SIGDIAL 2026
Long Paper #107

Bounded Conversational Memory with Hybrid Retrieval and Evidence Highlighting for Multi-Session Dialogue Systems

Manasa Mudunuri[1] Chandra Kiran Guntupalli[1] Shiraj Pokharel Murray Patterson[1] Mario M. Kubek[1]
[1] Department of Computer Science, Georgia State University, Atlanta, GA

THE FINDING Same retrieval. Substantially better answers.

Holding the retriever fixed, we changed only how retrieved context is **organized and presented** to the model, and answer accuracy rose by **+24.6 pp on LongMemEval** and **+15.9 pp on multi-hop LoCoMo**, with 4.8× fewer stored units.

1 · The Problem: Dialogue Systems Forget

- **Volatile memory:** when a session ends, everything the user explained is erased; the next session starts from a blank slate.
- **Context overload:** keeping every turn floods the context window, dilutes attention, and answers get worse ("lost in the middle").

A user who spent an hour explaining project requirements returns the next day to a system that remembers nothing.

2 · Problem Formulation

Goal. A memory layer that is (a) bounded, with capped storage and principled eviction, and (b) accurate: the model answers correctly from what is recalled.
The field's default assumption. Retrieval quality is the primary lever for accuracy.
We revisit that assumption. Does the presentation of retrieved content matter just as much?

3 · Related Systems All Bet on Retrieval

MemGPT OS-style hierarchical memory (~16,900 tokens per query)

Mem0 atomic fact extraction, but stores without bound

MemoryOS bounded via heat-based eviction

SimpleMem entropy-gated compression

GAM / A-MEM graph memory, never tested under a storage budget

None isolates **presentation quality** as an accuracy driver. CPP is designed to test exactly that.

4 · The Controlled Test

Two systems share the identical retrieval stack (BGE-M3 + BM25 + RRF fusion + MiniLM reranking) and differ only in structure:

C2 · Flat Retrieval
unbounded baseline
- every raw turn (~601 / conv.)
- no segmentation
- no evidence highlighting

C3 · CPP (ours)
bounded, structured
- topical segments (~124 / conv.)
- dedup + 400-segment cap
- evidence highlighting + typed prompts

If retrieval metrics come out identical but accuracy does not, **the difference must come from organization and presentation**: the question prior systems never isolated.

5 · Hypotheses

H1 Bounded segments match or beat unbounded flat retrieval with far fewer stored items.

H2 Largest gain on multi-hop questions: segments keep topical context together.

H3 Possible small deficit on diffuse open-domain questions.

H4 Identical Recall@k ⇒ the advantage is presentation-driven, not retrieval-driven.

6 · Method: The Context Persistence Protocol

A modular memory layer of four composable phases; every component is independently replaceable, dataset- and model-agnostic.

1 Segmentation
Group turns into topically coherent 3–10-turn segments (boundary when cosine < δ = 0.65). Coreference: "I" → speaker name at ingest. Compresses the store 4–5×.

2 Indexing
Dual index per segment: BGE-M3 dense (dim 1024) + sparse BM25. Near-duplicates discarded; hard budget cap B = 400 segments, FIFO eviction.

3 Retrieval
Hybrid RRF fusion of dense + BM25 scores, temporal (+0.10) and entity (+0.05) boosts, then cross-encoder reranking of top-50 → top-k.

4 Evidence Highlighting
The best line of each retrieved segment is appended as "Evidence: …", giving full context plus an explicit pointer at a high-attention position.

5 Generation
Segments + session datetimes + question-type-aware prompt → local LLM. **Model-agnostic.**

7 · What the Model Actually Sees

```
[Session 12 · 2024-03-08]
Alice: I finally joined that gym near work…
Bob: Nice! How often are you going?
Alice: Three mornings a week, before my shifts.
Evidence: "Alice goes to the gym three mornings a week."
```

Narrative coherence is preserved; the single most informative sentence is made explicit, countering the "lost-in-the-middle" failure (Liu et al., 2024).

8 · Experiments

LoCoMo
human–human · MC-10
1,986 questions · 10 conversations · ~600 turns each · single-hop, multi-hop, temporal, open-domain, adversarial

LongMemEval
human–AI · free-form
500 questions · ~40 sessions (~115K tokens) each · adds knowledge-update + abstention · LLM-judged

Conditions. C1 no memory · C2 unbounded flat retrieval · C3 full CPP.
Setup. Qwen2.5-14B-Instruct (FP16), greedy decoding, seed 42, single RTX 4090 (24 GB). k = 10 (LoCoMo) / 15 (LongMemEval), tuned on dev split.

Paper, code & prompts
github.com/MudunuriManasa/Bounded-Conversational-Memory-with-Hybrid-Retrieval-and-Evidence-Highlighting
mmudunuri3@student.gsu.edu
cguntupalli1@student.gsu.edu

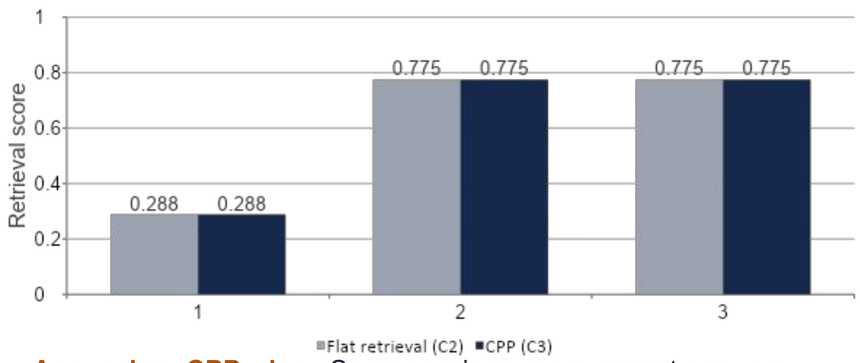
9 · Results

81.4%
LoCoMo accuracy (+3.4 pp vs. flat retrieval)

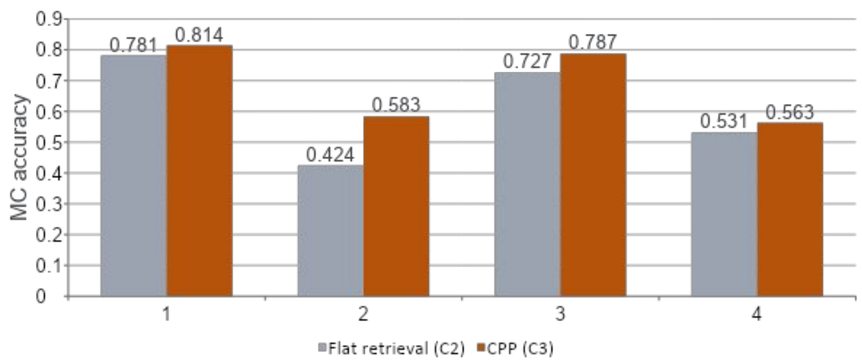
+24.6 pp
LongMemEval accuracy (0.606 vs. 0.360)

4.8×
fewer indexed units (124 segments vs. 601 turns)

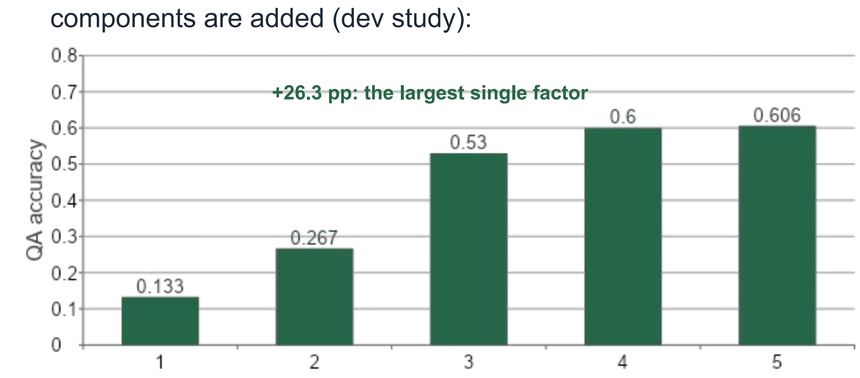
Finding the sessions: tied. Recall@k, Hit@k, and MRR are identical on LoCoMo; on LongMemEval, CPP's Recall@15 is slightly lower (0.881 vs. 0.919).



Answering: CPP wins. Same sessions, more correct answers (LoCoMo MC accuracy; gains $p < 0.05$, McNemar). Multi-hop: +15.9 pp (H2 ✓).



Where the gain comes from. LongMemEval accuracy as CPP components are added (dev study):



10 · Discussion and the Hard Question

- **Retrieval is matched, so the gain is presentation-driven:** segments carry richer semantic fingerprints, and the evidence line sits at a high-attention position.
- At k = 15, fifteen raw turns of noise sink the flat baseline; CPP stays stable.
- **Ask us the hard question:** CPP's prompt is ~3.4× longer at equal k. A token-matched baseline (C2 at k≈34) is our decisive next test; the 400-segment cap never fires (27–31% used), so boundedness is a production guarantee, not a

11 · Conclusion

★ **Take-home: if you build LLM memory, invest in how retrieved context is organized and presented, not only in what is retrieved.** One appended evidence sentence was worth more than any retrieval upgrade we tried, while storing 4.8× less.

References

[1] Maharana et al. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. ACL.
[2] Wu et al. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. ICLR.
[3] Packer et al. 2023. MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560.
[4] Chhikara et al. 2025. Mem0: Scalable Long-Term Memory for AI Agents. ECAI.
[5] Kang et al. 2025. Memory OS of AI Agent. EMNLP.
[6] Liu et al. 2026. SimpleMem: Efficient Lifelong Memory for LLM Agents. arXiv:2601.02553.
[7] Liu et al. 2024. Lost in the Middle: How Language Models Use Long Contexts. TACL.
[8] Pan et al. 2025. SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents. ICLR.
[9] Cormack et al. 2009. Reciprocal Rank Fusion. SIGIR.