Cook Memory: Finding the Right Context Across Conversations

Austin Worthington

Client Acquisition.io

trycook.ai/research

Abstract

Cook Memory combines planned keyword searches, semantic retrieval, reciprocal rank fusion, cross-encoder reranking, and answer-oriented ordering. On the recorded LongMemEval-S evaluation, the final pipeline reached 98.5% Recall@5, 98.9% Recall@10, and 90.2 mean reciprocal rank on a 0-100 scale. A single keyword-query baseline recorded 83.6% Recall@5. (Cook, 2026)

The measured improvement is 14.9 percentage points in top-five recall within Cook's evaluation. The research asks a practical question: can the agent retrieve the earlier exchange that contains the evidence needed now, despite unrelated conversations around it?

1 Published comparison

agentmemory reports 95.2% Recall@5, 98.6% Recall@10, and 88.2 MRR. Cook's recorded values are higher on all three measures. However, Cook scores answer-bearing excerpts on 476 questions; agentmemory reports session retrieval on 500 questions. These are separate evaluation setups, not a matched-harness experiment. (Cook, 2026; agentmemory, 2026)

This report concerns retrieval. Finding relevant evidence does not mean the final answer is correct. Answer-generation results, variability, and business-context experiments are reported separately in Section 5. No new benchmark execution was performed for this publication.

2 Architecture: a notebook backed by the original conversation

Long-term memory has two different jobs. It must keep useful context compact enough to carry forward, and it must find the original evidence when a later question needs detail. A short note can retain a standing preference, but it may omit a date or condition. The conversation archive retains that detail, but searching all of it indiscriminately can bury the useful passage. Cook combines these roles rather than asking a single summary to serve every future question. (Cook, 2026)

2.1 Preserve the difference between a decision and a suggestion

An exchange can contain a user's instruction, a fact about the user, and an assistant's proposal. Treating all three as equivalent can change what the system later believes was decided. Cook's memory design distinguishes user directions, durable personal context, and the state of work. It also retains a relationship between important directions and their supporting conversation. (Cook, 2026)

For example, "we approved this budget" and "we could consider this budget" may mention the same amount while having different consequences. A useful memory should preserve that distinction when the amount is recalled. This is an illustrative example of attribution, not an

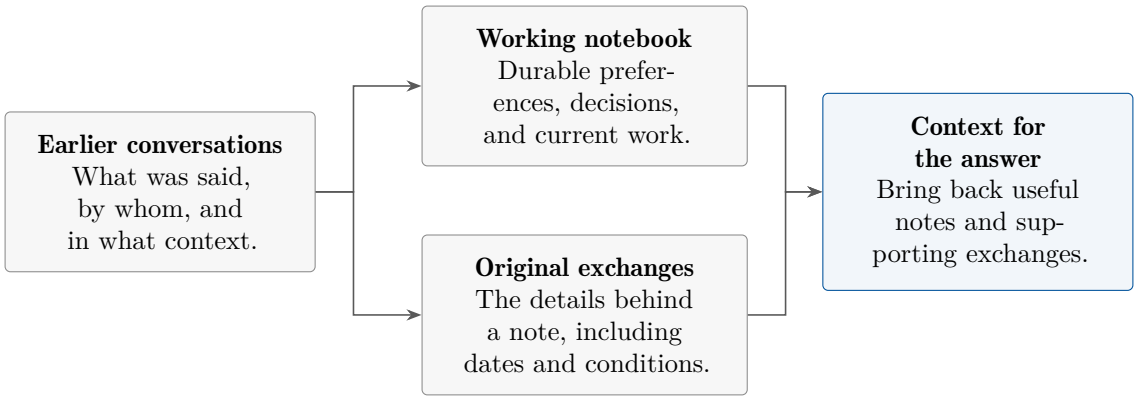

Figure 1: A plain-English view of Cook Memory: a working notebook backed by the original conversations. Notes preserve continuity; retrieved exchanges supply the evidence a particular question needs.

additional measured result. The reported retrieval scores measure finding annotated evidence; they do not by themselves measure whether every stored decision is attributed correctly.

2.2 Allow new information to replace old context

A memory system that only accumulates facts can return the wrong answer while accurately quoting an old conversation. Budgets change, plans are revised, and work moves forward. The architectural requirement is therefore not simply retention but *continuity through change*: keep the context that still applies and recognize when a newer statement supersedes it. (Cook, 2026)

The original exchange remains useful even when a compact note changes. It explains when a value applied and what conditions accompanied it. This is why an archive and a notebook are complementary: the notebook provides a current working view, while the archive allows the agent to reconstruct the context of a specific question.

2.3 Search for meaning as well as matching words

People rarely repeat the exact wording of an earlier conversation. A question about "what we can spend" may need an exchange that discussed a "monthly budget." Word matching is useful for exact names, numbers, and unusual terms; meaning-based search helps when the phrasing changes. The recorded retrieval research combines these complementary signals. (Cook, 2026)

The benefit is coverage. A later ranking step cannot promote an answer-bearing passage that was never retrieved. Broader discovery gives the system more plausible evidence to consider before it narrows the result set. This does not mean presenting more and more history to the answering model: discovery and selection have different jobs.

2.4 Choose evidence for the question, not merely the subject

A passage can be highly relevant to a topic and still fail to answer the question. A discussion of advertising budgets may be relevant to "what amount did we approve?" without containing any approved amount. Cook's final evidence selection emphasizes whether a passage supplies the requested information. The distinction is between *about the topic* and *useful for the answer*. (Cook, 2026)

2.5 Keep an exchange together when context carries the meaning

A short answer such as "yes, that one" can be meaningless without the question before it. Likewise, an assistant's recommendation can depend on a constraint in the user's preceding turn. Cook's scored retrieval unit retains the user/assistant exchange, giving the selected evidence more of the context needed to interpret it. The research explored finer-grained alternatives, but additional fragmentation did not consistently improve ranking. (Cook, 2026)

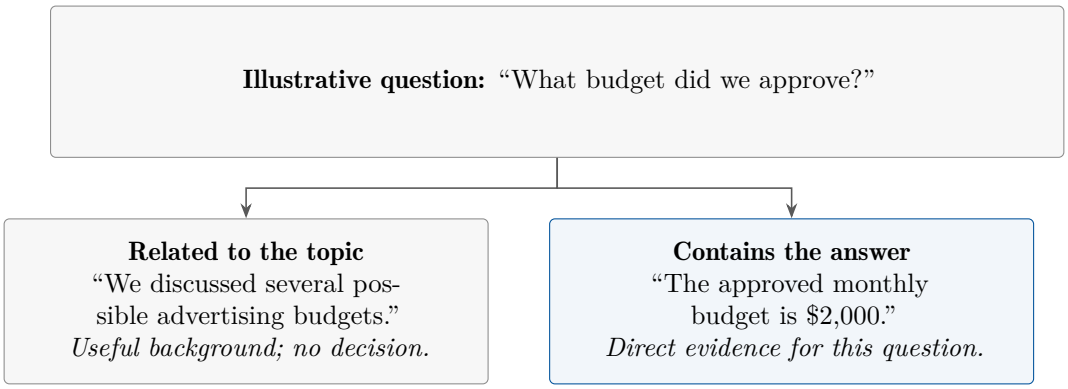

Figure 2: Why topical relevance is insufficient. Both passages concern budgets, but only one resolves the question. The sentences and amount are invented for explanation and are not benchmark examples.

These ideas describe what the architecture is trying to preserve: attribution, current context, complementary search coverage, and answer-bearing evidence. Internal prompts, ranking policies, retrieval budgets, and tuning configurations are outside the scope of this report.

3 Evaluation Methodology and Results

Metric	Cook Memory	agentmemory
Recall @ 5	98.5%	95.2%
Recall @ 10	98.9%	98.6%
MRR (x100)	90.2	88.2
Evaluation cohort	476 answerable questions	500 questions
Retrieved unit	User/assistant exchange	Conversation session

Table 1: Retrieval results from separate Cook and agentmemory evaluation setups (agentmemory, 2026).

3.1 Metric definitions

Recall@K is the share of scored questions with at least one relevant result among the first K retrieved results. It does not require retrieving every fact needed for a multi-session answer. MRR averages the reciprocal rank of the first relevant result, counting a miss as zero. A first-place result contributes 1; a second-place result contributes 0.5. This report multiplies MRR by 100 for readability.

3.2 Evaluation unit

Cook evaluates one excerpt per user/assistant exchange, preserving its date and using the dataset's answer-bearing annotations to establish relevance. Questions without a gold-bearing excerpt are excluded. The recorded cohort is 476 of the full 500 LongMemEval-S questions. The metric is computed from these annotations, without an answer-generation judge. (Cook, 2026)

3.3 Comparison boundary

agentmemory checks whether any gold session appears in its top-K session results. Its published method uses BM25 plus vector search with no LLM in the retrieval loop. Cook includes model-based query planning and final ordering. Differences in granularity, cohort, and compute mean the numerical gaps are descriptive; they do not isolate the effect of a particular architecture. (agentmemory, 2026)

Recorded retrieval stage	R@5	R@10	MRR
Single keyword query	83.6	90.7	69.9
Planned queries, concatenated	83.6	92.1	70.1
Reciprocal rank fusion	92.9	95.0	72.2
Add dense retrieval	94.3	98.6	74.9
Add cross-encoder reranking	97.3	99.2	85.4
Order by answer likelihood	98.5	98.9	90.2

Table 2: Cook retrieval ablation on 476 questions. MRR is shown multiplied by 100.

4 Retrieval Architecture and Ablation

4.1 Retrieval-stage ablation

Several searches capture different ways the earlier conversation may have expressed a fact. Reciprocal rank fusion combines positions across result lists without assuming their raw scores are comparable. Dense retrieval adds matches by meaning. A cross-encoder reads the question and excerpt together; the last stage orders candidates by whether they directly answer the question. (Cook, 2026)

The ablation makes the tradeoff visible: final ordering improved R@5 by 1.2 points and MRR by 4.8 points, while R@10 fell from 99.2 to 98.9. The final pipeline optimizes earlier access to useful evidence, rather than improving every metric monotonically.

4.2 Why the measured improvements are plausible

The ablation separates two bottlenecks: finding evidence at all, and moving the strongest evidence toward the front. Combining discovery signals raised Recall@5 from 83.6% in the single-query baseline to 94.3% after dense retrieval was added. Later selection stages raised it to 98.5%. These are observations from the recorded sequence of configurations, not universal gains that every application should expect. (Cook, 2026)

Rank quality tells a related story. Before the cross-encoder stage, MRR was 74.9; after it, MRR was 85.4. Answer-oriented ordering brought it to 90.2. In everyday terms, the improvement was not only that a useful exchange appeared somewhere in the shortlist, but that the agent was more likely to encounter it earlier. Figure 2 illustrates why ranking for an answer can differ from ranking for topical similarity.

The final stage also shows a tradeoff: top-five recall improved while top-ten recall declined slightly. That matters because "better memory" is not one undifferentiated score. The design should be assessed against the actual job: finding useful evidence near the front, preserving enough context to interpret it, and then producing a correct answer.

4.3 Why adding more reasoning was not automatically better

The broader research tested stronger models, additional answer-time checking, and more elaborate representations. Several approaches were neutral or worse on the evaluated workload. The recurring lesson was that extra processing does not compensate reliably for missing or poorly selected evidence, and can introduce another opportunity to reinterpret a previously correct fact. (Cook, 2026)

This finding does not imply that reasoning is unnecessary. It identifies where the measured bottleneck lay in these experiments. Improving the evidence supplied to the answerer produced a clearer retrieval signal than simply adding more machinery around the final answer. Counting and multi-conversation reasoning remained difficult and variable even when relevant material was available.

4.4 Reproducibility boundary

Scoring is deterministic for a fixed ranked list, but constructing that list uses models and external services. The research therefore distinguishes an observed retrieval result from a guarantee that the complete pipeline will produce identical rankings on every rerun. The public report provides the dataset variant, scored cohort, unit of retrieval, metrics, and architectural rationale; it does not provide an implementation recipe.

5 Discussion and Limitations

5.1 Answering and business-context results

On a separate 150-question LongMemEval-oracle evaluation, reported configuration means were approximately 92.5-94% answer accuracy. The best single run was 96.0%, but scores at or above 95% did not reproduce. Oracle provides the relevant sessions, so it cannot establish retrieval quality under distracting history. (Cook, 2026)

Cook also recorded 91.67% (55/60) on an internal business-context evaluation with eight invented clients. No cross-client mixing was observed in that evaluation. This is a separate synthetic workload, not a guarantee about every customer conversation. Its questions exercise changing values, dates, preferences, and facts spread across exchanges.

5.2 Limitations and conclusions

The recorded retrieval pipeline improved access to answer-bearing exchanges within Cook's evaluation. It does not establish end-to-end answer accuracy of 98.5%, an identical-harness win over agentmemory, or current production reliability. Four of the recorded retrieval misses were temporal-reasoning questions. Answering experiments also found substantial run-to-run variation. (Cook, 2026)

A production recall evaluation exists separately because offline scores cannot detect every integration failure. Future comparisons should align retrieval units and scored question IDs, freeze the candidate, repeat the full evaluation, and record both recall and answer quality together.

References

Cook. Memory results, architecture, and retrieval runner, August 2026. Cook Memory evaluation records and architectural research notes. Public chart summary: https://trycook.ai/research/results.json.

agentmemory. LongMemEval-S Benchmark Results. Accessed September 5, 2026. https://github.com/rohitg00/agentmemory/blob/main/benchmark/LONGMEMEVAL.md.

LongMemEval. Benchmark source and evaluation tools. https://github.com/xiaowu0162/LongMemEval.

LongMemEval cleaned dataset. https://huggingface.co/datasets/xiaowu0162/longmemeval-cleaned.

Data availability. The ablation and aggregates are reproduced from Cook's research records. Original per-question reports are retained outside the repository; this paper is not a new independently reproduced experiment.