Improving Long-Term Conversational Memory in LLMs with a Graph-Based Approach

Chaohui Guo[1] and Michel Klein[1,*]

Department of Computer Science, Vrije Universiteit Amsterdam, Netherlands
c.guo2@vu.nl michel.klein@vu.nl

Keywords: Long-term conversation · Graph-structured memory · Conversation coherence · Large language models.

1 Methodology

Large language models excel at generating coherent text within short contexts but struggle with maintaining long-term conversational coherence due to input token constraints [1]. This limitation often results in fragmented dialogues and loss of important context over extended interactions. To address these issues, we propose a Graph-Based Conversational Memory (GBCM), which uses a graph-structured memory approach to dynamically capture and connect key elements from past dialogues.

In Graph-Based Conversational Memory, nodes represent key dialogue segments or information, while edges capture their relationships, with the structure illustrated in Fig. 1. This structure allows for flexible retrieval and reintegration of previous conversations. A tiered query strategy ensures efficient access to historical information. We first search for the entities from past conversations that are related to the entities in the current query. If the answer is not found within these entities and the paths connecting them, we then expand the search to include the entities one layer beyond these entities. This approach enables the system to retrieve the most pertinent past data as background knowledge, resulting in more coherent and contextually informed responses [2].

We evaluated GBCM in four scenarios: long conversations, long narrative stories, continuous questioning, and multi-topic dialogues. For long conversations, we used the Llama model to rephrase The Diary of Samuel Pepys (1660–1669) into daily conversations between Samuel Pepys and a chatbot, also retaining The Diary of Samuel Pepys as test data to observe whether the Llama model's rephrasing affects the memory system's performance [3]. For narrative stories, we tested with Three-Body Problem [4], and for continuous questioning, we used publicly available medical Q&A data. The multi-topic dialogue combined the previous datasets into a mixed conversation. In this case, the Llama model generated transitions between topics [5].

For each dataset, we searched for 10 relevant questions, each with a unique correct answer, and created complex versions of questions by replacing key entities with referential expressions. For example, "In which laboratory did Wang

Miao first meet the core members of the Trisolaris organization?" became "The scientist who discovered the anomaly in the cosmic microwave background radiation first met the core members of the Three-Body Organization in which laboratory?"

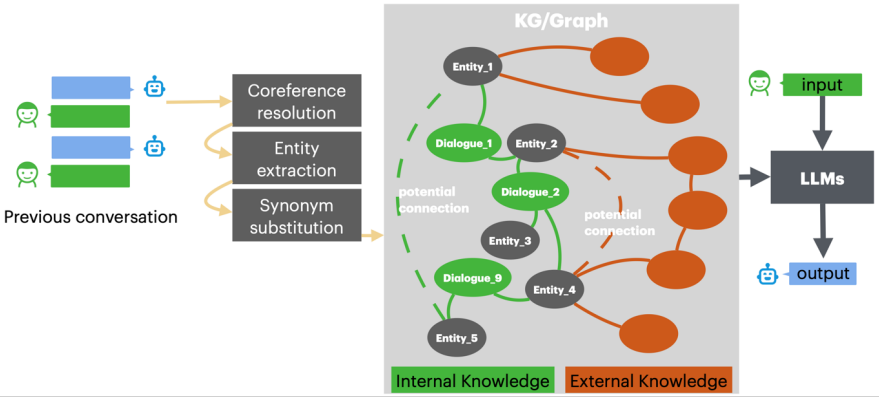

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

2 Experimental Setup and Results

In terms of experimental setup, we tested the narrative storytelling dataset using different methods: Llama 3 8B, Llama 3 8B + Chroma, Llama 3 8B + Sentence Transformer, Llama 3 8B + GraphRAG, and Llama 3 8B + GBCM. Our method achieved the highest accuracy in answering questions. The preliminary experimental results suggest that our GBCM system excels in understanding long texts, improving information retention, and connecting elements from past conversations.

3 Experimental Setup and Results

In future iterations of the system, we plan to refine the tests in all scenarios. In addition, we will adopt existing ontologies to better organize key information from previous conversations. By using a structured, semantic framework, we aim to improve the consistency and interpretability of the stored memory. Furthermore, by integrating external knowledge graphs into the memory system, we can expand the background knowledge available and enrich the conversation context with a broader range of information. This integration will enhance the system's ability to handle complex and diverse dialogues with a deeper contextual understanding.

References

1. McTear, M., & Ashurkina, M.: Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents. (2024)
2. Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., ... & Tang, S.: Graph retrieval-augmented generation: A survey. arXiv preprint arXiv:2408.08921 (2024)
3. Pepys, S.: The Diary of Samuel Pepys: Companion, vol. 10. University of California Press (1970)
4. Liu, C.: The Three-Body Problem, vol. 1. Macmillan (2014)
5. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)