Table 2: Comparison of our benchmark with existing long-term memory benchmarks. Memory abilities: IE = Information Extraction, MR = Multi-hop Reasoning, KU = Knowledge Update, TR = Temporal Reasoning, ABS = Abstention, CR = Contradiction Resolution, EO = Event Ordering, IF = Instruction Following, PF = Preference Following, SUM = Summarization.

Benchmark	**Domain**	**Chat Length**	**Memory Abilities**									
			IE	MR	KU	TR	ABS	CR	EO	IF	PF	SUM
MSC (Xu et al., 2021)	Casual	∼1K	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
DuLeMon (Xu et al., 2022)	Casual	∼1K	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
MemoryBank (Zhong et al., 2024)	Personal life	∼5K	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗
PerLTQA (Du et al., 2024)	Personal life	N/A	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
LoCoMo (Maharana et al., 2024)	Personal life	∼10K	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓
DialSim (Kim et al., 2024a)	TV/Film scripts	∼350K	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗
LongMemEval (Wu et al., 2024)	Personal life	115K, 1M	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗
MemBench (Tan et al., 2025)	Personal life	∼100K	✓	✓	✓	✓	✗	✗	✗	✗	✓	✗
BEAM (This work)	**Multi-domain:** Coding, Math, Health, Finance, Personal life, ...	**128K, 500K, 1M, 10M**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

B Benchmark Design

B.1 Dataset Statistics

Table 3 summarizes the statistics of the generated dataset, including averages of user messages, assistant messages, assistant and user follow-up questions, and dialogue turns across different chat sizes.

Table 3: Statistics of the dataset. Reported values are averages per chat in each chat size. # User Messages and # Assistant Messages denote the average number of utterances from the user and assistant, respectively. # Answer Assistant Questions is the number of times the assistant posed a question that the user answered. # Followup Questions is the number of follow-up questions asked by the user. # Turns refers to the total number of dialogue turns.

Chat Size	**# User Messages**	**# Assistant Messages**	**# Answer Assistant Questions**	**# Followup Questions**	**# turns**
128K	144	144	27	216	107
500K	544	544	79	51	416
1M	1067	1067	105	120	842
10M	10435	10435	1151	1528	7757

B.2 Benchmark Quality Evaluation

To evaluate the quality of the generated conversations, we conducted human assessment across all conversations. Two annotators rated each conversation on three dimensions using a 5-point Likert scale (1 = lowest, 5 = highest): *Coherence and Flow*, *Dialogue Realism*, and *Complexity and Depth*.

- **Coherence and Flow**: Conversation continuity (each turn follows naturally from the previous one), smooth transitions across topics and responses, and thread consistency without abrupt or jarring shifts.
- **Dialogue Realism**: Naturalness of user queries (messages sound authentic), realistic progression of topics over time, human-like interactions (appropriate clarifications, follow-ups, etc.), and believability of scenarios.
- **Complexity and Depth**: Handling of multi-layered, interconnected topics, progressive increase in difficulty, and demonstration of domain expertise when required.

The aggregated results are reported in Table 4.