Enhancing Memory and Context: Episodic Memory in LLMs

Ajmal Aksar
Division of Data Science and Cyber Security, Karunya Institute of Technology and Sciences, Coimbaore, Tamil Nadu, India

M. Bhuvaneshwari
Division of Data Science and Cyber Security, Karunya Institute of Technology and Sciences, Coimbaore, Tamil Nadu, India

Abstract

This study presents a novel episodic memory framework designed to overcome the fixed context limitations of large language models (LLMs). The proposed system integrates vector-based semantic search (via ChromaDB) with graph-based relationship analysis (using Neo4j) to store, manage, and decay conversational episodes. By dynamically prioritizing relevant interactions while discarding outdated data, the framework enhances LLMs' ability to maintain long-term contextual continuity. Experimental evaluations indicate that, although the integration introduces a modest inference overhead of approximately 12–15%, it yields significant gains in memory recall—achieving 95% accuracy for straightforward queries and 88% for more complex, context-dependent inquiries. Furthermore, the system demonstrates robust scalability, effectively managing up to 100,000 memory episodes with minimal performance degradation. The modular design also facilitates seamless integration with multiple LLM providers, such as Groq and OpenAI, while user feedback highlights the interface's intuitiveness and operational efficacy. Overall, the findings underscore the potential of episodic memory augmentation to significantly improve response accuracy and contextual awareness in extended, dynamic interactions, thereby broadening the practical applicability of LLMs in realworld, long-duration applications.

Keywords

Large Language Models, Episodic Memory, Context Retention, Memory Decay, Neural Networks.