

Long-term memory for AI assistants: capabilities and research

Loxley Labs. What the memory system does today, what it scored, and the research directions being worked on. October 16, 2026.

73.2

out of 100 on 1,000 sealed questions from a real seven-month archive; standard retrieval 67.5, no memory 5.9

76.2

on the public LoCoMo benchmark's 1,540 standard questions; standard retrieval 73.7

107

wrong answers out of 1,000, against 143 for standard retrieval

9 s

from a message to its memory: 4.4 s search, about 4.6 s selection

Summary

The system gives an AI assistant a long-term memory of its past conversations. On 1,000 sealed questions written from one person's real conversations, March to October 2026, it scored 73.2 out of 100 against 67.5 for standard retrieval and 5.9 with no memory, with the same local answering model each time: 166 questions better, 92 worse, $p = 0.0000048$. On LoCoMo, a public benchmark of generated conversations, it scored 76.2 on the 1,540 questions most published work reports and 81.0 on all 1,986.

Search, selection and answering run on one desktop PC with two consumer graphics cards. Two research directions are being worked on: memory held as model weights, and a small network that connects several models.

What the system does today

CAPABILITY	MEASURED
Answers questions about the user's life and the people in it	86.3 out of 100, against 77.0 for standard retrieval (244 questions)
Recalls what was decided and when	78.3 against 71.7 (210 questions)
Places events in time	LoCoMo temporal questions: 70.9 against 64.6 (321 questions)
Says "I don't remember" instead of inventing	50 of 50 questions about things never discussed
Delivers memory during a conversation	about 9 s from message to memory; the answer starts at once
Keeps up with new conversations	101 new conversations added in the first nightly update, 16 October

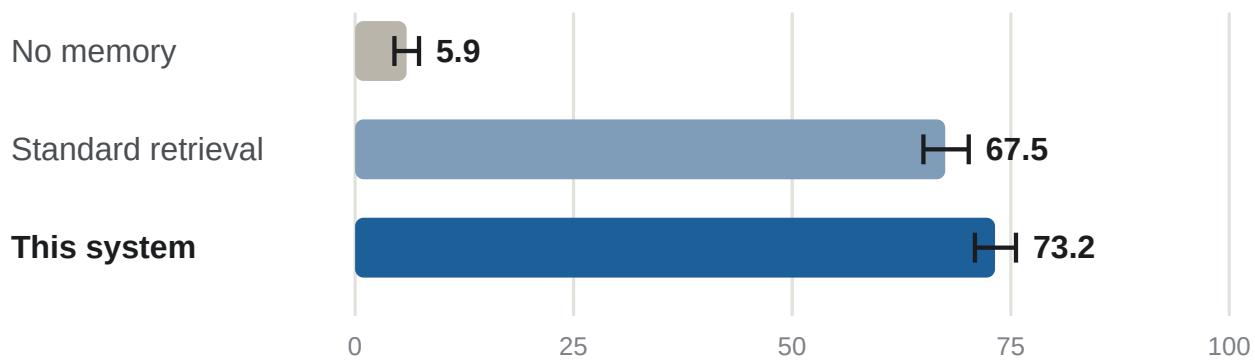
Measured results

1,000 questions from a real archive

The question set was sealed by fingerprint before any system answered. 20 AI writers wrote the questions from original passages, 5 reviewers checked each one against its source, 150 were traps, and 21 blind graders marked the answers without knowing which system wrote which.

Overall score, out of 100

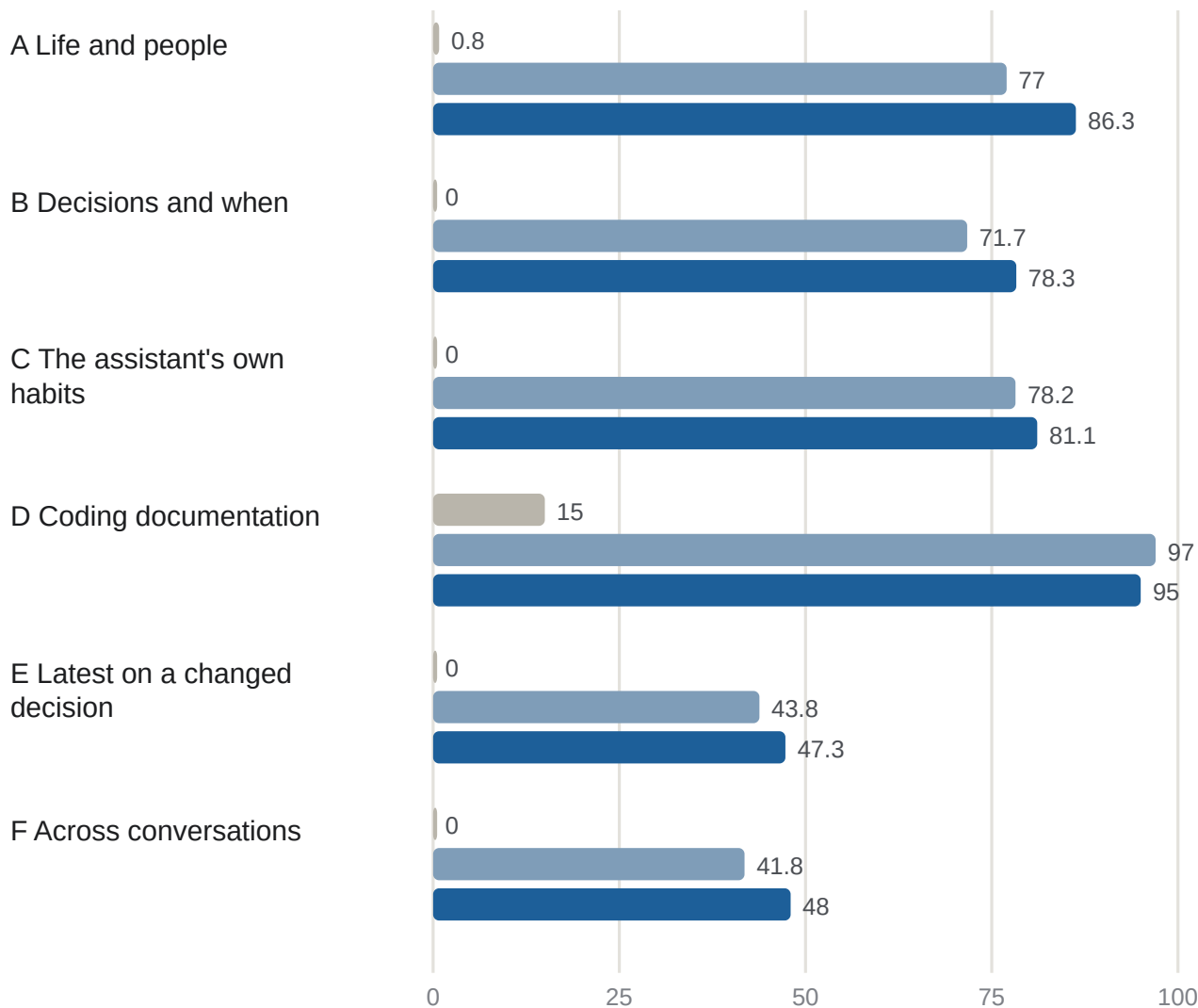
Black line: 95% confidence interval.



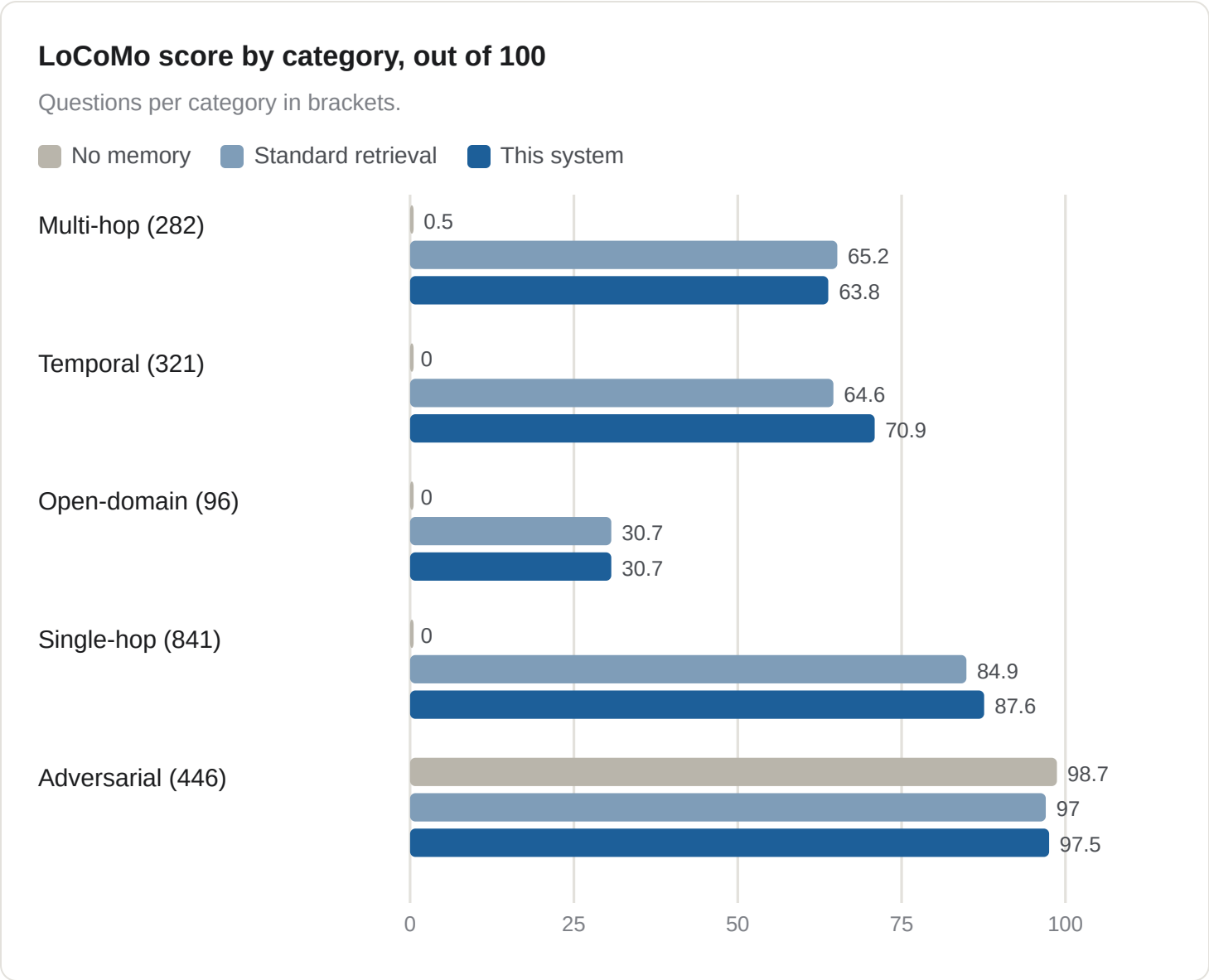
Score by category, out of 100

The two hardest categories, the latest on a changed decision and answers spread over several conversations, score under 50 for both methods.

■ No memory ■ Standard retrieval ■ This system



LoCoMo, the public benchmark



On all 1,986 questions: 81.0 against 78.9 for standard retrieval (192 better, 154 worse, $p = 0.047$). On the four standard categories: 76.2 against 73.7 ($p = 0.098$). Independent runs of published systems under one judge range from 67 to 76 ; our graders are not the judge those runs used, and about 6.4% of LoCoMo's answer key is wrong, so a gap of under 3 points on this benchmark is within its noise.

The live loop

STEP	TIME	ACCURACY
Search and rerank	4.4 s	the answer passage is among the candidates 98% of the time
Selection of the passages that answer	about 4.6 s	keeps the answer passage 91 to 94% of the time; 86% without the reranking step

On an earlier 200-question test the full loop scored 89.2 against 87.5 for standard retrieval.

How it keeps up

The memory is brought up to date every night with the day's new conversations. The first update of this kind, on 16 October, added 101 conversations.

Research directions

Memory held as weights

Part of the memory would be learned into the model itself, so the assistant knows a topic directly as well as looking it up.

A small network connecting several models

A small network would join several different models into one system that works as a whole.

Limits

- The 1,000-question test uses one person's history; other users' histories are untested.

- Questions were written and graded by AI models, with review and audit, not by people.
- The latest on a changed decision (47.3) and answers spread over several conversations (48.0) are the weakest categories.
- LoCoMo used different graders from published work; on its adversarial questions no memory scores 98.7 by abstaining.
- The nightly update currently uses cloud AI models as well as local ones.
- Neither research direction has a result reported here.

Statistics: bootstrap 95% confidence intervals over questions; two-sided sign tests on questions where two conditions received different grades. Figures for other systems are from their authors or from independent runs, as cited, and were not re-measured by us.

Further detail is available under a non-disclosure agreement.