arxiv.org/abs/2507.05257

Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions

MemoryAgentBench: 专门测试Agent的memory能力

Yuanzhe Hu[1]*, **Yu Wang**[1]*, **Julian McAuley**[1]
[1]UC San Diego
[1]{yuh127,yuw164,jmcauley}@ucsd.edu

Datasets Source Code

简介

本文提出MemoryAgentBench专门用于评测Memory Agent在记忆方面的能力，MemoryAgentBench从四个维度进行评测：准确检索（AR）、测试时学习（TTL）、长程理解（LRU）和冲突解决（CR）。实验中评测了Long-Context Agent、RAG Agent和Agentic Memory Agent这三大类不同的记忆机制表现，具体来说通过“连续注入chunks + prompt指导保存记忆 + 多问题查询”的实验方法统一评测几类Memory Agent。

背景

Memory是Agent的重要能力，目前对于Agent的能力评测主要聚焦推理、规划和工具调用等方面，忽略了Memory评测，因此，本文设计了专门评测Agent Memory能力的基准。本文中将具有Memory能力的Agent称为Memory Agent。

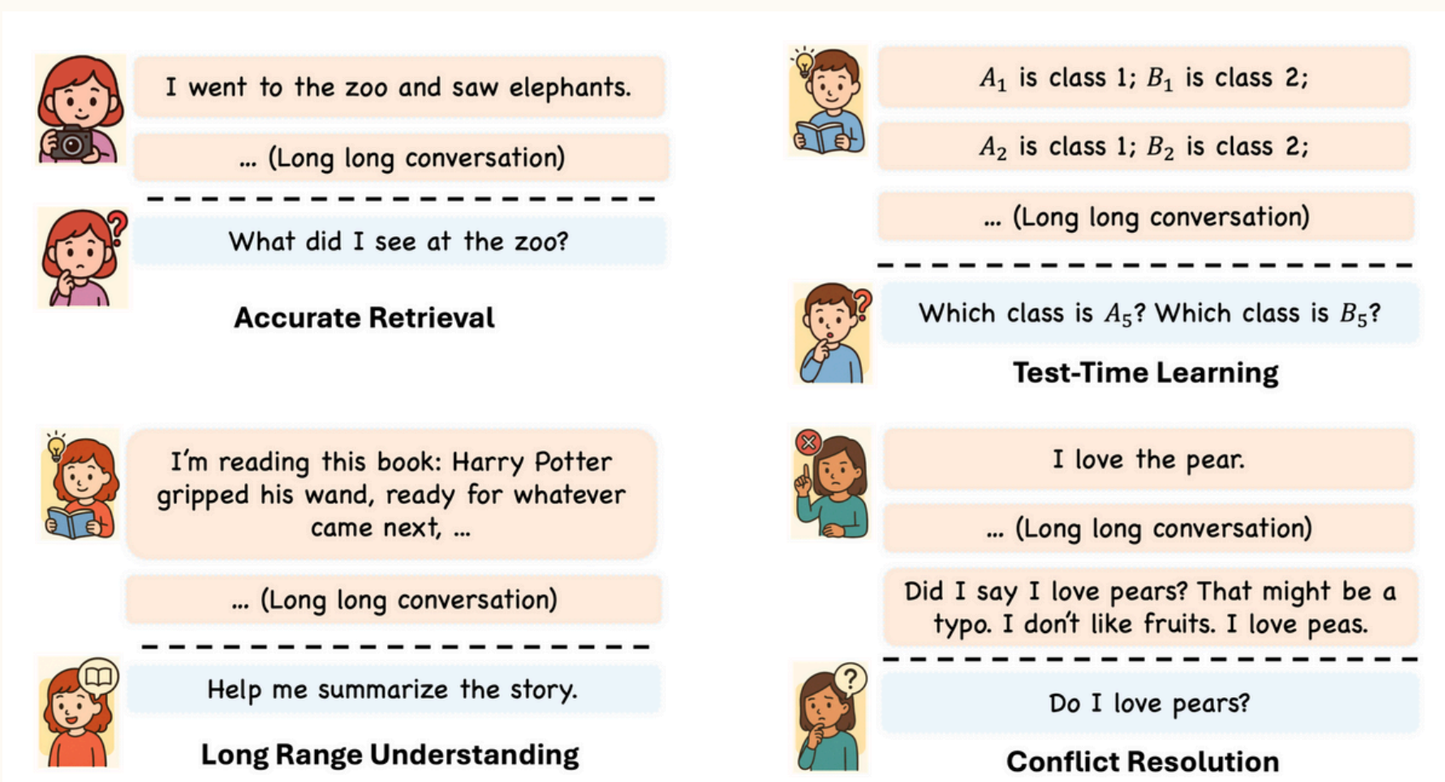
Memory Agent类型

- Long-Context Agents: 用最近的chunk填满context sequence，简单直接
- Simple RAG Agents: 保存所有chunk文本，使用时用BM25检索
- Embedding-based RAG Agents: 每个chunk向量化后保存，使用时query先向量化再检索
- Structure-Augmented RAG Agents: 对所有chunk结构化处理再保存，比如典型的GraphRAG
- Agentic Memory Agents: Self-RAG和MemGPT

Memory Agent的记忆能力评测

评测Memory Agent的四个记忆能力维度：

- Accurate Retrieval (AR): Memory Agent能够从多轮增量积累的对话记录中，准确找到回答当前问题所需要的信息
- Test-Time Learning (TTL) Memory Agent是否能在 不tuning的情况下，通过与用户的对话交互动态学习新知识
- Long-Range Understanding (LRU): Memory Agent能够在多轮历史交互中，不仅仅回忆单点信息，而是形成对整体内容、主题和脉络的完整理解，以便回答需要综合整体上下文的高阶问题
- Conflict Resolution (CR): 更新记忆



实验

Agent Type	AR RULER-QA	AR NIAH-MQ	AR ∞Bench-QA	AR LME(S*)	AR EventQA	TTL MCC	TTL Recom	LRU ∞Bench-Sum	CR FactCon-SH	CR FactCon-MH
Long-Context Agents										
GPT-4o	61.5	25.0	**55.4**	32.0	77.2	87.6	12.3	32.2	**60.0**	**5.0**
GPT-4o-mini	53.5	22.8	44.9	30.7	59.0	82.4	15.1	28.9	45.0	5.0
GPT-4.1-mini	**74.5**	**94.8**	45.8	**55.7**	**82.6**	75.6	16.7	41.9	36.0	5.0
Gemini-2.0-Flash	73.0	83.8	53.2	47.0	67.2	84.0	8.7	23.9	30.0	3.0
Claude-3.7-Sonnet	65.0	38.0	50.6	34.0	74.6	**89.4**	**18.3**	**52.5**	43.0	2.0
GPT-4o-mini	53.5	22.8	44.9	30.7	59.0	82.0	15.1	28.9	45.0	5.0
Simple RAG Agents										
BM25	61.0	**100.0**	45.6	45.3	**74.6**	**75.4**	13.6	20.9	**56.0**	3.0
Embedding RAG Agents										
Contriever	26.5	2.5	38.1	15.7	66.8	70.6	15.2	21.2	18.0	**7.0**
Text-Embed-3-Small	52.0	7.2	44.4	48.3	63.0	70.0	15.3	**25.7**	28.0	3.0
Text-Embed-3-Large	49.0	19.5	50.1	52.3	70.0	72.4	**16.2**	21.6	28.0	4.0
NV-Embed-v2	**83.0**	73.5	**51.4**	**55.0**	72.8	69.4	13.5	20.7	55.0	6.0
Structure-Augmented RAG Agents										
RAPTOR	33.5	15.8	31.3	34.3	45.8	59.4	12.3	13.4	14.0	1.0
GraphRAG	47.0	38.3	35.8	35.0	34.4	39.8	9.8	0.4	14.0	2.0
HippoRAG-v2	71.0	67.5	45.7	50.7	67.6	61.4	10.2	14.6	54.0	5.0
Mem0	28.0	4.8	22.4	36.0	37.5	3.4	10.0	0.8	18.0	2.0
Cognee	33.5	4.0	19.7	29.3	26.8	35.4	10.1	2.3	28.0	3.0
Agentic Memory Agents										
Self-RAG	38.5	8.0	28.5	25.7	31.8	11.6	12.8	0.9	19.0	3.0
MemGPT	39.5	8.8	20.8	32.0	26.2	67.6	14.0	2.5	28.0	3.0

思考

实验结果说实话看点很多啊，首先所有的agent都用GPT-4o-mini，可以看到，BM25简直乱杀啊？long-context的方式也好的离谱？GraphRAG比BM25差那么多？向量RAG也一般，mem0和memgpt无语😒
我在想，先对数据集切分chunk，然后通过prompt template让agent强行记忆，再出问题考试的评测方式，是否偏向检索模型呢？

@机器爱学习