SHIFTBENCH: MEASURING RECOVERY OF AGENT MEMORY UNDER DISTRIBUTION SHIFT

Teresa Zhang

ABSTRACT

Selecting memory policies by long-horizon accuracy can be misleading under shift, because rankings may reverse when evaluated by post-shift recovery. We introduce ShiftBench, a lightweight protocol defining shift segments and Recovery@T on LoCoMo and HaluMem-Long. On LoCoMo, lexical baselines (TF–IDF methods) show reversal under interruption (Spearman $\rho = -0.30$, inversion 0.60), and alignment drops from 0.94 to 0.70 ($\Delta\rho = 0.24$, 95% CI $[0.12, 0.37]$). On HaluMem-Long, reversal is smaller but still present ($\rho = 0.02$, inversion 0.50). Overall, ShiftBench shows that post-shift recovery is a distinct evaluation axis that can change memory-policy selection.

1 PROTOCOL AND METRICS

Motivation. LLM agents rely on external memory beyond the context window, motivating explicit memory policies (Packer et al., 2023). Benchmarks such as LoCoMo test long-term conversational memory (Maharana et al., 2024), but current evaluation emphasizes average long-horizon scores that can obscure slow post-shift recovery. ShiftBench is a lightweight reporting protocol that makes recovery explicit and can alter model selection under distribution shift.

Protocol. We mark shift points via dataset session boundaries; on LoCoMo we also inject context-pollution bursts to stress memory. These inserted off-topic turns serve as a controlled proxy for context pollution (e.g., tool logs or off-topic interleavings) in long-running agents. Letting t_0 be the shift time and Score(t) the per-question evidence-hit score (binary Hit@k at turn t), we define

$$\text{Recovery@T} = \frac{1}{T} \sum_{t=t_0}^{t_0+T-1} \text{Score}(t), \tag{1}$$

which exposes slow recovery even when average accuracy is high. $\mathrm{Score}(t)$ is binary $\mathrm{Hit@}k$ (1 if any gold evidence appears in the retrieved top-k) with fixed $k = 5$. We report overall Hit@k on the clean stream and Recovery@T as early-window performance after shift; this tests whether clean-stream selection predicts post-shift recovery. Inversion rate and Spearman rank correlation (ρ) measure ranking disagreement between these metrics (full definitions in Appendix A).

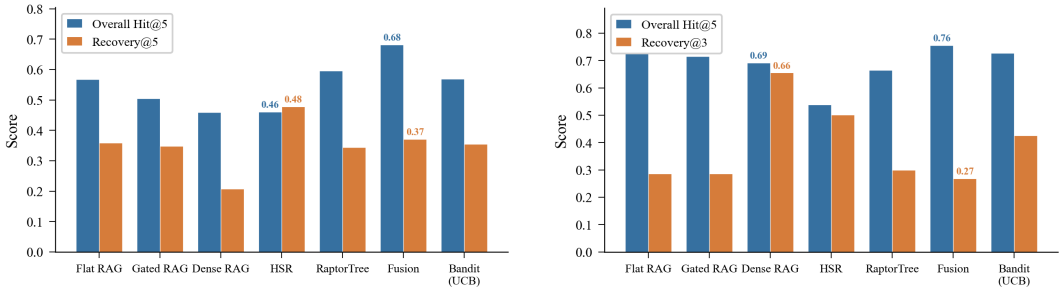
Baselines. We compare Flat RAG; Gated RAG, a heuristic retrieve/skip variant (not learned Self-RAG; Asai et al., 2023); Dense RAG (Gemini embeddings); HSR; RaptorTree, a shallow TF–IDF clustering tree inspired by RAPTOR but not a full RAPTOR implementation (Sarthi et al., 2024); and Fusion ($\alpha = 0.5$) (details in Appx. B). We also report a contextual UCB bandit over Flat/Dense/HSR/Fusion as a simple adaptive routing baseline, without policy analysis.

2 EXPERIMENTAL SETUP AND RESULTS

We evaluate on LoCoMo (10 conversations, 1,977 QA turns, 1,911 post-shift) (Maharana et al., 2024) and a fixed HaluMem-Long subset selected before evaluation (20 conversations, 720 QA turns) (Chen et al., 2025). For LoCoMo, we inject context pollution (m=100) by sampling off-topic turns from other sessions before the first N=3 post-boundary questions (see Appx. A). Runs use memory budget C and k=5. Evidence matching is deterministic (LoCoMo uses dataset evidence IDs; HaluMem-Long uses fixed normalized string matching; see Appx. A for rules). For HaluMem-Long, we use m=50 and report Recovery@3.

	Flat RAG	Gated RAG	Dense RAG	HSR	RaptorTree	Fusion	Bandit (UCB)
Overall Hit@5	0.568	0.504	0.459	0.461	0.596	**0.682**	0.569
Recovery@5 (m=0)	0.571	0.506	0.388	0.483	0.605	**0.676**	0.557
Recovery@5 (m=100)	0.359	0.348	0.208	**0.478**	0.345	0.371	0.354

Table 1: LoCoMo results ($C = 50$, retrieval $k = 5$). Overall Hit@5 is computed on the clean stream; Recovery@5 ($m = 0$) is measured at natural boundaries, while Recovery@5 ($m = 100$) is measured under interruption. Under interruption, HSR overtakes Fusion on recovery despite weaker overall accuracy; lexical per-conversation ρ drops from 0.94 to 0.70.



(a) LoCoMo (m=100) (b) HaluMem-Long (m=50)

Figure 1: Rank disagreement under distribution shift (overall Hit@k measured on the clean stream; Recovery@T measured under interruption). (a) LoCoMo: Fusion is best overall (0.682) but HSR leads on Recovery@5 (0.478). (b) HaluMem-Long: Fusion is best overall (0.756) but Dense RAG leads on Recovery@3 (0.661).

Table 1 shows the central finding: model rankings based on overall accuracy can change when evaluated on post-shift recovery. At natural boundaries ($m = 0$), overall Hit@5 and Recovery@5 are highly aligned across conversations (per-conversation $\rho = 0.94$, 95% CI $[0.88, 0.99]$). Under interruption ($m = 100$), this alignment weakens: although Fusion remains best overall (0.682), HSR achieves the highest Recovery@5 (0.478). Among lexical baselines (Flat/Gated/HSR/RaptorTree/Fusion), interruption produces substantial aggregate rank disagreement ($\rho = -0.30$, inversion 0.60), and per-conversation alignment drops to $\rho = 0.70$ (95% CI $[0.53, 0.84]$). Dense RAG also shows sensitivity to shift, degrading from Recovery@5 0.388 to 0.208 under interruption. Results on HaluMem-Long show smaller but non-trivial disagreement ($\rho = 0.02$, inversion 0.50), with Fusion best overall (0.756) and Dense RAG best on Recovery@3 (0.661). Together, these results show that recovery-focused evaluation can change memory-policy selection.

3 Discussion, Limitations, and Future Directions

ShiftBench separates long-horizon accuracy from post-shift recovery, revealing ranking differences that aggregate metrics can obscure. The protocol is intentionally lightweight, using evidence-hit scoring rather than full end-to-end generation; while this underestimates downstream generation errors, a small validation check shows that Recovery@T ordering agrees with answer-level F1 outcomes (Appendix D). Our results also suggest a possible mechanism: hierarchical retrieval (e.g., HSR) behaves like a low-pass filter, where summary-level selection is less sensitive to localized interruption bursts, yielding faster recovery than flat retrieval despite weaker overall accuracy. Related robustness benchmarks target retrieval perturbations (RARE, Zeng et al., 2025) or agent recovery from execution errors (Recovery-Bench, Tan et al., 2025b); in contrast, ShiftBench measures recovery dynamics following temporal distribution shifts in memory policies, treating adaptation speed as an explicit evaluation axis. Overall, ShiftBench aims to provide a reusable reporting lens that exposes adaptation failures masked by average performance.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. Halumem: Evaluating hallucinations in memory systems of agents, 2025. URL https://arxiv.org/abs/2511.03506.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025. URL https://arxiv.org/abs/2507.05257.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*, 2024. URL https://arxiv.org/abs/2403.14403.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. *arXiv preprint arXiv:2402.17753*, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*, 2024.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, Vienna, Austria, 2025a. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.989. URL https://aclanthology.org/2025.findings-acl.989/.

Shangyin Tan, Kevin Lin, Koushik Sen, and Matei A. Zaharia. Recovery-bench: Evaluating agentic recovery from mistakes. In *NeurIPS 2025 Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025b. URL https://neurips.cc/virtual/2025/122490.

Xiaqiang Tang, Jian Li, Nan Du, and Sihong Xie. Adapting to non-stationary environments: Multi-armed bandit enhanced retrieval-augmented generation on knowledge graphs. *arXiv preprint arXiv:2412.07618*, 2024. URL https://arxiv.org/abs/2412.07618.

Xiaqiang Tang, Qiang Gao, Jian Li, Nan Du, Qi Li, and Sihong Xie. MBA-RAG: a bandit approach for adaptive retrieval-augmented generation through question complexity. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3248–3254, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.218/.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. Evo-memory: Benchmarking LLM agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*, 2025. URL https://arxiv.org/abs/2511.20857.

Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and Lei Li. RARE: Retrieval-aware robustness evaluation for retrieval-augmented generation systems, 2025. URL https://arxiv.org/abs/2506.00789.

APPENDIX

A EVALUATION PROTOCOL AND DEFINITIONS

A.1 TERMINOLOGY AND METRIC DEFINITIONS

Term	Definition
Shift point	Annotated session boundary in LoCoMo conversations.
Interruption	m off-topic turns sampled from other sessions (fixed seed) and inserted before first N post-boundary QA turns; indexed identically and competing under memory budget C.
Memory budget C	Maximum number of turn-level memory chunks (USER–ASSISTANT exchanges) stored externally.
Score(t)	Binary Hit@k at QA turn t.
Overall Hit@k	Mean Score(t) on clean stream $(m{=}0)$.
Recovery@T	Mean Score(t) over first T post-shift QA turns.
Sparse baselines	Flat, Gated, HSR, RaptorTree, Fusion (see Sec. B).
Inversion	Fraction of method pairs with differing rankings (ties assigned average ranks; pairs tied in either metric are treated as non-inversions).
Spearman ρ	Spearman rank correlation between Overall Hit@k and Recovery@T rankings using average ranks under ties.

Table 2: Summary of evaluation terminology and metrics.

A.2 EVIDENCE MATCHING PROTOCOL

Evidence matching is deterministic and fixed across runs.
Retrieval unit: turn-level chunk (USER–ASSISTANT exchange).
LoCoMo: Hit@k if any retrieved chunk contains a gold evidence ID provided by the dataset.
HaluMem-Long: exact equality after normalization (lowercase, strip punctuation, collapse whitespace, Unicode NFKC).
Multiple matches: any matching retrieved item counts as a hit.
Unmatched cases: unmatched rate after preprocessing = 0%.

B BASELINE IMPLEMENTATIONS

Method	Implementation details
Shared setup	Turn-level chunks (no overlap); Hit@k with $k{=}5$. Sparse retrieval uses TF–IDF cosine (recency tie-break); dense retrieval uses normalized cosine with `gemini-embedding-001`.
Flat RAG	Index all turn chunks; retrieve top-k via TF–IDF cosine.
Gated RAG	Heuristic retrieve/skip gate: max TF–IDF similarity $< \tau{=}0.2$ returns empty; otherwise Flat RAG ranking (not learned Self-RAG).
Dense RAG	Index all turn chunks with dense embeddings; retrieve top-k via embedding cosine.
HSR	Two-stage hierarchical retrieval: retrieve top session summaries (summary_k$=2$), then retrieve top-k turns within selected sessions.
RaptorTree	Hierarchical TF–IDF clustering (depth=2, branching=4, min leaf=20, top branches=2) + recent-turn buffer (20); rerank via TF–IDF similarity (RAPTOR-inspired, not full RAPTOR).
Fusion	Union candidates from Recency and RaptorTree; score $\text{sim}(q,i)((1-\alpha)+\alpha e^{-\text{age}(i)/\tau})$ with $\tau{=}50$.
Bandit (UCB)	Contextual UCB router over expert subset using boundary label, interruption flag, query length bin, and session position bin.
Fairness note	Capacity C applies only to bounded-memory policies; RAG-style methods use full external index.

Table 3: Baseline implementations used in experiments.

C Recovery Dynamics

To verify that rank reversal reflects sustained recovery behavior rather than an artifact of a fixed evaluation window, we visualize full Recovery@T trajectories under interruption. These curves reveal how methods adapt immediately after shift events and help explain why hierarchical retrieval overtakes flat retrieval despite weaker steady-state accuracy. Figure 2 shows Recovery@T curves for m=100: HSR maintains the highest recovery across all T, while Fusion begins lower than other sparse baselines but converges by T=8, consistent with the observed early-window rank reversal.

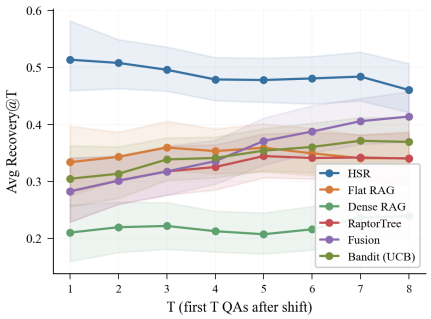


Figure 2: Recovery@T curves under interruption (m=100, LoCoMo). Lines show mean recovery across conversations; shaded regions denote 95% bootstrap confidence intervals (5000 samples).

D Pilot End-To-End Sanity Check

Because ShiftBench evaluates retrieval-level evidence hits rather than full end-to-end generation, we include a small sanity-check pilot to assess whether Recovery@T ordering aligns with downstream answer quality. Using Gemini-2.0-flash, we evaluate Fusion and HSR on 150 post-shift LoCoMo questions; Recovery@T and answer-level F1 yield the same method ordering. This pilot is not intended as validation, but confirms that the proxy metric does not contradict end-to-end outcomes in this setting. Answer F1 uses normalized token overlap (lowercase, punctuation stripped, whitespace collapsed) with dataset-provided canonical answers.

Method	n	EM	F1
Fusion	150	0.013	0.029
HSR	150	0.013	0.072

Table 4: Pilot end-to-end validation.

E Shift Sensitivity Analysis

We analyze how increasing interruption size m affects ranking stability. Table 5 shows that larger interruptions systematically increase disagreement between overall accuracy and recovery-based rankings.

Inversion rises from 0.00 at m=0 to 0.60 at $m \geq 100$, while Spearman correlation between overall Hit@k and Recovery@T drops from perfect alignment (ρ=1.00) to negative correlation (ρ=−0.30). This indicates that methods optimized for long-horizon averages may recover more slowly under interruption. Notably, moderate shifts ($m \approx 50$) already produce substantial ranking instability, suggesting recovery-focused evaluation becomes essential well before extreme stress levels.

These results reinforce the central claim of ShiftBench: adaptation speed represents an independent evaluation axis that aggregate accuracy alone cannot capture.

m (interrupt)	Inversion	Spearman ρ
0	0.000	1.000
50	0.500	-0.100
100	0.600	-0.300
150	0.600	-0.300

Table 5: Reversal metrics vs interruption size m (LoCoMo, lexical baselines).