AMA-Bench: Evaluating Long-Horizon Memory for Agentic Applications

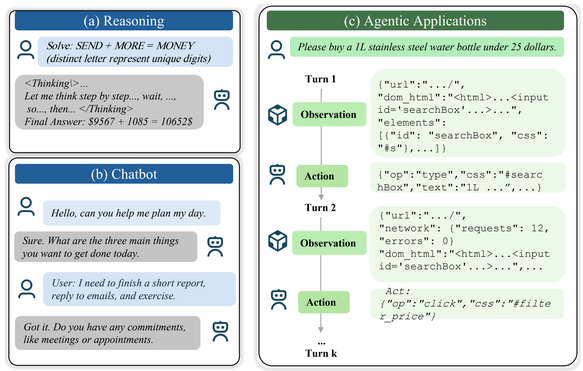
Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, Yuandong Tian, Jishen Zhao

University of California, San Diego and Independent Research

Motivation and Gap

Existing benchmarks are dialogue-centric and miss three key properties of agentic trajectories:

- **Representation diversity:** Machine-generated content (JSON, HTML, SQL, code).
- **Causal structure:** Actions induce state transitions.
- **Dense objectives:** Precise, sparse observations.



Contributions

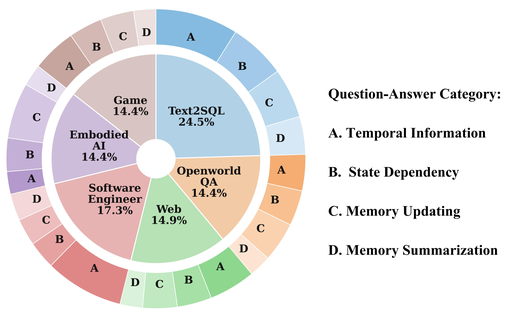
- **AMA-Bench:** first benchmark suite evaluating memory in real agentic applications.
- **Comprehensive evaluation:** shows many memory systems trail long-context baselines on agentic traces.
- **AMA-Agent:** causality graph for lossless construction and tool-augmented hybrid retrieval.

AMA-Bench Design

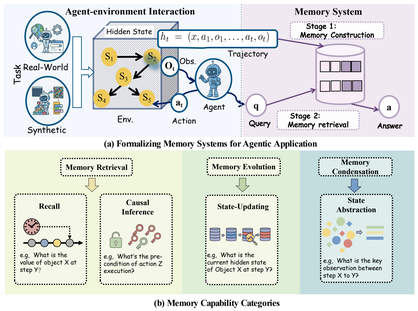
Two complementary subsets:

- **Real-world:** 6 domains, 2,496 expert-annotated QA pairs.
- **Synthetic:** 1,200 QA pairs across 5 trajectory lengths (8K–128K tokens) in TextWorld and BabyAI.

Evaluation capabilities: **Recall**, **Causal Inference**, **State Updating**, **State Abstraction**.

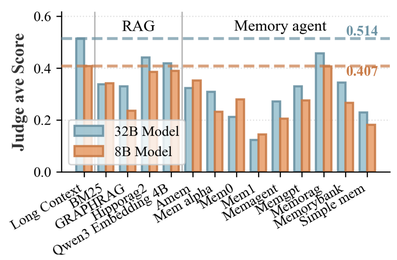


Memory Formulation



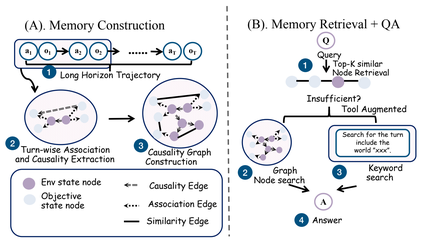
Empirical Motivation

- **Architecture > scale:** 8B→32B gives only +0.038 avg.; architecture variance is 0.45.
- **Compression loses signal:** MemoryBank drops 41.3% after construction; retrieval cannot recover it.
- **Memory still trails:** Long-context baselines outperform most memory systems.



The AMA-Agent

- **Causality Graph:** from (o_{t-1}, a_t, o_t), build state nodes and directed causal/association edges.
- **Tool-Augmented Retrieval:** top-K embedding retrieval with self-check, then graph traversal or keyword/script search if needed.



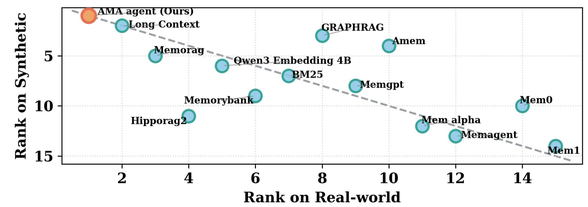
Main Results (Real-world Subset)

Method	Recall	Causal Inference	State Updating	State Abstraction	Average
Long-context (upper bound)					
GPT 5.2	0.774	0.805	0.656	0.604	0.723
GPT-5-mini	0.695	0.716	0.658	0.624	0.678
Qwen3-32B	0.554	0.564	0.510	0.443	0.518
Memory systems (Qwen3-32B backbone)					
AMA-Agent (ours)	0.624	0.615	0.531	0.472	0.572
MemoRAG	0.471	0.550	0.426	0.366	0.461
HippoRAG2	0.458	0.508	0.440	0.354	0.448
MemoryBank	0.323	0.410	0.301	0.333	0.340
Qwen3-Emb-4B	0.484	0.497	0.352	0.301	0.423

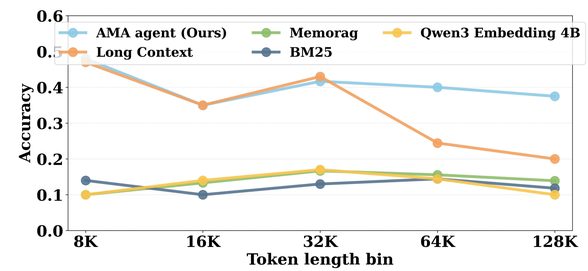
Gold: best; Gray: second best in each comparison group.

AMA-Agent surpasses the strongest memory baseline (MemoRAG) by **+11.16 percentage points**.

Scaling and Ablation



A. Correlation Analysis between Synthetic and Real-world Performance



B. Scalability Analysis Across Trajectory Lengths

Conclusion

- AMA-Bench is the first benchmark for long-horizon agent memory, covering 6 real-world domains (2,496 QA) and synthetic scaling to 128K tokens.
- Memory architecture is the primary bottleneck—not backbone scale—and existing systems underperform long-context baselines due to lossy compression and similarity-only retrieval.
- AMA-Agent's causality graph and tool-augmented retrieval achieve new SOTA, outperforming the best memory baseline by +11.16% and scaling robustly to 128K.

Scan for paper, code & leaderboard