BEYOND A MILLION TOKENS: BENCHMARKING AND ENHANCING LONG-TERM MEMORY IN LLMS

Mohammad Tavakoli[1], **Alireza Salemi**[2], **Carrie Ye**[1], **Mohamed Abdalla**[1], **Hamed Zamani**[2], **J. Ross Mitchell**[1]
[1]University of Alberta [2]University of Massachusetts Amherst
{tavakol5, cye, mabdall2, jmitche2}@ualberta.ca
{asalemi, zamani}@cs.umass.edu

ABSTRACT

Evaluating the abilities of large language models (LLMs) for tasks that require long-term memory and thus long-context reasoning, for example in conversational settings, is hampered by the existing benchmarks, which often lack narrative coherence, cover narrow domains, and only test simple recall-oriented tasks. This paper introduces a comprehensive solution to these challenges. First, we present a novel framework for automatically generating long (up to 10M tokens), coherent, and topically diverse conversations, accompanied by probing questions targeting a wide range of memory abilities. From this, we construct BEAM, a new benchmark comprising 100 conversations and 2,000 validated questions. Second, to enhance model performance, we propose LIGHT–a framework inspired by human cognition that equips LLMs with three complementary memory systems: a long-term episodic memory, a short-term working memory, and a scratchpad for accumulating salient facts. Our experiments on BEAM reveal that even LLMs with 1M token context windows (with and without retrieval-augmentation) struggle as dialogues lengthen. In contrast, LIGHT consistently improves performance across various models, achieving an average improvement of 3.5%–12.69% over the strongest baselines, depending on the backbone LLM. An ablation study further confirms the contribution of each memory component.

1 INTRODUCTION

Large language models (LLMs) have been deployed across diverse applications, including open-domain conversational agents (Laban et al., 2025; Chen et al., 2025), retrieval-augmented generation (RAG) for open-domain question answering and fact checking (Lewis et al., 2020; Salemi et al., 2025; Salemi & Zamani, 2025; Kim et al., 2024b), long-document and code analysis (Li et al., 2025; Jelodar et al., 2025; Fang et al., 2024), and scientific or legal research (Rueda et al., 2025; Nguyen et al., 2025). Many of these tasks demand models capable of processing long inputs, motivating LLMs such as Gemini (DeepMind, 2025) with input windows of up to 1M tokens. Among these domains, conversational systems present an intuitive and critical need for extended context, as users often engage in protracted, multi-session dialogues that require consistent memory across lengthy interactions (Zhong et al., 2024; Xu et al., 2022; Du et al., 2024; Tan et al., 2025). This highlights the importance of evaluating how well LLMs can reason over and utilize long conversational histories.

While there are many prior efforts on studying and evaluating long-term memory of LLMs (Kim et al., 2024a; Xu et al., 2021; Maharana et al., 2024; Zhong et al., 2024; Xu et al., 2022; Du et al., 2024; Tan et al., 2025), existing benchmarks have fundamental limitations. Most extend conversation length by artificially concatenating short sessions of different users, producing dialogues with abrupt topic shifts and weak narrative coherence. Such a construction artificially simplifies evaluation because distinct segments are easily separable, reducing the need for true long-range reasoning. Furthermore, these datasets typically target narrow domains—often limited to personal-life scenarios—leaving many real-world application areas underrepresented. Finally, they emphasize simple context recall, overlooking other critical memory abilities such as contradiction resolution, recognizing evolving information, and instruction following.