is appended as the $(|\mathcal{T}|+1)^{\text{th}}$ turn in the dialogue, and the system generates a response $\hat{y} = \pi(x; \mathcal{T})$ based on the conversation. The generated response is then evaluated using an ability-specific scoring function μ_m, producing a performance score $s = \mu_m(x, y, \hat{y})$. The goal of this work is to quantify the performance of conversational systems on each memory ability in $\mathcal{M}$.

2.2 BENCHMARK CREATION

Our goal is to evaluate how well LLMs can answer questions that depend on long-term conversational memory. We measure performance across ten complementary abilities, seven drawn from prior benchmarks and three newly introduced here—*Instruction Following*, *Event Ordering*, and *Contradiction Resolution* (see Table 2 in Appendix B.1). *Abstention* evaluates whether a model withholds answers when evidence is missing. *Contradiction Resolution* tests the capacity to detect and reconcile inconsistent statements across widely separated turns, maintaining global coherence. *Event Ordering* assesses whether a model can recognize and reconstruct the sequence of evolving information in the dialogue. *Information Extraction* measures recall of entities and factual details in long histories. *Instruction Following* examines sustained adherence to user-specified constraints over long contexts. *Information Update* evaluates revising stored facts as new ones appear. *Multi-hop Reasoning* probes inference that integrates evidence across multiple, non-adjacent dialogue segments. *Preference Following* captures personalized responses that adapt to evolving preferences. *Summarization* assesses the ability to abstract and compress dialogue content, while *Temporal Reasoning* tests reasoning about explicit and implicit time relations. Together, these abilities evaluate a system's capacity to maintain, update, and manipulate information throughout extended conversations (see Appendix B.6 for examples of each ability). Given these abilities and the formulation in Section 2.1, the benchmark requires three components: 1) a user–assistant conversation, 2) probing questions targeting key memory abilities, and 3) an evaluation methodology to assess the model's responses. The overall statistics of the constructed benchmark are summarized in Table 3 in Appendix B.1. The rest of this section details the process used to construct these components.

Overview: The overview of our framework for creating conversations, probing questions, and the evaluation strategy is illustrated in Figure 1. The process begins by generating a simulated conversation between a user and an assistant. Structured conversation plans are first produced to guide the flow of the synthetic interactions. Each plan specifies sufficient information to generate both user and assistant turns, ensuring a coherent and natural conversational trajectory. While a typical exchange consists of a user question followed by an assistant response, realistic dialogues often involve follow-ups for clarification, elaboration, or related subtopics. To capture this, we incorporate two interaction-control modules. The question-detection module identifies whether an assistant response includes a query that requires a user reply; if triggered, the system generates the corresponding user response. The follow-up detection module determines when the user would naturally pose a clarifying or elaborative question; if triggered, it produces an additional user query for the assistant. Together, these mechanisms produce conversations that exhibit interactive, bidirectional behavior beyond simple turn-taking. After the conversation is generated, an automated procedure constructs a candidate set of probing questions, each tailored to the specific memory abilities in the benchmark. These candidates are then reviewed by a human evaluator, who selects valid questions and formulates the associated evaluation rubrics used for subsequent benchmarking. A case study and an example of the different generated components of a conversation is provided in Appendix E.

2.2.1 CONVERSATION PLAN GENERATION

A *conversation plan* serves as the scaffold for each dialogue, providing a coherent storyline that unfolds chronologically. Each plan is generated using an LLM based on seed information, including: the conversation *domain*; a *title and theme*; *subtopics* outlining specific topics; a set of *narratives* defining evolving aspects (e.g., career progression, goals); a *user profile* with attributes such as name, age, gender, location, profession, and personality traits sampled from the Myers–Briggs Type Indicator (MBTI); a *relationship graph* linking the user to family, friends, and acquaintances, constrained for realism (e.g., age gaps); and an explicit *timeline* specifying the span of the conversation. To generate candidate titles and themes, human annotators specify target domains, then GPT-4.1 (OpenAI, 2025a) generates candidate titles, themes, and subtopics using Listing 22. Human reviewers refine outputs for topical diversity. For each conversation, we generate 15-20 narratives using the open-source LLaMA-3.3 70B model (AI, 2024) with the prompt in Listing 23 (Appendix G).