizheng Zhang et al. *Personalizing Dialogue Agents: I have a dog, do you have pets too?* 2018. arXiv: 1801.07243 [cs.AI].