Bridge Starvation: A Reproducible Failure Mode of Single-Pass Memory Retrieval, with a Bounded Repair

Anonymous Authors[1]

Abstract

Long-horizon agents can fail long before their final answer: a missed memory item at an early step quietly removes the evidence needed later. We isolate a reproducible instance of this problem in single-pass agentic-memory retrieval. The failure, which we call *bridge starvation*, occurs when a query names entity X but the answer requires a chain $X \to Y \to Z \to \ldots$; the first retrieval pass finds the X fact, but the bridge entities are absent from the query and the remaining links are never retrieved. We introduce DKMP, a controlled trigger that varies chain depth (2–5 hops) and context length (1K–128K tokens), and we pair it with trace diagnostics that separate retrieval coverage from reader composition. As a bounded repair, MULTIPASS reuses the same retriever but asks the LLM to emit one short bridge entity between retrieval passes. On the targeted 2/3/4-hop, 128K-token regime, MULTIPASS reaches 1.00, 0.90, and 0.97 accuracy (n=30 per cell), within 3pp of the gold-support oracle on 3/4-hop and above IR-CoT, ReAct, Self-Ask, and a passage-graph PPR variant under the same reader and index. The repair has clear boundaries: it does not solve 5-hop chains, where coverage falls to 0.75 and reader composition becomes a second failure mode; it does not help on short-context MuSiQue, where the trigger does not fire; and it does not target the cross-session aggregation failure that dominates LONGMEMEVAL. We release the trigger, traces, and per-pass diagnostics.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

A foundation-model agent often acts through a memory layer: it reads a long trace of prior turns, retrieves a few relevant items, and conditions the next step on those items. Reliability failures at this layer are hard to diagnose because the visible error may occur long after the missing memory item. A wrong final answer can therefore hide a simpler upstream question: did the agent ever retrieve the evidence needed to answer?

This paper isolates one retrieval-layer failure that is small enough to reproduce and sharp enough to repair. Many proposed agentic-memory systems expose a single-pass retrieval interface at inference: given the current query or state, retrieve a small set of memory items and pass them to a reader. The extraction, storage, and consolidation pipelines around this interface differ substantially. Examples where the single-pass interface is visible include Mem0 (Mem0 Team, 2026), MemoryOS (BAI-LAB, 2025), and A-Mem (Xu et al., 2025); our claim is about this interface, not about every component of these systems.

The failure appears when the query names the first entity in a chain but not the later bridge entities. A question may ask for an attribute of the city where X's purchase was originally cut. The first retrieval pass can find the sentence linking X to the purchase, but the city name is not in the query until that first sentence is read. A single-pass retriever therefore has no opportunity to search for the next link. We call this failure *bridge starvation*. Figure 1 shows the trace: retrieval stops after the first link, while the repair exposes the missing bridge entity and retrieves again.

To make the failure reproducible, DKMP varies chain depth $D \in \{2, 3, 4, 5\}$ and context length $N \in \{1K, 8K, 32K, 128K\}$ over 30 stories per cell. Synthetic chain entities prevent the reader from answering by prior knowledge. At N=128K, single-pass hybrid retrieval scores 0.40 / 0.47 / 0.57 / 0.13 on 2/3/4/5-hop chains, against a gold-support oracle of 1.00 / 0.93 / 1.00 / 0.76 (Table 3).

MULTIPASS repairs the specific retrieval failure without changing the retrieval index. Between passes, the LLM emits one short noun phrase ($\leq$ 6 words) naming the next entity to look up; the retriever runs again with that phrase;

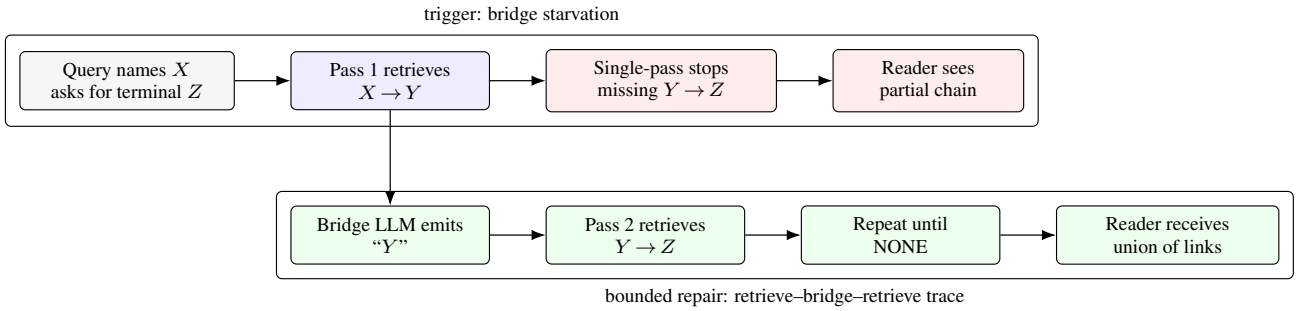

Figure 1. Bridge starvation and the MULTIPASS repair. Single-pass retrieval matches the query entity X and retrieves the first link, but bridge entities are absent from the query and therefore starve subsequent passes. MULTIPASS turns the failure into an observable trace: the LLM emits one short bridge entity, the same retriever is called again, and chain coverage can be measured after each pass.

and the final reader receives the union of retrieved chunks. The loop stops adaptively when the bridge call emits NONE. On the targeted 2/3/4-hop regime, MULTIPASS approaches oracle accuracy with 4–5 LLM calls per query and outperforms four iterative-retrieval baselines under the same reader.

The repair is intentionally bounded. It does not solve 5-hop chains: chain coverage drops to 0.75 and the reader must also compose more facts. It does not improve short-context MuSiQue, where the trigger does not fire. It also does not target LONGMEMEVAL's dominant cross-session aggregation failure, where there is no obvious next entity to retrieve.

Our contributions are:

- We define single-pass bridge starvation as an operational failure with explicit trigger and non-trigger conditions.
- We introduce DKMP, a controlled trigger that measures the failure across chain depth and context length.
- We provide trace diagnostics that separate retrieval coverage from reader composition.
- We evaluate a bounded repair with measured cost and report the cases where the repair does not apply.

2. Related Work

Several retrieval-time reasoning methods already revisit the corpus during generation, but they do not isolate bridge starvation as a failure mode. IRCoT (Trivedi et al., 2023) interleaves retrieval with one chain-of-thought sentence per step, originally using eight in-context demonstrations and a `remove_wh_words` BM25 reformulation. ReAct (Yao et al., 2023) alternates Thought / Action$_{\text{Search}}$ / Observation, Self-Ask (Press et al., 2023) decomposes the question into follow-ups, FLARE (Jiang et al., 2023) triggers retrieval on uncertainty, and HippoRAG (Gutiérrez et al., 2024) runs personalized PageRank over an OpenIE-derived passage graph. We re-implement IRCoT in both 0-shot and faithful 8-shot variants, plus ReAct, Self-Ask, and a passage-graph PPR variant of HippoRAG without OpenIE. All baselines share the same reader and retrieval index as MULTIPASS.

Agentic-memory systems differ in how they write and organize memory, so same-reader comparisons are important when evaluating the retrieval layer. Mem0 (Mem0 Team, 2026) extracts atomic facts and stores them in a vector store. MemoryOS (BAI-LAB, 2025) adds hierarchical operating-system layers. A-Mem (Xu et al., 2025) introduces note-link expansion at retrieval. We implement *Mem0-style* memory (LLM-extracted facts + cosine retrieval) and *A-Mem-style* retrieval (top-K + 1-hop note-link expansion) under the same GLM-4.7 reader as our chunk baseline (Table 7).

Long-context and multi-hop benchmarks provide useful controls but stress different mechanisms. Lost-in-the-middle (Liu et al., 2024) predicts that full-context readers degrade when relevant content appears mid-input; DKMP length curves quantify this effect for our reader. MuSiQue (Trivedi et al., 2022) is a natural multi-hop QA benchmark with short contexts, so we use it as a non-trigger control (Section 8.2). LONGMEMEVAL (Wu et al., 2025) is a long-term-memory benchmark whose dominant failure type is cross-session aggregation (Section 8.3).

3. Operational Failure and DKMP Trigger

Definition 3.1 (Single-pass bridge starvation)**.** A query q names entity X and asks for a terminal answer reachable only by traversing a chain $Y_0{=}X \to Y_1 \to \cdots \to Y_D{=}Z$. A single-pass top-$K$ retriever scores chunks against q and returns the chunk containing $Y_0{\to}Y_1$. Subsequent bridge facts ($Y_i{\to}Y_{i+1}$ for $i \geq 1$) are absent from q, so they are not retrieved. The reader answers from an incomplete chain.

Triggering preconditions. (a) chain depth $D \geq 2$; (b) context N large enough that full-context reading degrades — in

our setup, $N \gtrsim 8K$ tokens; (c) bridge entities Y_i are not lexically present in q; (d) retrieval is performed in one pass.

Non-trigger boundaries. short contexts where the reader can read everything; single-hop probes; aggregation-style questions with no bridge entity in q; settings where the reader is given gold support directly.

DKMP is built from 30 NarrativeQA stories. For each (story, D) pair with $D \in \{2, 3, 4, 5\}$, GLM-4.7 generates one chain of D needles using a directional prompt with synthetic entity names (e.g., "Vixenia Strode", "Velorum") chosen to avoid lexical collision with the source story. We validate each chain by entailment.

The contexts at $\{1K, 8K, 32K, 128K\}$ tokens are built by sampling distractor passages from *other* stories and inserting needle sentences at separate non-edge positions. The full trigger has 4 depths $\times$ 4 lengths $\times$ 30 stories = 480 contexts. We also include five single-hop probe types (lexical, paraphrase, coreference, temporal-order, and causal-direction) as sanity checks. These probes saturate at oracle for any reasonable retriever (avg. 0.97), so we omit them from the main analysis.

GLM-4.7 (T=0, `thinking=off`) reads the retrieved context and answers in one short sentence. A separate fresh GLM-4.7 call judges YES/NO against the gold reference, so all reported accuracies are LLM-judge accuracies. Synthetic entities remove a world-knowledge shortcut: an entity such as "Crystal of Mar" has no Wikipedia article, so the only path to the answer is retrieval and chain traversal.

4. Bounded Repair: MULTIPASS

Given query q and a chunked corpus, MULTIPASS:

1. Initialize $\mathcal{C} \leftarrow \emptyset$, $q_0 \leftarrow q$.
2. For $i \in \{0, \ldots, D_{\max}-1\}$:
   - Retrieve $\mathrm{top}_K(q_i)$, dedup against $\mathcal{C}$, append.
   - Bridge prompt: "Given q and $\mathcal{C}$, name the next intermediate entity to look up ($\leq$ 6 words), or NONE if done."
   - Parse $b_i \leftarrow$ LLM. If NONE or empty, break.
   - $q_{i+1} \leftarrow b_i$.
3. Final reader call: "Answer q using $\mathcal{C}$."

For a chain that stops adaptively after $D^\star \leq D_{\max}$ passes, MULTIPASS issues $D^\star$ retrievals, $D^\star$ short bridge LLM calls (the last may emit NONE), and one final reader call, for $D^\star$+1 LLM calls in total. Each bridge call has output capped at 30 tokens. Only the final reader sees the full retrieved context.

Table 1. Chain coverage and answer accuracy for MULTIPASS at N=128K (n=30/cell). For 2/3/4-hop, coverage is essentially complete and the small accuracy gap to oracle is reader-composition error on cases where all support is retrieved. For 5-hop, coverage itself drops to 0.75 — retrieval is the larger error source.

Hop	Avg cov.	Acc	Oracle acc	Cov= 1, correct	Cov< 1, correct
2	1.00	1.00	1.00	30/30	—
3	0.98	0.90	0.93	26/30	1/30
4	0.97	0.97	1.00	27/30	2/30
5	0.75	0.53	0.76	11/30	5/30

This design targets bridge starvation as defined in Section 3. The LLM names the next bridge entity, and the retriever locates the next link. The method does *not* directly solve reader composition, because the reader still has to chain D facts at the end. It also does not solve aggregation queries when there is no obvious bridge entity to emit.

5. Trace Diagnostics

We log every retrieval, every bridge-LLM call, and every reader call. The trace exposes failure modes that final-answer accuracy hides.

5.1. Schematic minimal trace

Figure 1 shows the minimal trace. On a 3-hop chain $X \to Y \to Z \to W$, single-pass retrieval stops after retrieving $X \to Y$; MULTIPASS emits "Y", retrieves $Y \to Z$, emits "Z", retrieves $Z \to W$, then stops with NONE. Appendix A gives a real DKMP trace.

5.2. Chain-coverage diagnostic

For each (D, method) cell at N=128K we measure Chain-FactRecall = fraction of gold needle sentences that appear in any retrieved chunk after the method finishes. Table 1 gives the breakdown.

5.3. Bridge-entity accuracy

When MULTIPASS emits a non-NONE bridge, does it name a gold chain entity? Table 2 shows the rate is 0.95–1.00 across all four depths. The 5-hop drop in chain coverage is therefore not bridge-LLM drift; it is that retrieving on a correctly-named bridge entity sometimes fails to return the chain link, presumably because the bridge entity recurs in distractor passages and dilutes the top-K.

6. Results: Trigger, Repair, and Boundary

The targeted regime is 2/3/4-hop reasoning at N=128K (Figure 2a). In this regime, MULTIPASS is within 3pp of or-

Table 2. Bridge-LLM accuracy: of all non-NONE bridge phrases emitted, fraction that mention a gold chain entity. Computed on the same n=30/cell at N=128K used in Tables 3 and 1.

Hop	Bridges emitted	NONE rate	Hits gold entity
2	88	30/88	55/58 (0.95)
3	110	25/110	82/85 (0.96)
4	116	28/116	88/88 (1.00)
5	91	29/91	61/62 (0.98)

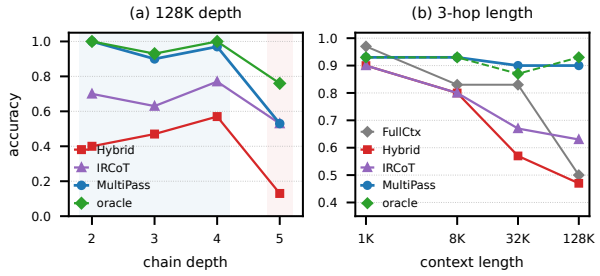

Figure 2. Audited DKMP results used in the main text. (a) At N=128K, MULTIPASS repairs 2/3/4-hop bridge starvation and approaches the gold-support oracle; 5-hop is a boundary case. (b) On 3-hop chains, the failure is context-length conditional: single-pass Hybrid-RRF and FullCtx degrade as context grows, while MULTIPASS remains close to oracle through 128K.

acle on every depth. Compared with IRCoT 8-shot faithful, the gap is +30pp on 2-hop, +27pp on 3-hop, and +20pp on 4-hop. The corresponding gaps over ReAct are +37 / +37 / +44pp, and the gaps over Self-Ask are +70 / +57 / +67pp. The confidence intervals do not overlap on 2-hop and touch only at the endpoint on 3-hop.

A natural concern is that 0-shot IRCoT may understate the original method by omitting demonstrations. We therefore include an 8-shot faithful IRCoT variant with eight hand-crafted DKMP-style demonstrations, `remove_wh_words`, reasoning-sentence skip for next-query selection, and `max_steps=8`. Both IRCoT variants lose to MULTIPASS on 2/3/4-hop by 20–30pp. The 8-shot variant is worse than 0-shot on 2-hop and 5-hop, likely because the demonstrations bias the model toward their chain structure even when the synthetic entities differ.

6.1. Length Curve: the Trigger is Context-Length Conditional

Figure 2b and Table 4 show that the failure is not a generic multi-hop deficit. At 1K, FullCtx, Hybrid-RRF, IRCoT, and MULTIPASS all score at least 0.90 on 3-hop chains except ReAct. The gap opens as the context grows: Hybrid-RRF drops to 0.57 at 32K and 0.47 at 128K, while MULTIPASS stays at 0.90. This pattern supports the trigger definition in Section 3: bridge starvation requires both a multi-hop chain and a context long enough that simply reading everything is

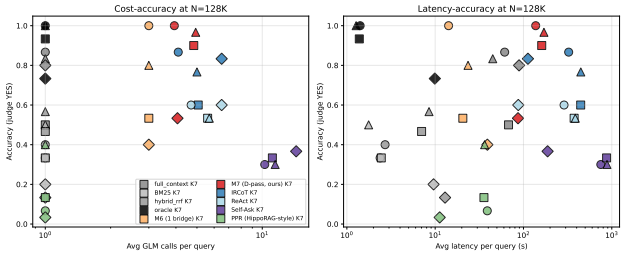

Figure 3. Cost–accuracy points on multi-hop chains at N=128K. Each marker is a (method, D) cell. Among iterative-retrieval methods, MULTIPASS is upper-left on both axes. FullCtx has fewer LLM calls and lower latency but uses ∼22× more reader tokens and is less accurate in the targeted regime.

unreliable.

7. Cost and Repair Budget

The latency gap between MULTIPASS (156s) and IRCoT 8-shot (470s) is ≈3×. The gap arises because MULTIPASS's bridge call output is capped at 30 tokens while IRCoT emits a full CoT sentence per step; asynchronous queueing under load amplifies the per-step difference. Self-Ask is the most expensive because it issues both a decomposition call *and* a sub-question answer call per follow-up. We report calls, tokens, and latency; we do not convert to dollar cost because the provider's token rates may change.

8. Boundary and Negative Results

8.1. 5-hop: a boundary, not a tuned success

MULTIPASS 5-hop accuracy is 0.53 [0.37, 0.70], tied with IRCoT 8-shot and below FullCtx (0.80) and IRCoT 0-shot (0.84). MULTIPASS does not solve 5-hop.

Table 1 shows that on 5-hop MULTIPASS reaches full chain coverage on only 12 of 30 cases (40%), versus 27–30 of 30 on 2/3/4-hop. Of the 12 full-coverage cases, the reader is correct on 11 (0.92); of the 18 partial-coverage cases, only 5 are correct (0.28). The aggregate gold-support oracle accuracy of 0.76 also shows that 5-step composition is difficult even when the support is provided directly.

The 5-hop boundary therefore has two causes. The larger one in our setup is retrieval: the bridge LLM is accurate (0.98 hit rate, Table 2), but a correctly named bridge entity can recur in distractor passages and fill the top-K window with non-gold matches. The second cause is reader composition: even with gold support, 5-step reasoning is near the reader's ability boundary. MULTIPASS targets bridge starvation, not these two later failure modes.

Table 3. Multi-hop accuracy at N=128K, GLM-4.7 LLM-judged (n=30/cell). Bootstrap 95% CIs in brackets (10K resamples) for the three methods most discussed in the text. MULTIPASS matches oracle on 2-hop and is within 3pp of oracle on 3/4-hop. At 5-hop, MULTIPASS no longer wins (Section 8.1).

Method	2-hop	3-hop	4-hop	5-hop
FullCtx (8K trunc)	0.87 [0.73,0.97]	0.50 [0.33,0.67]	0.83 [0.70,0.97]	0.80 [0.63,0.93]
BM25 top-10	0.33	0.33	0.50	0.20
Hybrid-RRF top-10	0.40	0.47	0.57	0.13
PPR (HippoRAG-style)	0.07	0.13	0.40	0.03
IRCoT 0-shot	0.87	0.60	0.77	0.84
IRCoT 8-shot faithful	0.70 [0.53,0.87]	0.63 [0.47,0.80]	0.77 [0.60,0.90]	0.53 [0.37,0.70]
ReAct	0.60	0.53	0.53	0.60
Self-Ask	0.30	0.33	0.30	0.37
MULTIPASS (ours)	**1.00 [1.00,1.00]**	**0.90 [0.80,1.00]**	**0.97 [0.90,1.00]**	0.53 [0.37,0.70]
Oracle (gold needles)	1.00	0.93	1.00	0.76

Table 4. 3-hop accuracy across context lengths. The single-pass failure emerges in the 8K–32K band; MULTIPASS is the only method that holds at N=128K.

Method	N=1K	N=8K	N=32K	N=128K
FullCtx	0.97	0.83	0.83	0.50
Hybrid-RRF	0.90	0.80	0.57	0.47
IRCoT 8-shot	0.90	0.80	0.67	0.63
ReAct	0.60	0.63	0.37	0.53
MULTIPASS	**0.93**	**0.93**	**0.90**	**0.90**
Oracle	0.93	0.93	0.87	0.93

Table 5. Per-query cost on 2/3/4-hop at N=128K (averaged over 90 queries). MULTIPASS is the cheapest among iterative-retrieval methods on both LLM calls and end-to-end latency, and the highest-accuracy method overall except oracle. FullCtx has lower latency than MULTIPASS but uses $\sim$22$\times$ more reader tokens.

Method	Calls	Tokens	Lat. (s)	Acc
FullCtx	1.0	128,007	65.8	0.73
Hybrid-RRF	1.0	2,044	7.8	0.48
PPR	1.0	5,600	37.0	0.20
MULTIPASS	**4.6**	5,650	**156**	**0.96**
IRCoT 8-shot	5.5	6,500	470	0.70
ReAct	5.3	2,378	350	0.55
Self-Ask	11.0	2,689	836	0.31
Oracle (gold)	1.0	41	3.6	0.98

8.2. Short context: the trigger does not fire

The MuSiQue tie is non-trigger evidence: DKMP's length curve places the trigger boundary in the 8K–32K range. MuSiQue at 5K is below it, so single-pass retrieval and FullCtx work and MULTIPASS has nothing to repair. We do not claim improvement where the failure does not exist.

8.3. LongMemEval: same-reader fairness check

LONGMEMEVAL's dominant failure type is cross-session aggregation (e.g., "How many model kits have I bought?"), not chain traversal. MULTIPASS's bridge mechanism does not help because there is no obvious next-entity to retrieve. We include LONGMEMEVAL as a same-reader fairness check on agentic-memory baselines.

Table 6. MuSiQue val (n=300 balanced 2/3/4-hop, GLM-4.7 LLM-judged). Each MuSiQue question has 20 short paragraphs, total context $\sim 5K$ characters. Below the trigger preconditions in Section 3; single-pass retrieval, FullCtx, and MULTIPASS cluster around 0.49.

Method	All	2-hop	3-hop	4-hop
MULTIPASS	0.490	0.550	0.560	0.360
ReAct	0.493	0.460	0.550	0.470
FullCtx	0.493	0.530	0.520	0.430
IRCoT	0.473	0.580	0.500	0.340
Hybrid-RRF	0.387	0.440	0.450	0.270
Self-Ask	0.223	0.400	0.170	0.100
Oracle (supporting)	0.753	0.760	0.770	0.730

Mem0-style wins on KU and Temp (where extraction filters chitchat) but loses on SA/SU (where the reader needs the original surface form). Aggregate Hybrid-RRF wins by 5.2pp. Read this as a fairness reminder: published agentic-memory numbers often change both the memory representation and the reader; same-reader comparisons are needed to isolate the memory-side contribution.

8.4. Cross-reader check

9. Limitations

The main DKMP cells have n=30, so several confidence intervals remain wide; the paper should be read as a reproducible trigger and diagnostic study, not as a final leaderboard. We test GLM-4.7 fully and GPT-4o-mini on the cross-reader slice, but not Claude- or Llama-class readers. The trigger uses synthetic-entity needles embedded in real fictional contexts; this prevents world-knowledge shortcuts, but it does not replace a natural long-context multi-hop benchmark. Finally, the HippoRAG-style PPR baseline uses chunk-cosine edges rather than a fully faithful OpenIE

Table 7. LONGMEMEVAL N=500, GLM-4.7 LLM-judged. This is a same-reader fairness check, not a claim against the published Mem0 or A-Mem systems. Under the fixed reader used here, chunk retrieval over raw conversation transcripts is strongest overall; Mem0-style wins KU and Temp but loses SA/SU, showing that memory representation and reader choice must be separated.

Method	Overall	KU	MS	Temp	SA	SU
FullCtx (8K trunc)	0.360	0.128	0.120	0.195	0.982	0.814
Oracle (gold sessions)	0.384	0.141	0.105	0.203	0.982	0.957
BM25 chunks K=10	0.612	0.692	0.481	0.398	0.982	0.929
Dense chunks K=10	0.608	0.603	0.579	0.316	0.982	0.929
Hybrid-RRF K=10	**0.614**	0.667	0.526	0.361	0.982	0.943
Mem0-style	0.562	0.718	0.481	0.541	0.464	0.714
A-Mem-style	0.520	0.603	0.436	0.474	0.482	0.657

Table 8. Cross-reader: MULTIPASS at N=128K with GLM-4.7 vs. GPT-4o-mini (n=30). Single-pass Hybrid-RRF (GLM-4.7) shown for reference. The repair transfers across the two readers we tested.

Hop	GLM	GPT	Hybrid
2-hop	1.00	0.97	0.40
3-hop	0.90	0.80	0.47
4-hop	0.97	0.90	0.57

graph, and our IRCoT 8-shot demonstrations are DKMP-style rather than unchanged HotpotQA prompts because the synthetic entity vocabulary is incompatible. We use a separate LLM call for judging, but the judge is still an automatic judge from the same model family as the reader; a human-audited subset would make the error analysis stronger.

10. Conclusion

We isolated and operationalized one failure of single-pass agentic-memory retrieval at long context: bridge starvation. DKMP reproduces the failure under a fixed protocol and exposes the missing links in the trace. MULTIPASS is a minimal repair; it works in the targeted regime, has measured cost, and has explicit boundaries where it does not work. We hope the trigger and the trace diagnostic are reusable.

References

BAI-LAB. MemoryOS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. Oral presentation.

Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M., and Su, Y. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, 2024.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Mem0 Team. State of AI agent memory 2026. https://mem0.ai/blog/state-of-ai-agent-memory-2026, 2026.

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., and Lewis, M. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023.

Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., and Yu, D. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025.

Xu, W., Zhao, M., Wang, L., Zhang, L., Liu, Y., and Wang, Y. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, 2025.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

A. Concrete trace example

A real DKMP 3-hop instance from our logs.

Story: `dkmp_1b82b150` (NarrativeQA-derived distractor, N=128K).

Chain: "Hartin Whatley purchased the Sapphire of Telemar." → "The Sapphire of Telemar was originally cut in the city of Bryandyl." → "Bryandyl is renowned for saffron-laced cordials."

Question: "What is the most famous export of the city where Hartin Whatley's purchase was originally cut?"

Single-pass top-10 hybrid retrieval: returns the Hartin Whatley needle and 9 distractor chunks; *misses* both Bryandyl needles. Reader: "unknown / unspecified."

MULTIPASS:

- Pass 1: same retrieval as single-pass.
- Bridge LLM, given Hartin's needle: "Sapphire of Telemar."
- Pass 2: retrieves the Bryandyl-cutting needle.
- Bridge LLM: "Bryandyl."
- Pass 3: retrieves the Bryandyl-cordials needle.
- Bridge LLM: "NONE."
- Reader: "saffron-laced cordials." Judge: YES.

B. Ablations

All on 3-hop chains, N=128K, n=30, lenient string match (faster than the GLM judge for sweeps; trends are robust to switching to the GLM judge on a 30-row spot check).

- MULTIPASS max-passes sweep: 1 / 2 / 3 / 4 / 5 / 6 / 7 → 0.43 / 0.50 / 0.93 / 0.90 / 0.87 / 0.83 / 0.90. Match-or-exceed-chain-depth is necessary; beyond that, the LLM-NONE early-stop kicks in and accuracy plateaus.
- Bridge-prompt length cap (words): 3 / 6 / 12 / 25 → 0.73 / 0.70 / 0.77 / 0.67. Caps from 3 to 12 are within noise; the 25-word "full sentence" variant degrades.
- Top-K per pass: 3 / 5 / 10 / 20 → 0.67 / 0.63 / 0.73 / 0.77. Diminishing returns appear above K=10.

C. Reproducibility

- All code, prompts, seeds, and per-cell predictions released alongside the paper (see supplementary).
- Total compute: ∼12 hours wall-time on a single laptop (M-series MPS for BGE-M3 chunk embeddings).
- Each method uses the same retrieval index (BGE-M3 + BM25 fused via reciprocal rank fusion) and the same reader prompt; the only difference is the bridge or reasoning loop.
- All judges use a fresh GLM-4.7 instance with a YES/NO prompt disjoint from the reader prompt.
- Bootstrap CIs use 10K resamples with seed 42.