(DeepMind, 2025; Anthropic, 2025; OpenAI, 2025a)) and even 10M (Llama 4; (Meta-AI, 2025)). This growth is driven by advances in efficient attention (sparse, linear, memory-optimized kernels; (Beltagy et al., 2020; Wang et al., 2020; Dao et al., 2022)), improved positional encodings (relative, rotary with scaling, ALiBi; (Dai et al., 2019; Peng et al., 2023b)), long-context training strategies (continued-training, curriculum learning; (Xiong et al., 2023; Ding et al., 2024)), and inference optimizations such as paged attention, KV-cache compression, and distributed attention (Kwon et al., 2023; Zhang et al., 2023; Li et al., 2024; Liu et al., 2023). Such capabilities are especially valuable for applications involving conversational histories, the main focus of our work.

Beyond expanding context windows, models incorporate additional mechanisms for persistent memory. These include recurrence and compression (Transformer-XL, Compressive Transformer; (Dai et al., 2019; Rae et al., 2019)), state-space architectures (RWKV, Mamba, Hyena; (Peng et al., 2023a; Gu & Dao, 2023; Poli et al., 2023)), external memory modules (Memformer, RETRO, RMT; (Wu et al., 2020; Borgeaud et al., 2022; Fan et al., 2024)), context summarization (AutoCompressor; (Chevalier et al., 2023)), and retrieval-augmented generation (REALM, RAG, HippoRAG; (Guu et al., 2020; Lewis et al., 2020; Jimenez Gutierrez et al., 2024)). These approaches complement larger windows by enabling scalable and persistent long-term reasoning.

Existing benchmarks such as DialSim, MSC, LoCoMo, MemoryBank, DuLeMon, PerLTQA, LongMemEval, and MemBench (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) evaluate recall, temporal reasoning, and multi-session reasoning, but typically span narrow domains, exhibit shallow dependencies, and concatenate separate user sessions to simulate long context, reducing realism. Our benchmark instead scales to 10M tokens across diverse topics and introduces new tasks such as contradiction resolution, event ordering, and instruction following, generating coherent, single-user conversations that preserve narrative continuity for a more faithful assessment of long-term conversational memory.

6 Conclusion

This paper addresses the shortcomings of existing benchmarks for evaluating long-term memory in conversational systems. We introduce a scalable framework to generate BEAM, a new benchmark with long, coherent dialogues (up to 10M tokens) and diverse memory probes. To improve LLMs performance, we develop LIGHT, a cognitive-inspired framework combining episodic, working, and scratchpad memories. Our experiments show that while standard LLMs' performance degrades over long contexts, LIGHT provides substantial improvements, boosting memory performance by an average of 3.5%-12.69%. By offering a more robust evaluation and an effective memory enhancement technique, this work helps the development of more reliable long-context conversational systems.

Acknowledgments

Ross Mitchell is the Alberta Health Services Chair in Artificial Intelligence in Health and is supported by CIFAR, the University Hospital Foundation, Amii, and the Canada Foundation for Innovation. Mohamed Abdalla is supported by a CIFAR AI Chair and an Amii grant. This research is supported by the Canadian Institutes of Health Research (FRF 196047). Carrie Ye is supported by the CRAF (CIORA)–Arthritis Society Canada Clinician Investigator Award (CI-24-0013).