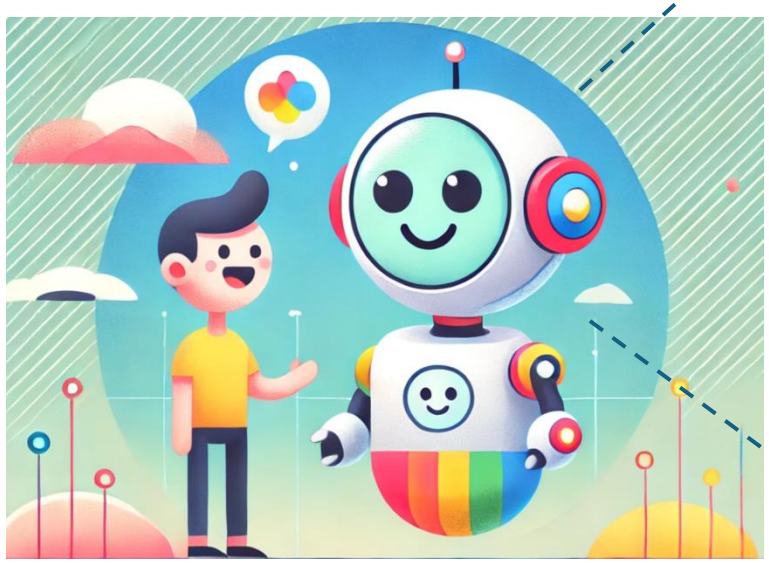
LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory

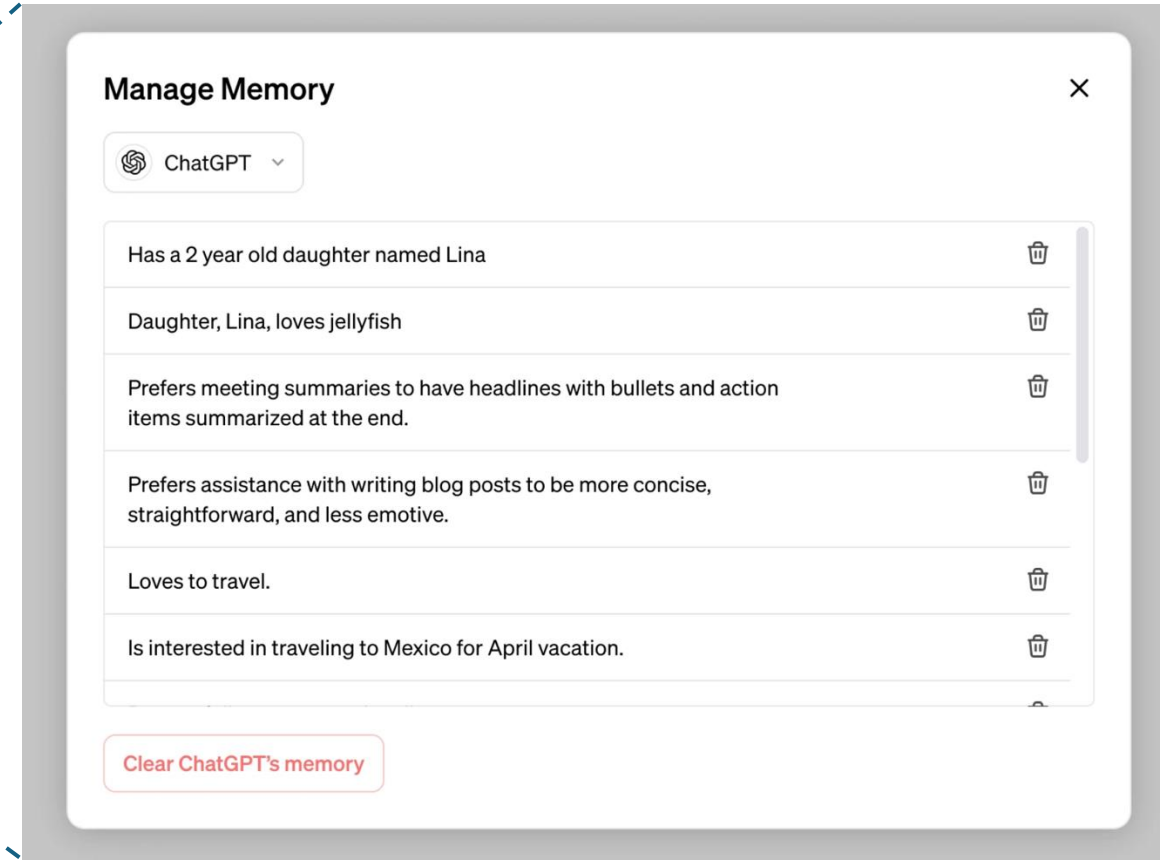
Di Wu[1], Hongwei Wang[3], Wenhao Yu[3], Yuwei Zhang[2],
Kai-Wei Chang[1], Dong Yu[3]
[1]UCLA, [2]UC San Diego, [3]Tencent AI Lab




Long-Term Memory

- **Long-term memory** is a crucial ability of conversational agents.
  - Personal Information
  - Preferences
  - Custom Instructions



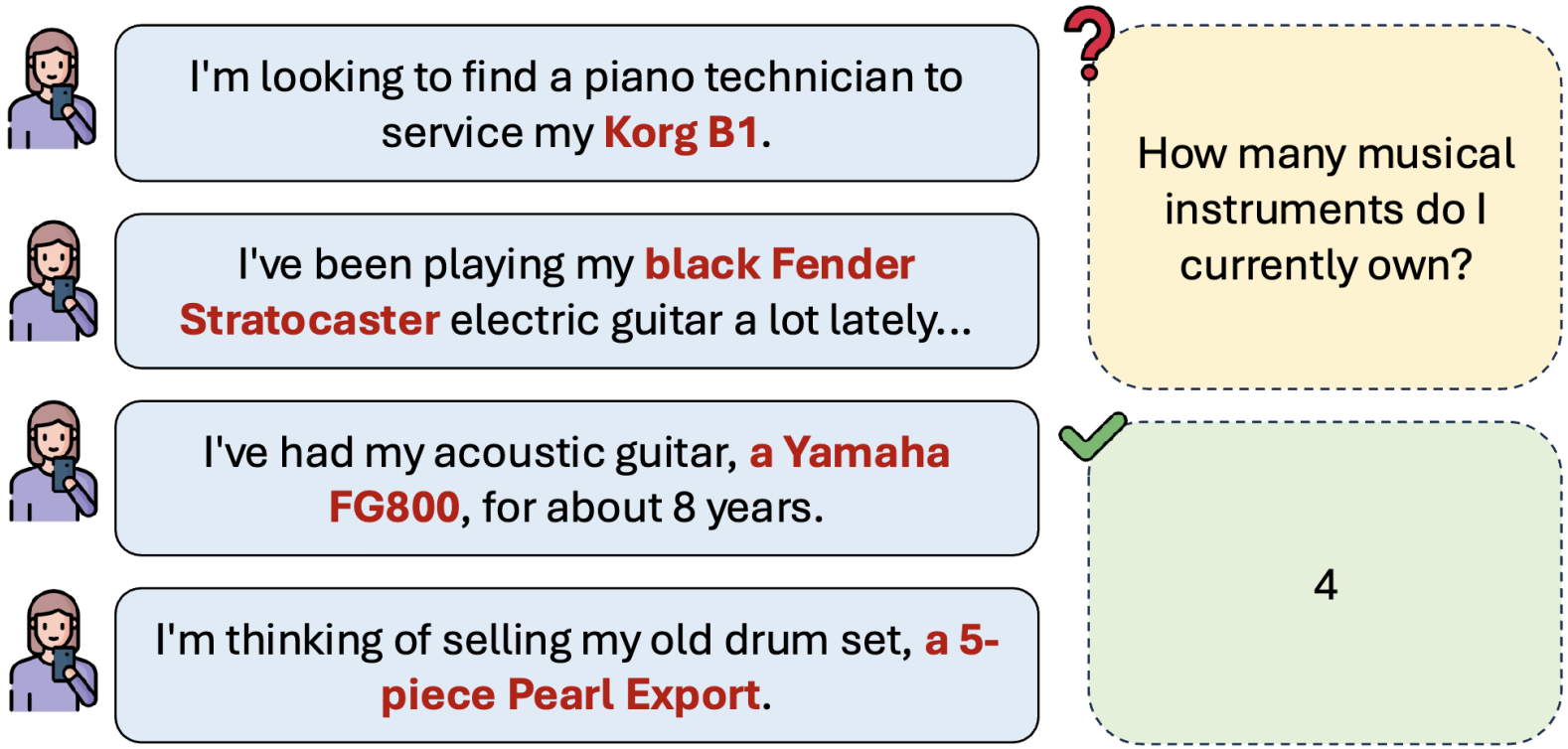


LongMemEval

- LongMemEval is a challenging benchmark for long-term memory of chat assistants.
- Five core abilities
  - Information Exaction
  - Multi-Session Reasoning
  - Knowledge Updates
  - Temporal Reasoning
  - Abstention

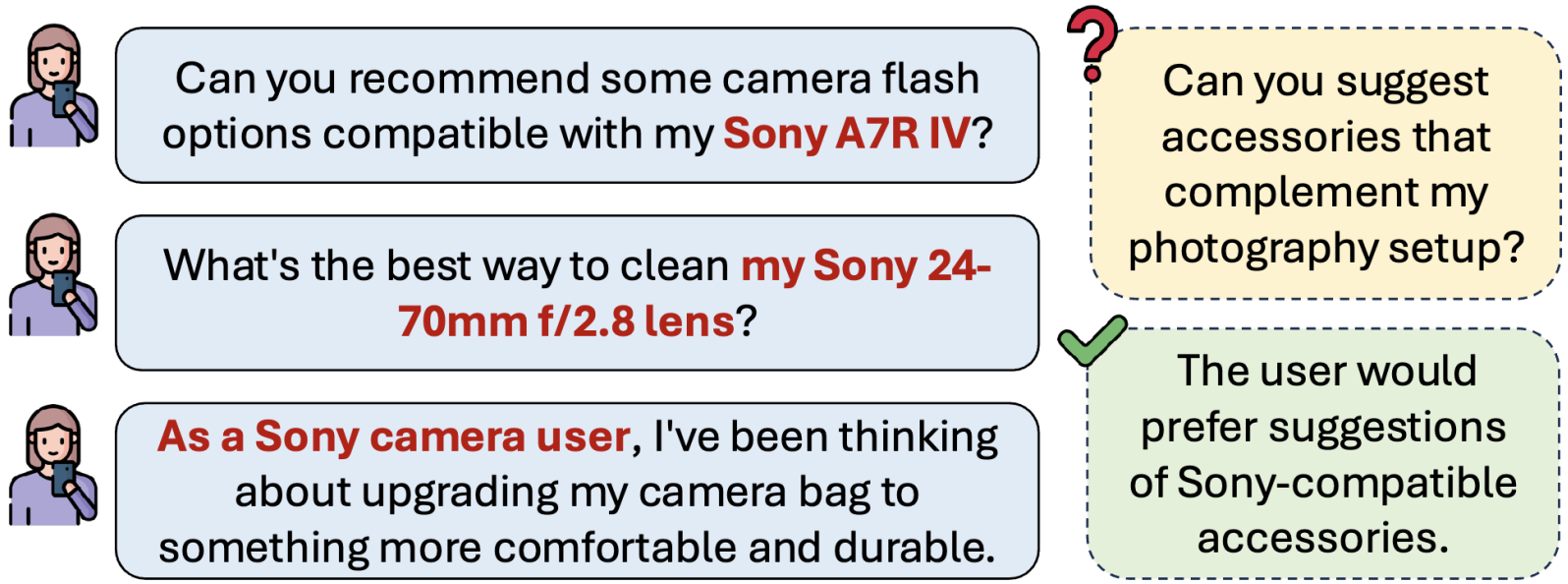
LongMemEval: Highlighted Features

- LongMemEval tests **aggregation over multiple sessions**.



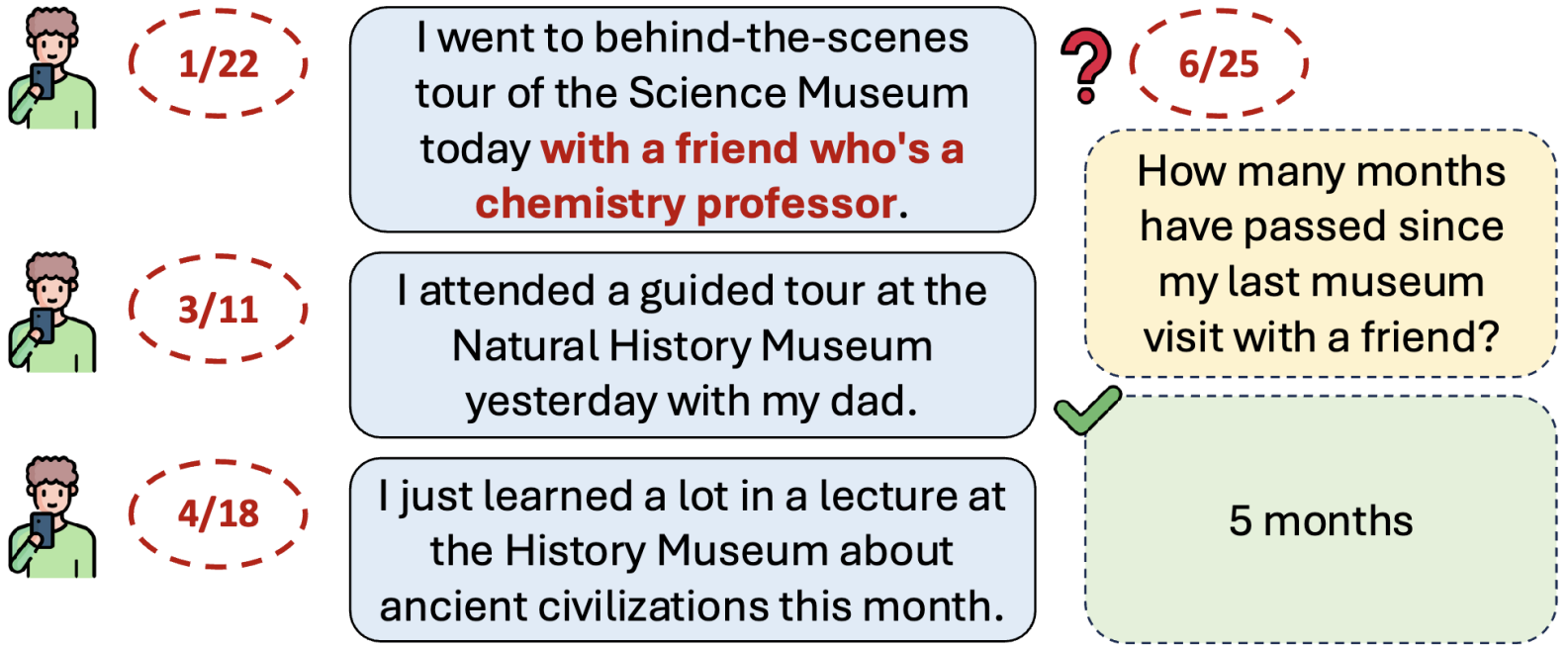
LongMemEval: Highlighted Features

- LongMemEval tests history understanding via **implicit preference**.

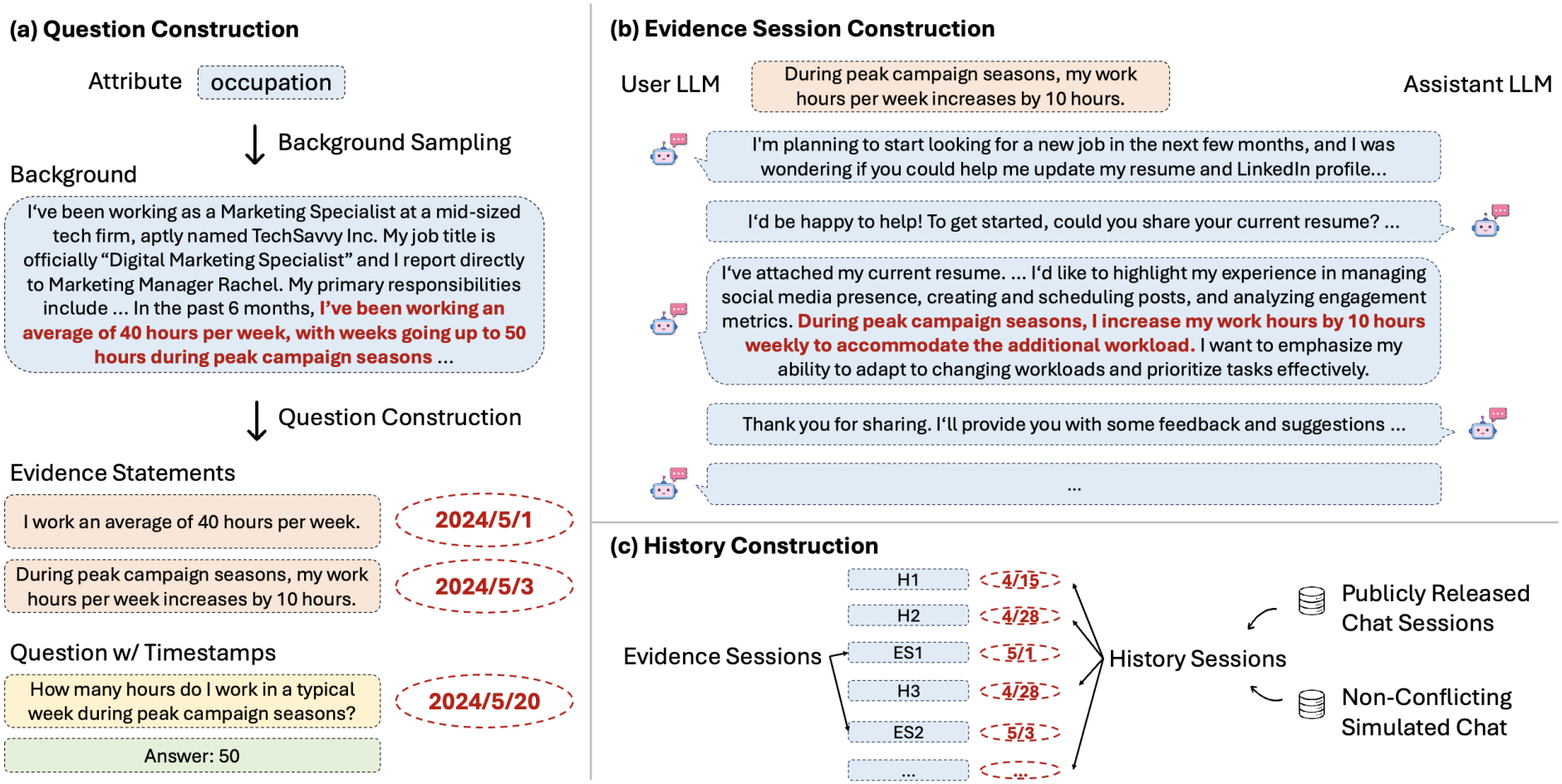


LongMemEval: Highlighted Features

- LongMemEval tests reasoning over **timestamp metadata**.



Freely-Extensible Chat History



LongMemEval is Challenging

Performance drops moving from offline to online memory

System	LLM	Accuracy
Offline Reading	GPT-4o	0.9184
ChatGPT	GPT-4o	0.5773
	GPT-4o-mini	0.7113
Coze	GPT-4o	0.3299
	GPT-3.5-turbo	0.2474

(a) Commercial memory-augmented chat assistants exhibit weak performance on LONGMEMEVAL. The accuracy of ChatGPT and Coze degrades by a large amount compared to directly reading the context ("Offline Reading") with the same LLM. Specifically, ChatGPT and Coze instantiated with GPT-4o exhibits 37% and 64% performance drop, respectively.

Model	Size	Oracle	S	% Drop
No Chain-of-Note				
GPT-4o	-	0.870	0.606	30.3%↓
Llama 3.1 Instruct	70B	0.744	0.334	55.1%↓
Llama 3.1 Instruct	8B	0.710	0.454	36.1%↓
Phi-3 128k Instruct	14B	0.702	0.380	45.9%↓
Phi-3.5 Mini Instruct	4B	0.660	0.342	48.1%↓
With Chain-of-Note				
GPT-4o	-	0.924	0.640	30.7%↓
Llama 3.1 Instruct	70B	0.848	0.286	66.3%↓
Llama 3.1 Instruct	8B	0.710	0.420	40.8%↓
Phi-3 128k Instruct	14B	0.722	0.344	52.4%↓
Phi-3.5 Mini Instruct	4B	0.652	0.324	50.3%↓

(b) Long-context LLMs exhibit large QA performance drops on LONGMEMEVAL$_S$ (column "S"), compared to the accuracy of answering the questions based on only the evidence sessions (column "Oracle").

Figure 3: Pilot study of (a) commercial systems and (b) long-context LLMs on LONGMEMEVAL.

LongMemEval is Challenging

Performance drops moving from offline to online memory

System	LLM	Accuracy
Offline Reading	GPT-4o	0.9184
ChatGPT	GPT-4o	0.5773
	GPT-4o-mini	0.7113
Coze	GPT-4o	0.3299
	GPT-3.5-turbo	0.2474

(a) Commercial memory-augmented chat assistants exhibit weak performance on LONGMEMEVAL. The accuracy of ChatGPT and Coze degrades by a large amount compared to directly reading the context ("Offline Reading") with the same LLM. Specifically, ChatGPT and Coze instantiated with GPT-4o exhibits 37% and 64% performance drop, respectively.

Model	Size	Oracle	S	% Drop
No Chain-of-Note				
GPT-4o	-	0.870	0.606	30.3%↓
Llama 3.1 Instruct	70B	0.744	0.334	55.1%↓
Llama 3.1 Instruct	8B	0.710	0.454	36.1%↓
Phi-3 128k Instruct	14B	0.702	0.380	45.9%↓
Phi-3.5 Mini Instruct	4B	0.660	0.342	48.1%↓
With Chain-of-Note				
GPT-4o	-	0.924	0.640	30.7%↓
Llama 3.1 Instruct	70B	0.848	0.286	66.3%↓
Llama 3.1 Instruct	8B	0.710	0.420	40.8%↓
Phi-3 128k Instruct	14B	0.722	0.344	52.4%↓
Phi-3.5 Mini Instruct	4B	0.652	0.324	50.3%↓

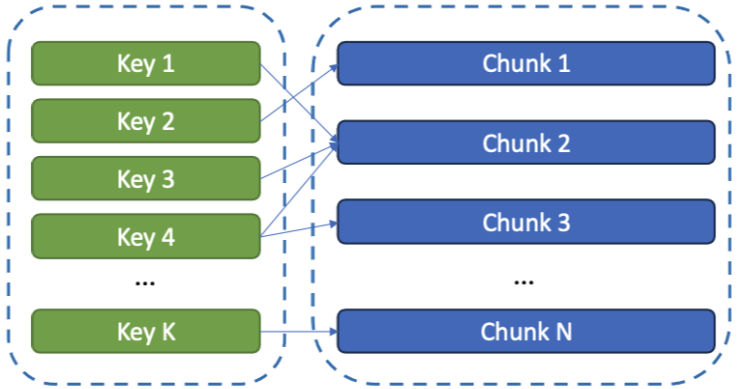
(b) Long-context LLMs exhibit large QA performance drops on LONGMEMEVAL$_S$ (column "S"), compared to the accuracy of answering the questions based on only the evidence sessions (column "Oracle").

Performance drops moving from oracle to short history

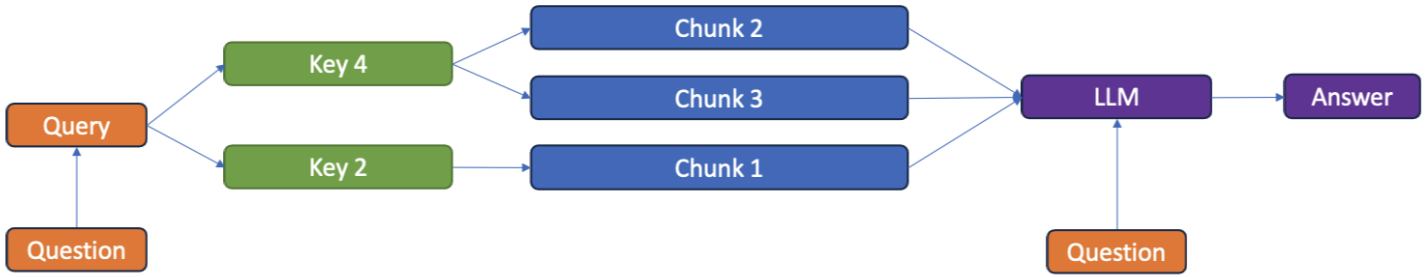
Figure 3: Pilot study of (a) commercial systems and (b) long-context LLMs on LONGMEMEVAL.

Memory Assistants: a Unified View

- With LongMemEval, we study the design choices of memory-augmented assistants in three abstract stages.



(1) Indexing



(2) Retrieval

(3) Reading

Empirical Insights

- **Round** is the more optimal value granularity for history storage.
- Extracting **user facts** for indexing improves both memory recall and downstream question answering.
- **Time-aware indexing and retrieval** improve temporal reasoning.
- Choosing a good memory **reading strategy** is important:
  - JSON formatting
  - Chain-of-Note

Tencent AI Lab

Thank you for listening!

Paper

Repo