A Detailed Related Work

Long-Context Large Language Models. The context window of LLMs has expanded from 512–2,048 tokens in early models (GPT-1/2/3, BERT, T5; (Radford et al., 2018; 2019; Brown et al., 2020; Devlin et al., 2019; Raffel et al., 2020)) to 128K–1M tokens in recent systems (Claude-3, GPT-4-Turbo, Gemini 1.5 Pro, Gemini 2.0 Flash, Claude-4, GPT-4.1; (Anthropic, 2024; Achiam et al., 2023; Team et al., 2024; DeepMind, 2025; Anthropic, 2025; OpenAI, 2025a)), with some reaching 10M tokens (Llama 4 Scout; (Meta-AI, 2025)). This growth has been enabled by innovations that address the quadratic cost of self-attention, including sparse mechanisms (Longformer, BigBird; (Beltagy et al., 2020; Zaheer et al., 2020)), linear approximations (Linformer, Performer; (Wang et al., 2020; Choromanski et al., 2020)) and memory-efficient kernels (FlashAttention; (Dao et al., 2022)). Advances in positional encoding, such as relative encodings (Transformer-XL; (Dai et al., 2019)), rotary embeddings (RoPE; (Su et al., 2024)) with scaling methods (YaRN, NTK; (Peng et al., 2023b)), and linear biases (ALiBi; (Press et al., 2021)), have extended usable context lengths. Training strategies like continued pre-training and curriculum learning (e.g., LLaMA-2-Long (Xiong et al., 2023), LongRoPE (Ding et al., 2024)) further expand capabilities, while inference optimizations such as PagedAttention (Kwon et al., 2023), KV-cache compression (H2O, SnapKV; (Zhang et al., 2023; Li et al., 2024)) and distributed approaches (Ring Attention; (Liu et al., 2023)) enable practical deployment at scale.

Long-Term Memory Methods. Researchers have developed approaches to enhance long-term memory beyond simply extending context windows. Architectural modifications include Transformer-XL (Dai et al., 2019), which introduced segment-level recurrence, and Compressive Transformer (Rae et al., 2019), which stored both recent states and compressed older information. State-space models such as RWKV (Peng et al., 2023a), Mamba (Gu & Dao, 2023), and Hyena (Poli et al., 2023) replace attention with recurrent dynamics, allowing linear scaling and theoretically unbounded memory. Memory-augmented transformers such as Memformer (Wu et al., 2020), RETRO (Borgeaud et al., 2022) and RMT (Fan et al., 2024) add external memory slots for explicit storage and recall. Context compression offers an orthogonal strategy by summarizing past information rather than storing it verbatim, as in AutoCompressor (Chevalier et al., 2023), which learns compact, information-preserving representations to reduce token usage. Retrieval-augmented generation (RAG) scales further by maintaining external knowledge stores: REALM (Guu et al., 2020) and RAG (Lewis et al., 2020) pioneered dense retrieval, RETRO (Borgeaud et al., 2022) integrated retrieval into transformers, and HippoRAG (Jimenez Gutierrez et al., 2024) incorporated structured knowledge graphs.

Building on these foundations, we propose a novel retrieval-augmented method that shows substantial improvements over baselines in long-memory evaluation.

Long-Term Memory Benchmarks. Several benchmarks have emerged to evaluate long-term memory capabilities in LLMs. DialSim (Kim et al., 2024a) derives evaluation data from multiparty television scripts, producing dialogues extending to 350K tokens with naturalistic patterns but limited topical diversity. MSC (Xu et al., 2021) introduces multisession human-assistant conversations testing memory across session boundaries, though with brief sessions and shallow dependencies. LoCoMo (Maharana et al., 2024) presents 50 conversations averaging 9K tokens in 35 sessions, while MemoryBank (Zhong et al., 2024) provides 300 sessions with 194 probing questions evaluating recall and temporal reasoning. DuLeMon (Xu et al., 2022) focuses on dialogue-level memory and forgetting curves, PerLTQA (Du et al., 2024) targets memory classification and retrieval, and LongMemEval (Wu et al., 2024) constructs multisession evaluations with 500 questions testing information extraction and temporal reasoning. More recently, MemBench (Tan et al., 2025) evaluates the memory of LLM-based agents by assessing their performance on information extraction, multi-hop reasoning, knowledge updating, preference following, and temporal reasoning.

As summarized in Table 2, the existing benchmarks are largely based on concatenated short sessions with limited coherence, narrow personal and casual domains, and few memory abilities. They also lack realistic bidirectional interactivity. In contrast, our benchmark spans diverse domains, scales up to 10M tokens, and introduces three additional dimensions—contradiction resolution, event ordering, and instruction following—yielding a more comprehensive framework for evaluating long-term memory in conversational systems.