



Figure 1: Overview of data generation.

To address these limitations, this paper presents a framework for automatically generating long coherent conversations between a user and an AI assistant—scaling up to 10M tokens on diverse domains—with a set of probing questions designed to evaluate diverse memory abilities of any LLM on the generated dialogues. An overview of the data generation framework is shown in Figure 1. This framework begins by defining a high-level conversation plan—a narrative for a particular domain and a simulated user with generated attributes—that outlines the overall flow of the dialogue. This plan is recursively decomposed into finer sub-plans that specify the storyline and its progression. From these sub-plans we generate chronologically ordered user turns, which are then expanded with corresponding assistant responses. To increase realism, the system injects follow-up questions and clarifications from both sides. Finally, we automatically create a set of probing questions that target ten distinct memory dimensions, with a focus on complicated and multi-hop reasoning, which are then validated by human annotators to ensure high quality. Using this pipeline, we construct the BEAM dataset: 100 diverse conversations ranging from 100 K to 10 M tokens each, accompanied by 2000 probing questions to evaluate the memory capabilities of LLMs.

To improve LLM performance on probing questions, we introduce the LIGHT framework (Figure 2), which is applicable to both open-source and proprietary LLMs, inspired by research in human cognitive science and human’s memorization and recall process (Sridhar et al., 2023; Binder & Desai, 2011). This framework integrates three complementary memories: (1) episodic memory, a long-term index of the full conversation used for retrieval; (2) working memory, capturing the most recent user–assistant turns; and (3) a scratchpad, where after each turn the model reasons over the dialogue and records salient facts for future use. At inference, the LLM draws jointly on retrieved episodic content, the working memory, and the accumulated scratchpad to generate accurate answers.

To evaluate LLM memory capabilities and the effectiveness of our method, we conduct experiments on the constructed dataset, BEAM, using both open-source and proprietary models. Results show that even LLMs with long context windows perform substantially worse as conversation length increases. Our method improves the LLM’s performance in answering the probing questions by 3.5%–12.69% on average over the best-performing baseline, depending on the backbone model and conversation length. An ablation study further reveals the contribution of each LIGHT component on the performance. To support future work, we release all code, data, and evaluation scripts.¹

2 BEAM: BENCHMARKING MEMORY CAPABILITIES OF LLMs

2.1 PROBLEM FORMULATION

Let $\mathcal{D} = \{T_i\}_{i=1}^{|\mathcal{D}|}$ denote a collection of $|\mathcal{D}|$ conversations between users and a conversational agent π . Each conversation is represented as $\mathcal{T} = \{t_i\}_{i=1}^{|\mathcal{T}|}$, where $t_i \in \mathcal{T}$ corresponds to the i^{th} utterance (turn) in the dialogue. The objective of this work is to systematically evaluate a predefined set of memory abilities \mathcal{M} exhibited by π across conversations. For each memory ability $m \in \mathcal{M}$, we construct a probing dataset of size N , denoted as $\mathcal{Q}_m = \{(x_i, y_i)\}_{i=1}^N$, where x_i is a probing question and y_i is the corresponding ground-truth answer set. Each probing question $(x, y) \in \mathcal{Q}_m$

¹ Available at: <https://github.com/mohammadtavakoli78/BEAM>