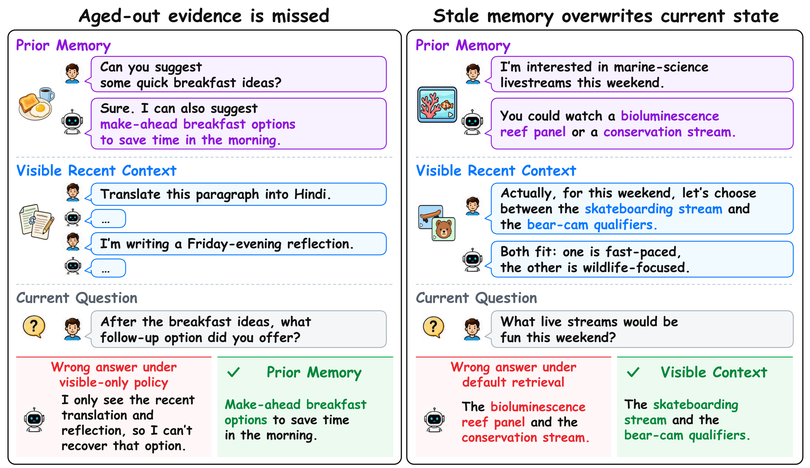
Know Where You Stand: Memory-Source Choice in Long-Context Dialogue Agents

Caishen Zhou, Yihong Tang, Xuefeng Bai, Kehai Chen*, Min Zhang
Harbin Institute of Technology, Shenzhen

NEURAL INFORMATION PROCESSING SYSTEMS

01–03 | BENCHMARK AND PHENOMENON

01 Two ways memory fails



The **breakfast question** needs an earlier offer.
The **livestream question** needs the user's updated plan.

02 MSBench: two questions per checkpoint

45	450	900
scenes	checkpoints	questions

Setup: Each pair shares 16 recent turns. Future turns are hidden.

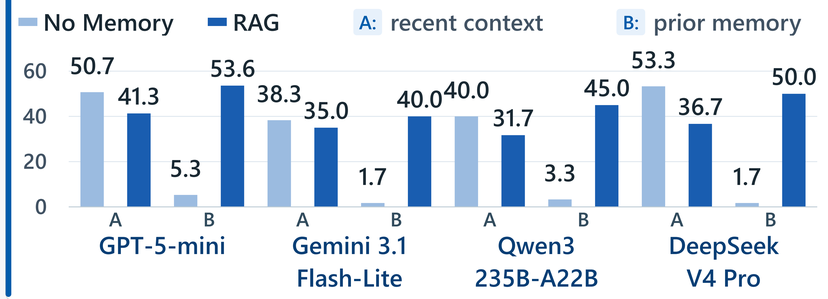
A Answer from the **recent window**.

B Recover evidence from an **earlier exchange**.

For **184 of 450** B questions (**40.9%**), recent turns match the topic but refer to the wrong instance or version.

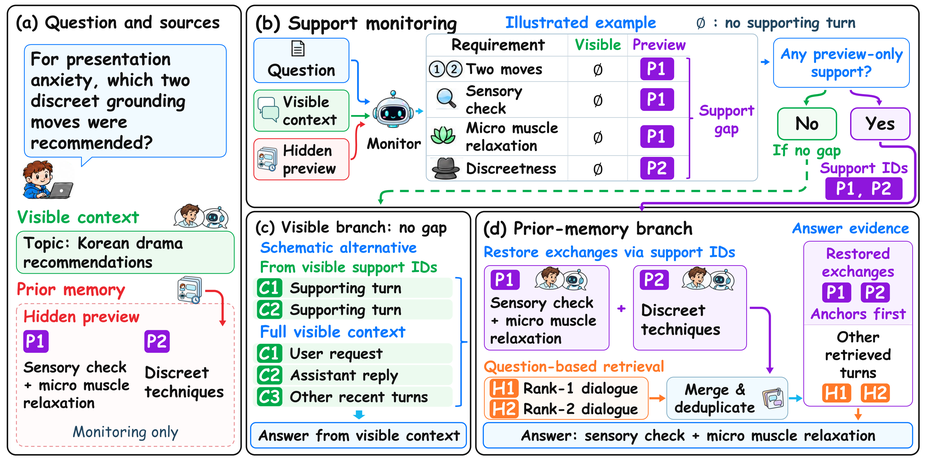
03 Retrieval helps B but hurts A

Answer accuracy (%)



04–06 | METHOD AND RESULTS

04 The Memory-Source Reasoner



MSR checks which facts the **visible turns** support.
If only the **memory preview** fills a gap, it restores the supporting exchange.

05 Main results on MSBench

GPT-5-mini

Policy	Answer accuracy (%)	Source selection (%)
No Memory	28.0	50.0
RAG	47.4	50.0
Prompted Gate	46.8	75.4
MSR	60.0	81.7

MSR: **+12.6 points** vs. RAG on answers; **+6.3** vs. Prompted Gate on selection.

Other backbones	Answer accuracy (%)		Source selection (%)	
	Best baseline	MSR	Best baseline	MSR
Gemini 3.1 Flash-Lite	37.5	**44.2**	66.7	**79.2**
Qwen3 235B-A22B	38.3	**41.7**	55.0	**72.5**
DeepSeek V4 Pro	49.2	**55.8**	81.7	**82.5**

06 Transfer to LoCoMo and QMSum

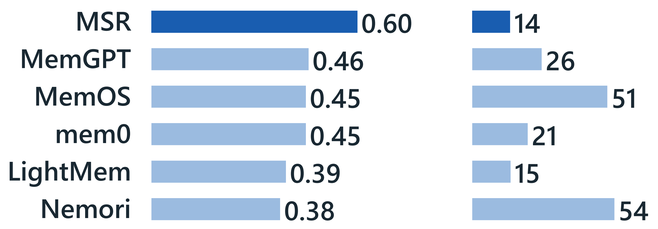
Prompted Gate → MSR (%). Bars show the gain in percentage points.

	LoCoMo		QMSum	
Answer	78.6 → 80.7	+2.1	41.3 → 55.3	+14.0
Selection	77.0 → 85.9	+8.9	71.3 → 88.7	+17.4

07–09 | SYSTEMS AND DIAGNOSTICS

07 Memory-system comparison

MSBench results. Latency is per question.



Native configurations differ across systems.

08 What each component adds

GPT-5-mini answer accuracy (%)

Variant	Recent	Prior	Overall
Full MSR	**61.7**	**58.3**	**60.0**
No atom split	**53.3**	**48.3**	**50.8**
No gap selection	**53.3**	**58.3**	**55.8**
No visible support pack	**48.3**	**55.0**	**51.7**
No exchange expansion	**61.7**	**50.0**	**55.8**

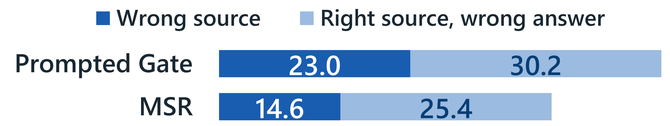
Darker blue = larger loss vs. full MSR.

−13.4 pts
on recent questions without visible support

−8.3 pts
on prior questions without exchange expansion

09 Which errors remain?

Wrong answers per 100 GPT-5-mini questions



On **25.4%** of all questions, MSR chooses the **right source** but still gives the **wrong answer**.

TAKEAWAY | **External memory can be a source of error, not just a source of evidence.**

Scan for the conference paper

* Correspondence: chenkehai@hit.edu.cn