NEURAL INFORMATION PROCESSING SYSTEMS

EvoMemBench

Benchmarking Agent Memory from a Self-Evolving Perspective

Yuyao Wang, Zhongjian Zhang, Mo Chi, Kaichi Yu,
Yuhan Li, Miao Peng, Bing Tong, Chen Zhang, Yan Zhou, Jia Li

HKUST (Guangzhou), Createlink Technology,
Beijing University of Posts and Telecommunications, Beijing Institute of Technology

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

创邻科技 CREATE LINK

北京郵電大學 Beijing University of Posts and Telecommunications

Why evaluate self-evolving memory?

NEURAL INFORMATION PROCESSING SYSTEMS

Agent memory must adapt as interaction unfolds

Growing histories	Finite context windows limit access to earlier observations and tool outputs.
Changing information	Agents must retain valid evidence and revise information that becomes outdated.
Reusable experience	Past episodes can provide knowledge and procedures for later tasks.

Existing benchmarks cover these demands only in part.

Two axes, four memory settings

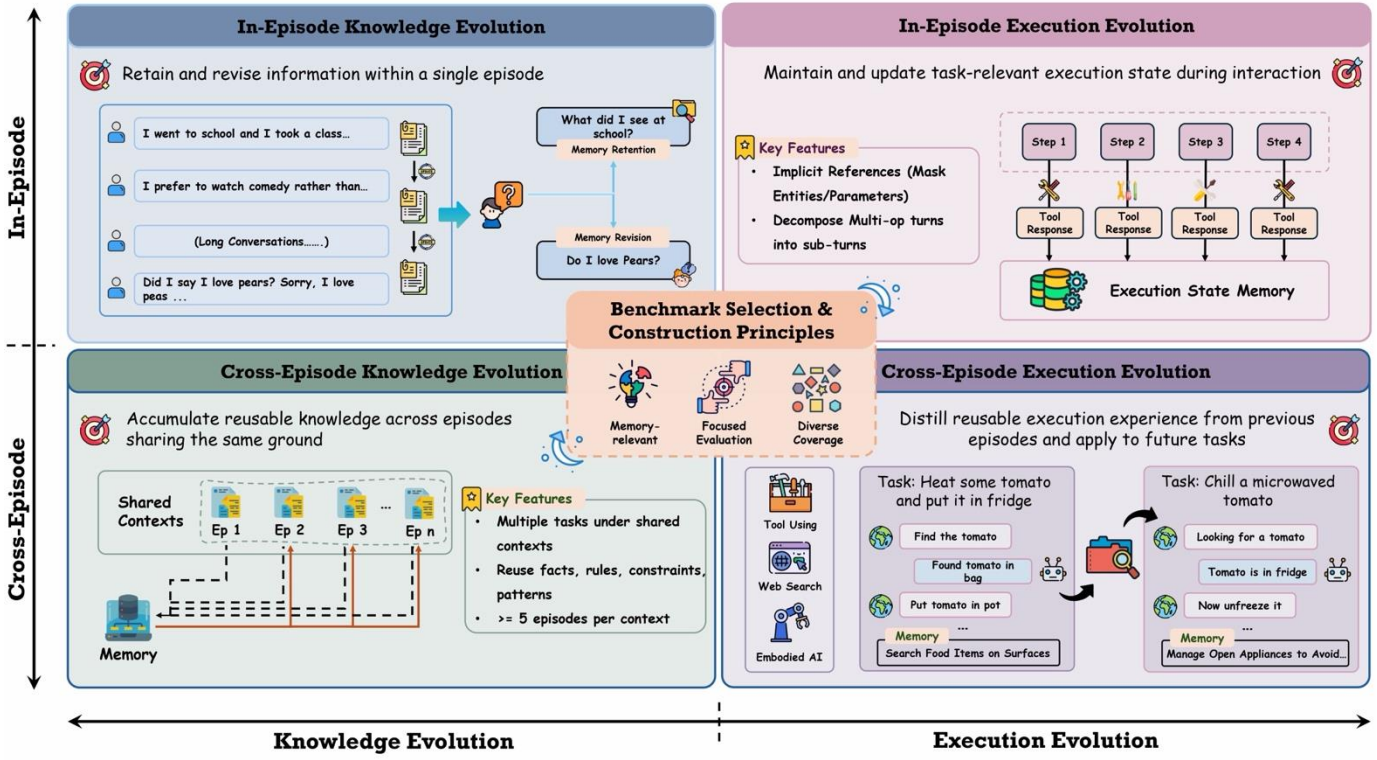


Figure 1 EvoMemBench overview

Memory scope

In-episode: adapt within one task attempt.
Cross-episode: reuse experience in later attempts.

Memory content

Knowledge: facts, constraints, rules.
Execution: task state, tool outputs, procedures.

Six datasets cover the four settings

Table 1 Dataset statistics

Setting	Dataset	Source	Subset / Domain (#Samples)
In-Episode Knowledge	INEP-KNOW	MemoryAgentBench	Accurate Retrieval (1000); Selective Forgetting (200)
In-Episode Execution	INEP-EXEC	BFCL-MultiTurn-LongContext	Gorilla File System (50); Vehicle Control (50); Trading Bots (50); Travel Booking (50)
Cross-Episode Knowledge	CROSSEP-KNOW	CL-Bench	Domain Knowledge Reasoning (294); Rule System Application (257); Procedural Task Execution (306); Empirical Discovery & Simulation (27)
Cross-Episode Execution	CROSSEP-TOOL	BFCL-MultiTurn-Base	Gorilla File System (50); Vehicle Control (50); Trading Bots (50); Travel Booking (50)
	CROSSEP-WEB	xbench-DeepSearch	Deep Search (100)
		WebWalkerQA	Web QA (170)
	CROSSEP-EMB	ALFWorld	Examine in Light (19); Pick&Place (46); Clean&Place (37); Cool&Place (28); Heat&Place (25); Pick Two&Place (45)

Within an episode

INEP-KNOW: retention and revision
INEP-EXEC: stateful tool use

Across episodes

CROSSEP-KNOW: shared-context knowledge
CROSSEP-TOOL / WEB / EMB: execution experience

Selection criteria: memory relevance, verifiable outcomes, diverse coverage.

A standardized evaluation protocol

NEURAL INFORMATION PROCESSING SYSTEMS

15 memory methods across five categories

Retrieval-augmented / Short-term / General long-term
Procedural long-term / Meta-evolution

Controlled backbone

All memory-augmented agents use DeepSeek-V3.2.
Gemini-3-Flash and GPT-5-mini provide additional long-context baselines.

Shared interfaces

utilize retrieves or supplies relevant memory. update incorporates new information.

Setting-specific lifecycle

Within-episode memory resets after each task. Cross-episode memory persists within a context or environment group.

Metrics: answer accuracy, task success rate, total inference tokens.

In-Episode Knowledge

NEURAL INFORMATION PROCESSING SYSTEMS

Table 2 INEP-KNOW retention and revision

Category	Method	Memory Retention					Memory Revision			Overall Rank
		Event QA	LME (S*)	Ruler qa1	Ruler qa2	Avg. Rank	FC MH	FC SH	Avg. Rank	
Long-Context LLM	Gemini-3-Flash	**98.20**	**83.00**	**94.00**	**77.00**	**1.00**	**58.00**	**96.00**	**1.00**	**1.00**
	GPT-5-mini	80.60	62.33	85.00	73.00	3.50	24.00	77.00	2.00	3.00
	DeepSeek-V3.2	83.00	32.33	64.00	42.00	6.75	10.00	65.00	3.00	5.50
Retrieval-Augmented Memory	BM25	88.80	50.33	73.00	55.00	**4.25**	**7.00**	**54.00**	**4.00**	**4.17**
	Qwen3-Emb-4B	83.60	21.33	38.00	38.00	8.50	4.00	30.00	7.50	8.17
	GraphRAG	74.20	22.00	31.00	31.00	10.50	4.00	18.00	9.50	10.17
Short-Term Memory	MemAgent	71.80	12.67	26.00	39.00	10.75	3.00	19.00	10.00	10.50
	MemoBrain	77.60	10.67	30.00	32.00	11.00	1.00	20.00	11.00	11.00
General Long-Term Memory	Mem0	79.20	**52.00**	**75.00**	**68.00**	4.50	**7.00**	49.00	4.50	4.50
	A-MEM	**91.20**	51.00	60.00	46.00	**4.25**	5.00	35.00	6.00	4.83
	MemOS	90.40	41.33	55.00	43.00	5.50	3.00	35.00	7.50	6.17
	MemoryOS	84.80	38.67	41.00	37.00	7.50	2.00	30.00	9.50	8.17

Long-context baseline

Gemini-3-Flash ranks first for both retention and revision.

Retention improves

BM25 on EventQA: 83.0% to 88.8% with the same backbone.

Revision remains hard

Multi-hop FactConsolidation: DeepSeek 10.0%, BM25 7.0%

In-Episode Execution

NEURAL INFORMATION PROCESSING SYSTEMS

Table 3 INEP-EXEC across context budgets

Method	Gorilla File System				Vehicle Control				Trading Bots				Travel Booking				Overall			
	16K	32K	64K	128K	16K	32K	64K	128K	16K	32K	64K	128K	16K	32K	64K	128K	16K	32K	64K	128K
Long-Context LLM																				
Gemini-3-Flash	**54.0**	**50.0**	**52.0**	**52.0**	**60.0**	**70.0**	**77.6**	**60.0**	**28.0**	**80.0**	**78.0**	**80.0**	**60.0**	**62.0**	**60.0**	**56.0**	**50.5**	**65.5**	**66.9**	**62.0**
GPT-5-mini	34.0	34.0	40.0	34.0	46.0	44.0	63.7	44.0	20.0	62.0	62.0	56.0	40.0	30.0	32.0	36.0	35.0	42.5	49.4	42.5
DeepSeek-V3.2	52.0	46.0	46.0	48.0	38.0	40.0	73.2	44.0	8.0	40.0	46.0	46.0	16.0	16.0	16.0	20.0	28.5	35.5	45.3	39.5
Retrieval-Augmented Memory																				
BM25	50.0	56.0	48.0	54.0	50.0	46.0	72.8	40.0	20.0	54.0	48.0	42.0	26.0	22.0	20.0	26.0	36.5	44.5	47.2	40.5
Qwen3-Emb-4B	48.0	**58.0**	44.0	58.0	42.0	48.0	68.7	42.0	16.0	32.0	46.0	44.0	32.0	20.0	26.0	24.0	34.5	39.5	46.2	42.0
GraphRAG	56.0	56.0	54.0	54.0	48.0	38.0	**80.2**	44.0	12.0	24.0	46.0	42.0	24.0	28.0	**28.0**	24.0	35.0	36.5	52.0	41.0
Short-Term Memory																				
MemAgent	48.0	48.0	54.0	46.0	34.0	36.0	79.8	36.0	10.0	40.0	40.0	**52.0**	12.0	22.0	16.0	14.0	26.0	36.5	47.4	37.0
MemoBrain	40.0	46.0	54.0	46.0	40.0	32.0	38.0	42.0	4.0	32.0	50.0	42.0	16.0	12.0	20.0	26.0	25.0	30.5	40.5	39.0
General Long-Term Memory																				
Mem0	54.0	44.0	46.0	**60.0**	46.0	44.0	70.9	38.0	16.0	48.0	38.0	38.0	24.0	18.0	24.0	20.0	35.0	38.5	44.7	39.0
A-MEM	52.0	52.0	50.0	46.0	48.0	44.0	76.2	48.0	**34.0**	36.0	48.0	40.0	16.0	26.0	24.0	26.0	37.5	39.5	49.5	40.0
MemOS	52.0	52.0	54.0	50.0	44.0	36.0	79.0	38.0	12.0	40.0	48.0	46.0	22.0	12.0	14.0	16.0	32.5	35.0	48.8	37.5
MemoryOS	48.0	**58.0**	42.0	44.0	36.0	36.0	72.5	40.0	10.0	38.0	42.0	20.0	20.0	20.0	22.0	26.0	28.5	38.0	44.6	32.5
Procedural Long-Term Memory																				
AWM	50.0	46.0	48.0	48.0	46.0	46.0	78.3	32.0	10.0	**58.0**	**72.0**	42.0	20.0	22.0	14.0	18.0	31.5	43.0	**53.1**	35.0
SkillWeaver	48.0	50.0	48.0	54.0	44.0	42.0	73.8	46.0	14.0	42.0	58.0	46.0	30.0	28.0	26.0	32.0	34.0	40.5	51.5	44.5
AgentKB	46.0	50.0	54.0	50.0	8.0	40.0	77.2	**56.0**	18.0	26.0	56.0	50.0	18.0	8.0	10.0	14.0	22.5	31.0	49.3	42.5
ACE	52.0	42.0	**56.0**	54.0	42.0	34.0	78.4	44.0	8.0	42.0	44.0	48.0	26.0	22.0	24.0	20.0	32.0	35.0	50.6	41.5
ReasoningBank	**62.0**	52.0	50.0	54.0	**60.0**	**52.0**	74.6	54.0	16.0	44.0	36.0	48.0	**34.0**	**50.0**	**28.0**	**36.0**	**43.0**	**49.5**	47.1	**48.0**
Meta-Evolution Memory																				
MemEvolve	52.0	48.0	46.0	54.0	42.0	44.0	74.8	50.0	20.0	50.0	62.0	46.0	30.0	22.0	20.0	22.0	36.0	41.0	50.7	43.0

16K context

28.5% to 43.0% success
+14.5 percentage points

64K context

45.3% to 53.1% success
+7.8 percentage points

Execution details

Procedural guidance helps.
Compression can lose tool arguments and intermediate state.

Cross-Episode Knowledge

NEURAL INFORMATION PROCESSING SYSTEMS

Table 4 CROSSEP-KNOW by difficulty

Category	Method	Domain Knowledge Reasoning			Empirical Discovery & Simulation			Procedural Task Execution			Rule System Application			Overall		
		Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Long-Context LLM	Gemini-3-Flash	44.4	17.9	9.1	0.0	0.0	11.1	42.9	12.6	5.0	61.8	21.2	6.0	37.3	13.0	7.8
	GPT-5-mini	61.7	**37.6**	**16.5**	**50.0**	0.0	**22.2**	54.7	**22.3**	**8.8**	**76.7**	**26.6**	**9.9**	**60.8**	**21.6**	**14.4**
	DeepSeek-V3.2	**64.6**	21.4	0.0	10.0	0.0	0.0	**57.9**	10.0	0.0	75.7	18.8	0.0	52.1	12.6	0.0
Retrieval-Augmented Memory	BM25	60.9	23.5	4.8	0.0	0.0	22.2	54.5	13.2	**12.1**	73.5	18.4	1.4	47.2	13.8	10.1
	Qwen3-Emb-4B	63.5	23.7	6.7	**10.0**	0.0	11.1	56.3	13.0	8.1	70.3	18.3	5.9	**50.0**	13.7	7.9
	GraphRAG	60.4	25.4	7.6	0.0	0.0	11.1	59.2	13.4	9.2	72.4	15.9	2.8	48.0	13.7	7.7
General Long-Term Memory	Mem0	55.3	15.9	5.1	0.0	0.0	0.0	50.5	12.5	4.4	72.6	18.4	5.8	44.6	11.7	3.8
	A-MEM	49.4	13.8	2.9	**10.0**	0.0	11.1	50.4	16.1	7.4	65.3	10.1	5.5	43.8	10.0	6.7
	MemOS	55.7	25.4	4.8	**10.0**	0.0	11.1	57.5	16.3	7.4	63.9	17.3	4.0	46.8	14.8	6.8
	MemoryOS	51.5	19.0	1.9	**10.0**	0.0	22.2	52.7	12.4	6.6	62.7	12.7	5.4	44.2	11.0	9.0
Procedural Long-Term Memory	AWM	54.2	22.7	**9.0**	0.0	0.0	0.0	**60.9**	12.8	6.9	68.8	14.4	3.3	46.0	12.5	4.8
	SkillWeaver	**64.4**	23.5	4.8	**10.0**	0.0	11.1	58.8	7.1	2.2	60.5	14.8	4.5	48.4	11.3	5.6
	AgentKB	56.9	22.1	5.1	**10.0**	0.0	0.0	51.2	10.3	5.5	64.5	12.7	3.7	45.6	11.3	3.6
	ACE	58.0	24.6	2.9	0.0	0.0	**33.3**	46.2	17.2	8.1	**74.9**	**24.0**	**7.8**	44.8	**16.5**	**13.0**
	ReasoningBank	60.9	**27.5**	6.1	0.0	0.0	0.0	48.0	**18.9**	11.0	62.1	16.5	7.4	42.7	15.7	6.1
Meta-Evolution Memory	MemEvolve	54.3	16.7	3.8	**10.0**	0.0	11.1	47.3	12.4	11.4	72.5	18.9	3.1	46.0	12.0	7.4

Easy split

No memory: 52.1%
Best memory: 50.0%

Hard split

No memory: 0.0%
Best memory, ACE: 13.0%

Cross-Episode Execution

Table 5 Cross-episode execution results

Method	Tool Using					Web Search			Embodied AI							Overall Rank
	Gorilla File System	Vehicle Control	Trading Bots	Travel Booking	Avg. Rank	xbench -DS	WebWalker QA	Avg. Rank	Examine in Light	Pick &Place	Clean &Place	Cool &Place	Heat &Place	Pick Two &Place	Avg. Rank	
Long-Context LLM																
Gemini-3-Flash	**74.0**	**62.0**	**88.0**	**54.0**	**3.0**	**76.0**	**72.4**	**5.0**	68.4	76.1	43.2	**64.3**	**60.0**	26.7	12.0	**7.8**
GPT-5-mini	50.0	50.0	72.0	44.0	9.5	72.0	71.8	10.0	36.8	69.6	21.6	25.0	40.0	40.0	15.5	12.7
DeepSeek-V3.2	64.0	36.0	54.0	26.0	13.5	74.0	61.2	14.0	**73.7**	**93.5**	**56.8**	60.7	56.0	**57.8**	**10.0**	11.8
Retrieval-Augmented Memory																
BM25	68.0	**88.0**	66.0	32.0	5.8	79.0	65.9	6.5	68.4	**97.8**	62.2	60.7	64.0	71.1	7.8	6.9
Qwen3-Emb-4B	68.0	82.0	78.0	34.0	**4.0**	78.0	65.3	8.0	79.0	**97.8**	70.3	60.7	64.0	71.1	6.0	5.7
GraphRAG	66.0	68.0	72.0	40.0	6.5	73.0	70.0	11.0	63.2	**97.8**	75.7	60.7	72.0	77.8	5.8	6.9
General Long-Term Memory																
Mem0	**76.0**	58.0	48.0	30.0	8.5	**82.0**	70.6	**4.0**	79.0	89.1	67.6	57.1	68.0	53.3	8.2	7.9
A-MEM	68.0	62.0	70.0	26.0	8.8	76.0	71.8	6.0	**84.2**	91.9	**81.1**	64.3	**80.0**	**84.4**	**1.8**	**5.2**
MemOS	72.0	60.0	36.0	28.0	9.8	78.0	71.2	5.0	63.2	91.3	67.6	64.3	68.0	53.3	7.8	8.4
MemoryOS	72.0	72.0	56.0	36.0	5.8	78.0	69.4	6.5	73.7	95.7	75.7	64.3	60.0	57.8	6.5	6.5
Procedural Long-Term Memory																
AWM	64.0	56.0	74.0	24.0	11.0	76.0	72.4	5.0	73.7	91.3	78.4	71.4	76.0	64.4	4.3	7.2
SkillWeaver	66.0	48.0	44.0	28.0	12.3	74.0	64.7	13.5	63.2	84.8	59.5	53.6	64.0	66.7	10.2	12.2
AgentKB	62.0	71.4	74.0	30.0	8.3	79.0	68.8	6.0	68.4	89.1	67.6	**85.7**	**80.0**	73.3	4.2	6.5
ACE	68.0	56.0	14.0	34.0	9.8	77.0	65.3	9.5	**84.2**	78.3	75.7	21.4	76.0	82.2	4.8	8.2
ReasoningBank	64.0	82.0	**86.0**	**42.0**	4.8	76.0	**75.3**	4.5	79.0	87.0	**81.1**	67.9	60.0	73.3	4.5	5.4
Meta-Evolution Memory																
MemEvolve	66.0	70.0	66.0	28.0	8.8	73.0	65.3	13.0	57.9	89.1	64.9	78.6	76.0	75.6	6.0	8.8

Tool use

Retrieval-augmented memory
Best family average rank: 5.43

Web search

General long-term memory
Best family average rank: 5.38

Embodied interaction

Procedural long-term memory
Best family average rank: 5.60

Transfer depends on the target task

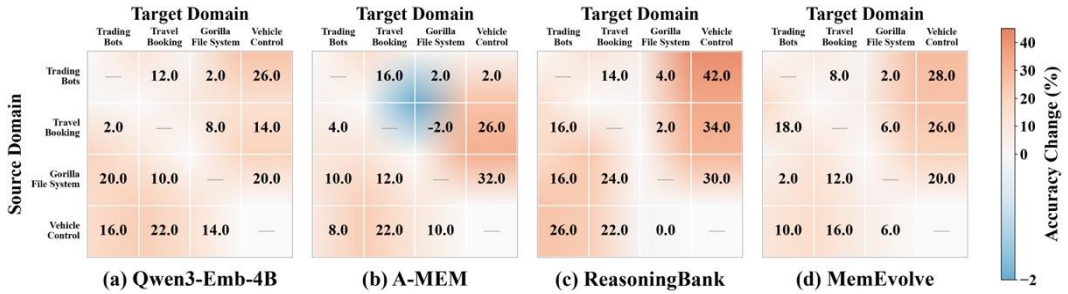


Figure 3 CROSSEP-TOOL transfer

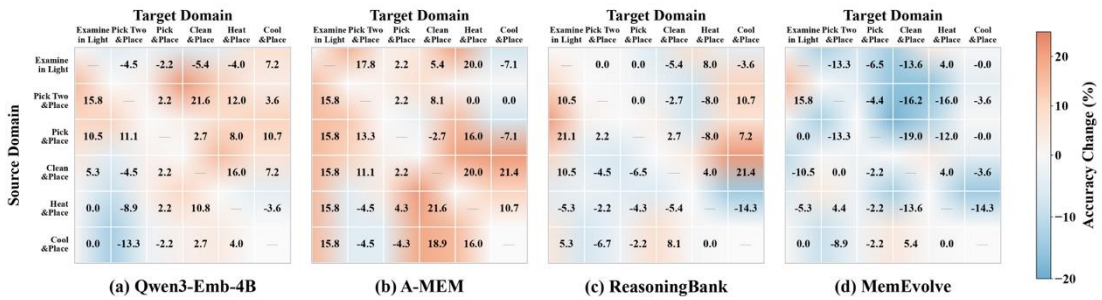


Figure 4 CROSSEP-EMB transfer

Tool-use transfer shows more stable gains.

Embodied transfer is less stable and can hurt.

Protocol: build memory on the source, then freeze it during target evaluation.

Takeaways and future directions

NEURAL INFORMATION PROCESSING SYSTEMS

A unified evaluation view

Four settings connect memory scope with knowledge and execution demands.

Conditional benefits

Memory helps most when context is insufficient, tasks are hard, or experience matches the target.

An open research problem

Selective memory use, reliable revision, and task-matched representations remain promising directions.

EvoMemBench evaluates existing methods. It does not introduce a new memory architecture.

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!

EvoMemBench: Benchmarking Agent Memory from a Self-Evolving Perspective

Yuyao Wang, Zhongjian Zhang, Mo Chi, Kaichi Yu,
Yuhan Li, Miao Peng, Bing Tong, Chen Zhang, Yan Zhou, Jia L

Code and data

github.com/DSAIL-Memory/EvoMemBench

Scan for code