Episodic Memory in Large Language Models

1st Gabriel B. Ladislau
Department of Informatics
Federal University of Espirito Santo
Vitória, ES, Brazil
gabriel.ladislau@edu.ufes.br

2nd Guilherme S. G, Brotto
Department of Informatics
Federal University of Espirito Santo
Vitória, ES, Brazil
guilherme.brotto@edu.ufes.br

3rd Marlon M. Amaral
Department of Informatics
Federal University of Espirito Santo
Vitória, ES, Brazil
marlon.amaral@edu.ufes.br

***Abstract*—Episodic memory enables Large Language Models (LLMs) to retain and retrieve information from past interactions, improving response consistency and personalization. This study explores the implementation and evaluation of episodic memory in LLMs using a question/answer dataset. The proposed methodology involves embedding extracted facts into a memory system and assessing its impact on question-answering performance. Evaluation is conducted by comparing LLM responses with and without episodic memory, using cosine similarity with Sentence-BERT (SBERT). By leveraging episodic memory, LLMs can move beyond static, context-limited interactions and demonstrate improved factual consistency over extended conversations. Preliminary findings aim to demonstrate the advantages of episodic memory in long-term knowledge retention and enhanced reasoning capabilities.**

***Index Terms*—Episodic Memory, Large Language Models, Question/Answer datasets,**

I. Introduction

Large Language Models (LLMs) exhibit remarkable performance in various natural language processing tasks but often lack long-term memory capabilities. Unlike human cognition, where episodic memory allows for learning from past experiences, LLMs traditionally rely on limited context windows that reset after each interaction. This limitation hinders their ability to recall user-specific information or accumulate knowledge over time. To address this, episodic memory mechanisms can be integrated into LLMs, allowing them to store and retrieve relevant information from previous conversations. This study investigates the implementation of episodic memory in LLMs, focusing on its impact on factual recall and contextual understanding.

This research proposes an experimental setup using a Question and Answer (QA) dataset, that contains a wide range of specific knowledge questions paired with answer sets and supporting context passages. Our study simulates episodic memory by extracting factual statements from these context passages and storing them in a structured memory system, by leveraging a vector embedding dataset. The LLM is then tested on related questions under two conditions: without episodic memory, where it relies solely on its pretrained knowledge, and with episodic memory, where it retrieves stored facts before generating a response.

II. Related Works

Memory-augmented LLMs have gained increasing attention in recent research. A comprehensive survey on the memory mechanisms of LLM-based agents by Z. Zhang [1] discusses various strategies for integrating memory into AI systems, categorizing memory architectures and evaluation techniques. This work provides a foundational understanding of memory augmentation techniques.

Human-like episodic memory mechanisms have been explored in EM-LLM [2], which introduces a novel approach to organizing memory using Bayesian surprise and graph-theoretic segmentation. EM-LLM demonstrates superior performance on long-context tasks, outperforming retrieval-based and full-context models. This research highlights the potential benefits of episodic memory in handling extended contexts efficiently.

Evaluating conversational memory is a critical area of study, as demonstrated by research on long-term conversational memory in LLM-based agents [3]. This work introduces LOCOMO, a dataset designed to assess the effectiveness of memory retention over extended dialogues. The findings indicate that existing models struggle with long-term context retention, reinforcing the need for robust memory mechanisms.

Cognitive architectures for language agents, such as CoALA [4], propose a modular framework that integrates memory, structured actions, and decision-making processes. This framework offers a solid theoretical foundation for designing memory-augmented LLMs and serves as a valuable guide for the implementation of episodic memory in our study.

The insights provided by these prior works highlight the critical role of memory in enhancing the performance of LLMs and offer various approaches for its integration. Our study primarily builds upon the CoALA framework [4], leveraging its principles to incorporate episodic memory into LLMs and explore its impact on model performance.

III. Methodology

This section is divided into four subsections, namely: (i) Reflection, where the reflection process of each conversation is explained; (ii) Recall, where for each new conversation topic data from the memory database is queried and used; (iii) Dataset, where the dataset used to evaluate our approach is detailed; (iv) Evaluation, where the evaluation process of the results is detailed.

A. *Reflection*

To establish a memory structure for the LLM, our study developed a reflection process. After each conversation, we

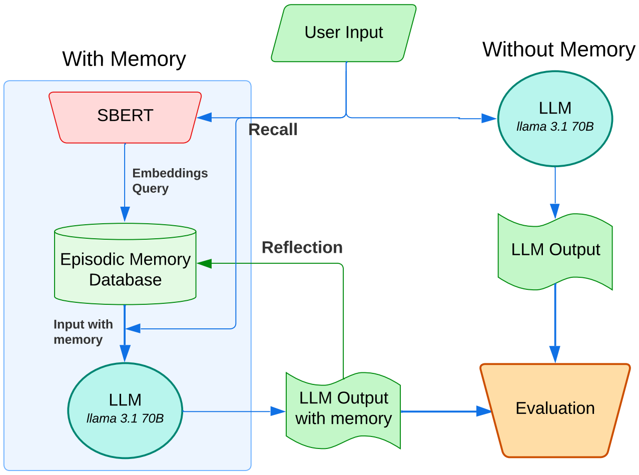

Fig. 1. Evaluation overview.

generate a prompt containing the conversation transcript and pass it to the LLM. The model then extracts key facts, identifies what worked well and what did not, and summarizes the overall interaction. These reflections are stored in an embeddings dataset (ChromaDB), allowing for efficient retrieval during the recall phase when a new conversation begins.

B. Recall

For episodic recall at each new conversation, our study proposes a process where, after receiving the user's input, the memory LLM retrieves relevant information from the episodic memory database. This retrieval is a smart query, ensuring that only useful context is returned. The LLM then generates a response based on both the user's message and the recalled memory, maintaining continuity across interactions. This approach ensures that each response is informed by previous conversations, allowing the LLM to provide more contextually aware and coherent replies, thus guaranteeing the LLM an episodic memory context.

C. Dataset

This study utilizes the NarrativeQA dataset [6], a benchmark for machine comprehension and question answering that evaluates a model's ability to understand and reason about long-form narratives. The dataset consists of stories, screenplays, and books, along with human-generated questions, answers, and corresponding contexts—passages that contain the necessary information to answer the questions. To enhance accuracy, we first feed the memory LLM with these contexts, ensuring it has access to relevant facts before attempting to generate responses. Since both LLMs are stateless, the memoryless model does not require this step and directly processes the questions without prior context.

D. Evaluation

The evaluation process, represented in Figure 1, involves comparing the answers generated by both the memory LLM and the memory-less LLM against the ground-truth answers provided in the NarrativeQA dataset. By doing so, we aim to assess how effectively each model comprehends and responds to questions based on the given contexts. Our evaluation follows two complementary perspectives: quantitative analysis, which focuses on result correctness, and qualitative analysis, which examines answer quality from a broader linguistic perspective.

For the quantitative evaluation, we measure the similarity between the model-generated answers and the actual answers using Sentence Transformer (SBERT) [9]. Specifically, we employ cosine similarity to quantify the closeness of each generated response to the ground-truth answers. By calculating the average similarity score and standard deviation, we gain insight into how consistently each model produces accurate responses. This numerical assessment allows for a direct comparison of performance between the two models.

In addition to the quantitative evaluation, we conduct a qualitative analysis by leveraging ChatGPT-4 Turbo [7] and DeepSeek-V3 [8] models to assess the quality of responses. We provide these models with all generated answers with the same prompt and ask them to analyze and determine which model produced better responses for each question.

The analysis was based on four key criteria: (i) Performance, which evaluates how well each model's answer matches the ground truth and whether the information is presented correctly and fully; (ii) Consistency, which examines whether the model maintains coherence in its responses over time, particularly in extended interactions where memory retention could impact answer stability; (iii) Contextual Relevance, assessing whether the model's response remains relevant to the given context and whether the memory-enabled model demonstrates superior contextual understanding due to its ability to recall past information; (iv) Coherence, which focuses on the reasoning structure behind each answer and how effectively the models handle complex reasoning tasks. These criteria provide a comprehensive framework for evaluating the strengths and limitations of each model, particularly in scenarios that demand high accuracy, adaptability, and contextual awareness.

This human-like evaluation helps capture nuances such as coherence, relevance, and informativeness, offering a more comprehensive understanding of each model's effectiveness beyond strict similarity metrics.

IV. Experiments & Results

To evaluate the performance of episodic and non-episodic LLMs, we randomly selected 50 question-answer pairs from the NarrativeQA dataset. Then, using another LLM, we generated contextual information for each question to populate the memory dataset.

We queried both an LLM with the memory module activated and another without it, using the same set of questions.

The boxplot in Figure 2 illustrates the similarity results for both models. Notably, integrating memory context resulted in a remarkable improvement, with the mean similarity score increasing by over 200%. Specifically, the memory-enhanced

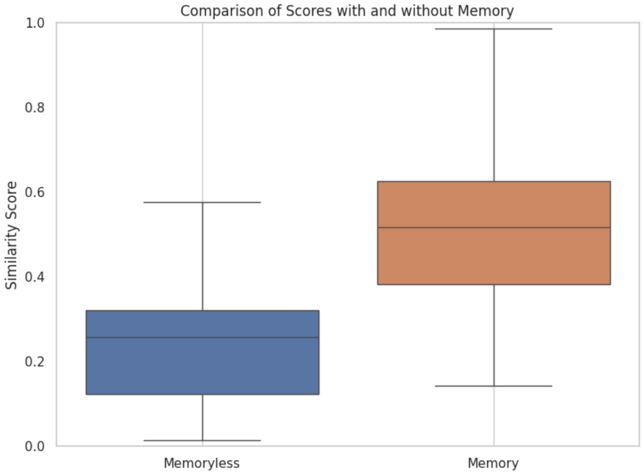

Fig. 2. Evaluation Boxplot with results for both the memoryless and memory LLM approach.

model achieved a mean score of 0.5023 (with a standard deviation of 0.1766), while the memoryless model had a mean score of 0.2510 (with a standard deviation of 0.1483). These findings clearly demonstrate that incorporating episodic memory significantly enhances the model's performance, enabling a stronger alignment with real-world contextual memory.

For the qualitative analysis, we prompted both ChatGPT and DeepSeek models with all the generated answers and the ground truth. The memory LLM is the superior performer across all four criteria. It provides correct, contextually relevant, and logically structured answers. The memoryless model fails in accuracy, relevance, and coherence. It frequently fails to meet the expected standards, often providing irrelevant or incomplete responses.The results show that in every aspect the memory LLM outperformed the memoryless version, giving it a notable boost as seen in Figure2.

V. Conclusion

In conclusion, the implementation of a basic episodic memory system that allows the LLM to store and retrieve relevant information from each prompt leads to a significant improvement in model performance. Both quantitative and qualitative analyses demonstrate that the memory-enabled LLM outperforms the memoryless version in all evaluated aspects.

Moving forward, future work will focus on developing a more sophisticated episodic memory system that not only enhances long-term memory retention but also supports dynamic adaptation, allowing the model to more effectively manage evolving contexts and better generalize to novel situations. Additionally, exploring ways to optimize memory retrieval and reduce the risk of information overload will be crucial to further boosting overall performance of LLMs in general.

References

[1] Z. Zhang, X. Bo, C. Ma, R. Li, X. Chen, Q. Dai, J. Zhu, Z. Dong, and J.-R. Wen, "A Survey on the Memory Mechanism of Large Language Model based Agents," arXiv preprint arXiv:2404.13501, 2024.

[2] Z. Fountas, M. A. Benfeghoul, A. Oomerjee, F. Christopoulou, G. Lampouras, H. Bou-Ammar, and J. Wang, "Human-like Episodic Memory for Infinite Context LLMs," arXiv preprint arXiv:2407.09450, 2024. .

[3] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, "Evaluating Very Long-Term Conversational Memory of LLM Agents," arXiv preprint arXiv:2402.17753, 2024.

[4] T. R. Sumers, S. Yao, K. Narasimhan, and T. L. Griffiths, "Cognitive Architectures for Language Agents," arXiv preprint arXiv:2309.02427, 2024.

[5] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," arXiv preprint arXiv:1705.03551, 2017.

[6] T. Kočisk'y, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette, "The NarrativeQA Reading Comprehension Challenge," Transactions of the Association for Computational Linguistics, vol. 6, pp. 317–328, 2018.

[7] OpenAI, "ChatGPT-4 Turbo: A Large-Scale Conversational AI Model," OpenAI preprint, 2025.[Online]. Available: https://chatgpt.com [Accessed: 22/03/2025].

[8] DeepSeek, "DeepSeek-V3: An Advanced Artificial Intelligence Model for Natural Language Processing and Comprehension", [Online]. Available: https://www.deepseek.com. [Accessed: 22/03/2025].

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," arXiv preprint arXiv:1908.10084, 2019.