HeurChain and LoCoMo with Recursive Improvement: The memory retrieval advantage

11 November 2026. LoCoMo DEV split plus two sealed-holdout confirmations.

Abstract

How much of a conversational assistant's accuracy comes from its memory? We answer this on LoCoMo with a memory grid. The grid holds the question set, answer prompt, evidence budget and judge fixed, and varies only what the reader is given: nothing, the Golden Prompt (only the turns LoCoMo labels as evidence), a memory system's retrieval, or the whole conversation. With no memory, gpt-6-luna, Muse Spark 1.3 and DeepSeek V4.1 Flash each answer 0 of 543 scored questions, so every correct answer depends on what memory delivers.

Reading the whole conversation is the ceiling (426 for gpt-6-luna, 495 for Muse, 492 for DeepSeek), but it costs about 117K input tokens per question.

A retrieval configuration we call S2 adds 33 to 38 correct answers for every reader against the previous HeurChain, under same-window controls. S2 uses a deeper ranked list, neighbouring dialogue turns, session openers for time questions, and compact metadata. With S2, HeurChain scores 427 with gpt-6-luna, 446 with Muse and 443 with DeepSeek, on about 5,273 input tokens per question. For gpt-6-luna it is level with reading the whole conversation, on 4.5% of the tokens. A single pre-registered run on six sealed conversations (997 scored questions) confirmed the gain: S2 beat the previous HeurChain by 68 to 18 questions where both runs of each arm agree (net +50, $p < 0.0001$). Most of S2's gain is on single-hop questions.

A pre-registered recursive improvement campaign then targeted multi-hop and temporal questions. It found 'dialogue windows' (configuration p04a), which present each top result together with its neighbouring turns as one gap-free stretch of dialogue. On DEV this raises gpt-6-luna to 436.3 correct against 418.7 for S2 in the same window (three runs each; multi-hop +8.3, temporal +4.3), above the full-conversation score of 426. A second pre-registered holdout run, with a stricter bar ($p < 0.01$), confirmed it against S2:

757.5 against 725.5 correct of 997, net +22 where both runs agree (p = 0.004), at 8% fewer input tokens.

Run in the same window as Mem0, HeurChain p04a scores 436.3 against Mem0's 384.3 with gpt-6-luna, 469 against Mem0's 404 with Muse Spark 1.3, 467.5 against Mem0's 410 with DeepSeek V4.1 Flash, using about 4,932 input tokens per question against Mem0's 8,106. The lead is mostly single-hop; on temporal questions the two are close, with Mem0 slightly ahead for gpt-6-luna.

1. Experimental conditions

Every number in this paper comes from the same harness. The table of conditions below says what was held fixed, what was varied and what was measured, so that each comparison changes one thing at a time.

Fixed (held constant across every comparison)

- Dataset: LoCoMo, 10 long multi-session conversations. A seeded hash splits them into 4 development (DEV) conversations (701 questions, 543 scored) and 6 sealed holdout conversations (1,285 questions, 997 scored). Categories 1-4 are scored (multi-hop 106, temporal 122, open-domain 28, single-hop 287 on DEV); adversarial category 5 is answered but not scored.
- Stored text: each dialogue turn is stored once, as one unit: session label, session timestamp, dialogue id, speaker and text, plus the dataset's own image caption where present. Every memory system ingests the same turns.
- Answer prompt: one fixed instruction to answer only from the evidence, at most 512 characters, UNKNOWN if the evidence is insufficient.
- Evidence budget: retrieved units are packed in rank order into a 16 KB budget (whole units only) by one packer shared by every arm. The full-conversation baseline is the only exception.
- Judge: gpt-6-luna with one pinned rubric. It labels each answer CORRECT or WRONG against the gold answer and is identical for every arm and reader.
- Generation settings per reader (reasoning effort, verbosity, output cap) are pinned in configuration.

- HeurChain embeddings: one embedding model, with its responses recorded once and replayed exactly, so HeurChain variants differ only in retrieval logic.

Independent (what we varied)

- Evidence arm: none; Golden Prompt (labelled evidence turns only); a memory system's retrieval (HeurChain v2, HeurChain S2, HeurChain p04a, Mem0, Zep/Graphiti, Hindsight, OpenViking); the whole conversation.
- Retrieval configuration within HeurChain: ranked-list depth, neighbouring-turn expansion, session openers for time questions, metadata projection, fusion method and weights, candidate pool, date matching, dialogue windows, result header format, date rendering.
- Answer model (reader): gpt-6-luna, Muse Spark 1.3, DeepSeek V4.1 Flash. More readers appear in the full working report.

Dependent (what we measured)

- Correct answers per run, overall and by LoCoMo category.
- Question-level paired outcomes against a same-window control (which arm is right where the other is wrong).
- Input tokens per question as reported by the provider, and measured cost.

Controls and statistics

- Every improvement is read against a control run in the same time window, because the same reader drifts by several questions between windows.
- Tests are one-sided sign tests on per-question outcomes. With two runs per arm, a question counts only when both runs of one arm are right and both of the other are wrong.
- Decision rules were written down before each stage ran.
- The holdout was used twice, each time under its own pre-registration and one-shot approval, with aggregate-only output: once for S2, then once for p04a with a stricter bar, reported as a second use.

2. The role of memory: the memory grid

The grid gives each of three readers five kinds of evidence, with everything else fixed. No memory scores 0 for all three readers: LoCoMo's answers are facts about these particular conversations, so a reader without memory cannot know them. The Golden Prompt, a few hundred tokens of exactly the right dialogue (the turns LoCoMo labels as evidence), already reach 382 to 438. They fall short of the full conversation mainly on temporal questions, where the turns carry relative dates ("last week") that only the session context resolves. The full conversation is the ceiling, at about 117K tokens per question. A memory system's job is to land as close to that ceiling as possible at the cost of the Golden Prompt.

No memory scores 0; the full conversation scores highest, at 25 times HeurChain's prompt size

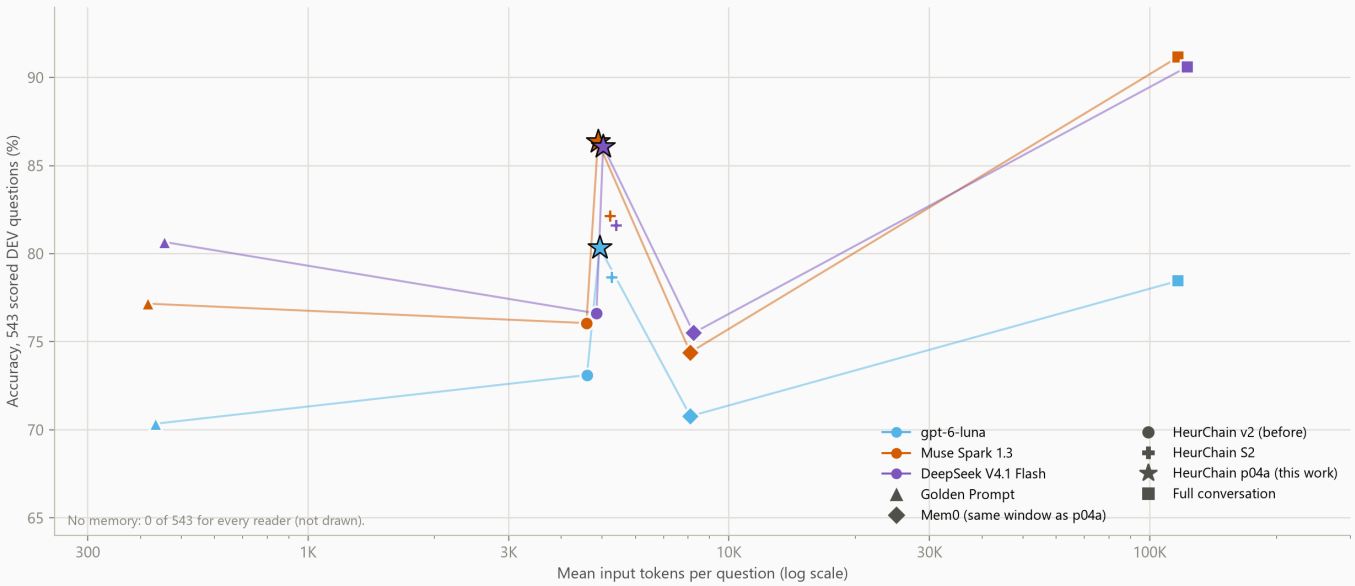
Correct of 543 scored questions, gpt-6-luna judge with the pinned rubric. Columns run from the smallest prompt to the largest. HeurChain and Mem0 pack to the same 16 KB evidence budget.

	No memory ≈ 107 input tokens per question	Golden Prompt ≈ 436 input tokens per question	HeurChain ≈ 4.7K input tokens per question	Mem0 ≈ 8.2K input tokens per question	Full conversation ≈ 119K input tokens per question
gpt-6-luna	**0** 0.0% 112 tokens	**382** 70.3% 435 tokens	**397** 73.1% 4.6K tokens · mean of 2 runs	**387** 71.3% 8.1K tokens · mean of 2 runs	**426** 78.5% 117K tokens
Muse Spark 1.3 Contributor	**0** 0.0% 93 tokens	**419** 77.2% 416 tokens	**413** 76.1% 4.6K tokens	**408** 75.1% 8.1K tokens	**495** 91.2% 117K tokens
DeepSeek V4.1 Flash	**0** 0.0% 116 tokens	**438** 80.7% 456 tokens	**416** 76.6% 4.8K tokens · mean of 2 runs	**407** 75.0% 8.2K tokens	**492** 90.6% 123K tokens

Cell shade follows accuracy from 65% (light) to 93% (dark). Tokens are the provider-reported mean input tokens per scored question.

Correct of 543 scored DEV questions for each evidence arm and reader (the HeurChain column is v2, before S2).

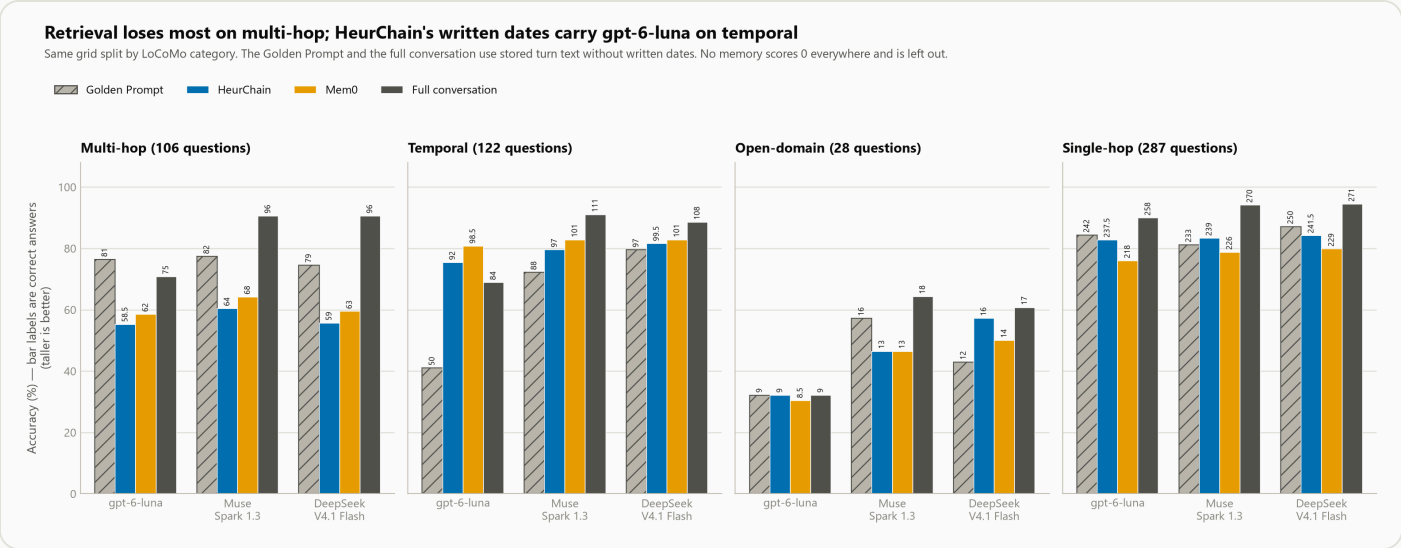
What memory buys: HeurChain p04a beats Mem0 for every reader in the same window, on about 60% of Mem0's tokens



S2 and p04a are HeurChain broker test versions (retrieval configurations). S2: deeper ranked list, neighbouring turns, session openers for time questions, compact metadata. p04a: S2 plus dialogue windows.

Accuracy against input tokens per question. The large stars are HeurChain p04a and the crosses S2; Mem0 points are from the same window as p04a; the other markers are the grid arms. Each colour is one reader.

Reader	No memory	Golden Prompt	Mem0	HeurChain v2	HeurChain S2	HeurChain p04a
gpt-6-luna	0	382	384.3	393	427	436.3
Muse Spark 1.3	0	419	404	413	446	469
DeepSeek V4.1 Flash	0	438	410	405	443	467.5



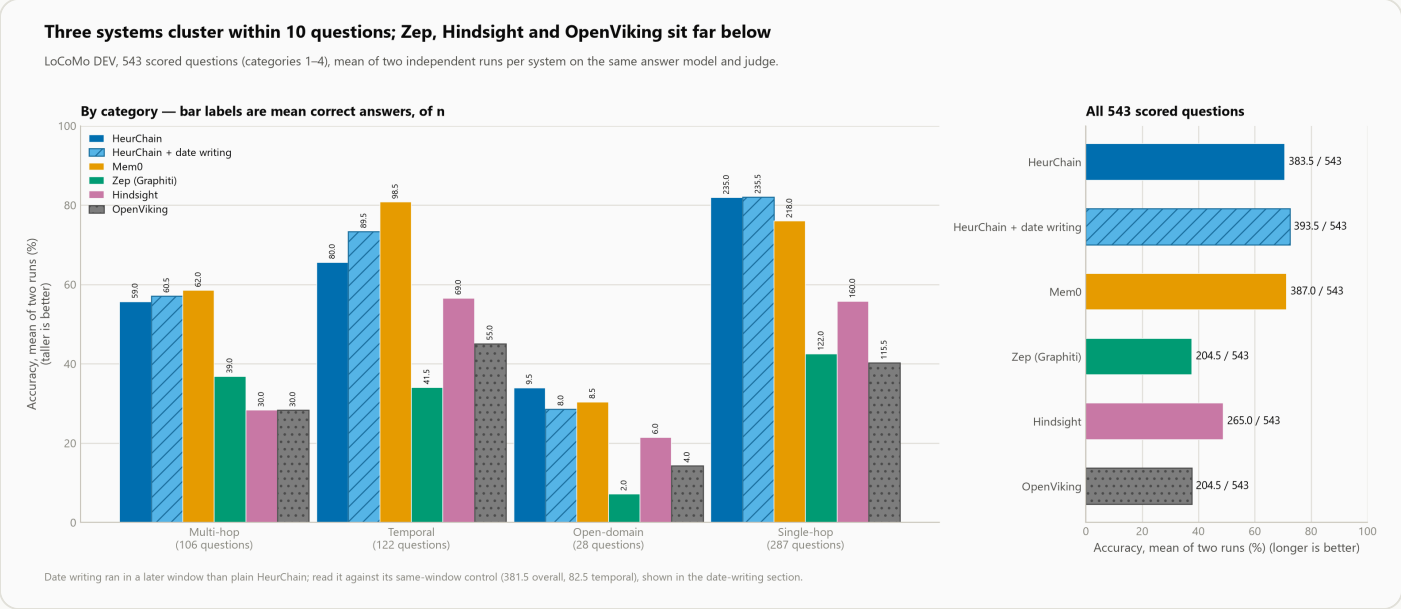
The grid by LoCoMo category (HeurChain column is v2).

3. Memory systems compared

Five memory systems were run with gpt-6-luna as reader, two runs each, under the conditions above. The two leaders were close; the others were hampered by how much of each conversation they retained or returned.

System (gpt-6-luna, two-run mean)	Correct of 543	Accuracy
HeurChain (v1 dates)	393.5	72.5%
Mem0	387	71.3%
Hindsight	265	48.8%
Zep (Graphiti)	204.5	37.7%

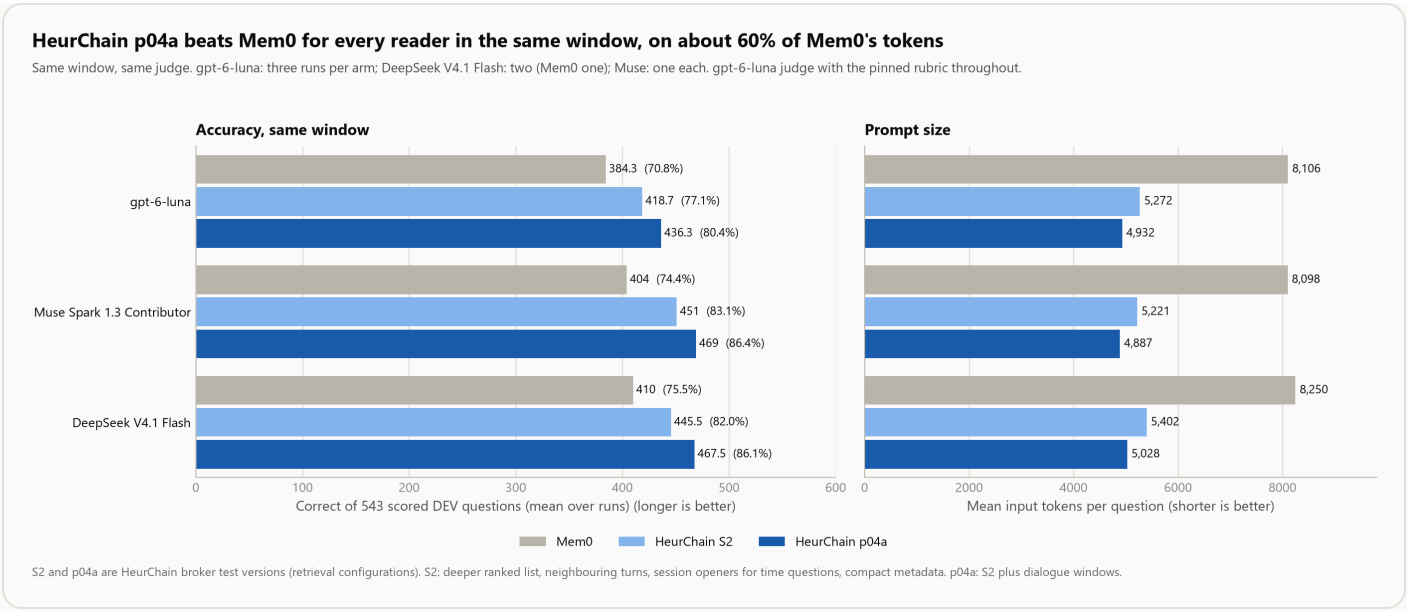
System (gpt-6-luna, two-run mean)	Correct of 543	Accuracy
OpenViking	204.5	37.7%



Mean correct answers by category for the memory systems (earlier window; HeurChain before S2).

3.1 Same-window head-to-head with Mem0

Mem0 was re-run in the same window as S2 and p04a, so the head-to-head below is not affected by drift between windows: three runs per arm for gpt-6-luna, two for DeepSeek (one for Mem0), and one each for Muse. Each count is a mean over runs. p04a leads Mem0 for every reader, on about 61% of Mem0's input tokens. Most of the lead is on single-hop questions. On temporal questions the two are close: p04a is ahead with DeepSeek and Muse, and with gpt-6-luna Mem0 scores 99.7 against p04a's 97 and S2's 92.7 (S2 against Mem0: 10 to 22). Temporal questions are where HeurChain has the most room to improve. The paired tests compare each question's mean correctness over runs; they are one-sided in the direction shown.



Mem0, S2 and p04a in the same window: correct answers (left) and input tokens per question (right) for each reader.

Reader	Memory	Runs	Multi-hop	Temporal	Open-domain	Single-hop	All	I
gpt-6-luna	Mem0	3	59.7	99.7	8	217	384.3	8
gpt-6-luna	HeurChain S2	3	62	92.7	9	255	418.7	5
gpt-6-luna	HeurChain p04a	3	70.3	97	10.3	258.7	436.3	4
Muse Spark 1.3	Mem0	1	67	99	11	227	404	8
Muse Spark 1.3	HeurChain S2	1	74	102	14	261	451	5
Muse Spark 1.3	HeurChain p04a	1	78	105	18	268	469	4
DeepSeek V4.1 Flash	Mem0	1	67	104	13	226	410	8
DeepSeek V4.1 Flash	HeurChain S2	2	69.5	103.5	15.5	257	445.5	5
DeepSeek V4.1 Flash	HeurChain p04a	2	77.5	107.5	16	266.5	467.5	5

Questions where p04a is right and Mem0 wrong, against the reverse (mean over runs):

Reader	Multi-hop	Temporal	Open-domain	Single-hop
gpt-6-luna	32 / 19, p = 0.05	10 / 17, Mem0 ahead, p = 0.12	4 / 2, p = 0.34	58 / 10, p < 0.0001
Muse Spark 1.3	21 / 10, p = 0.04	13 / 7, p = 0.13	7 / 0, p = 0.008	46 / 5, p < 0.0001

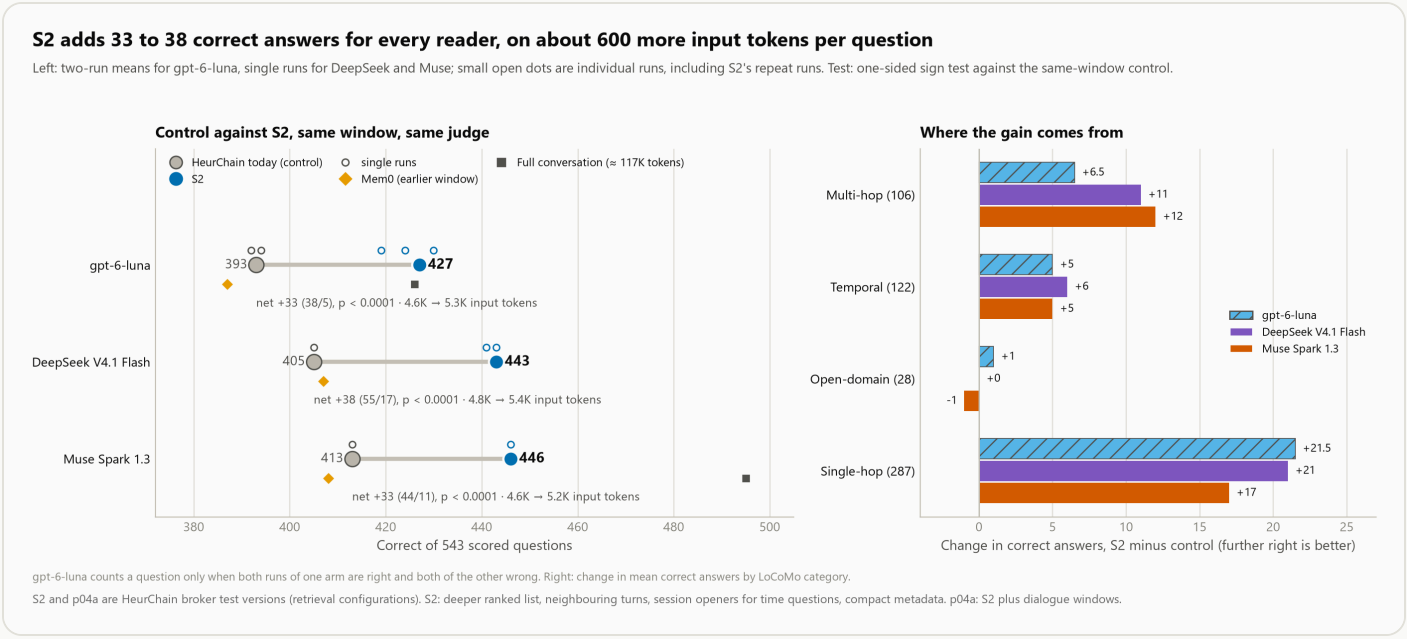
Reader	Multi-hop	Temporal	Open-domain	Single-hop
DeepSeek V4.1 Flash	24 / 14, p = 0.07	11 / 8, p = 0.32	6 / 2, p = 0.14	46 / 6, p < 0.0001

4. Closing the gap: retrieval configuration S2

S2 changes only how HeurChain retrieves and returns evidence:

- a ranked list of 30 instead of 20;
- the turns either side of the top 5 results, from the same session;
- the opening turns of the first three sessions, for questions that ask for a time;
- each result's metadata trimmed to its conversation id, so more evidence fits the same 16 KB budget.

Against the previous HeurChain in the same window it adds 33 (gpt-6-luna, both runs agree 38/5), 38 (DeepSeek) and 33 (Muse) correct answers, all at p < 0.0001. It sends about 600 more input tokens per question.

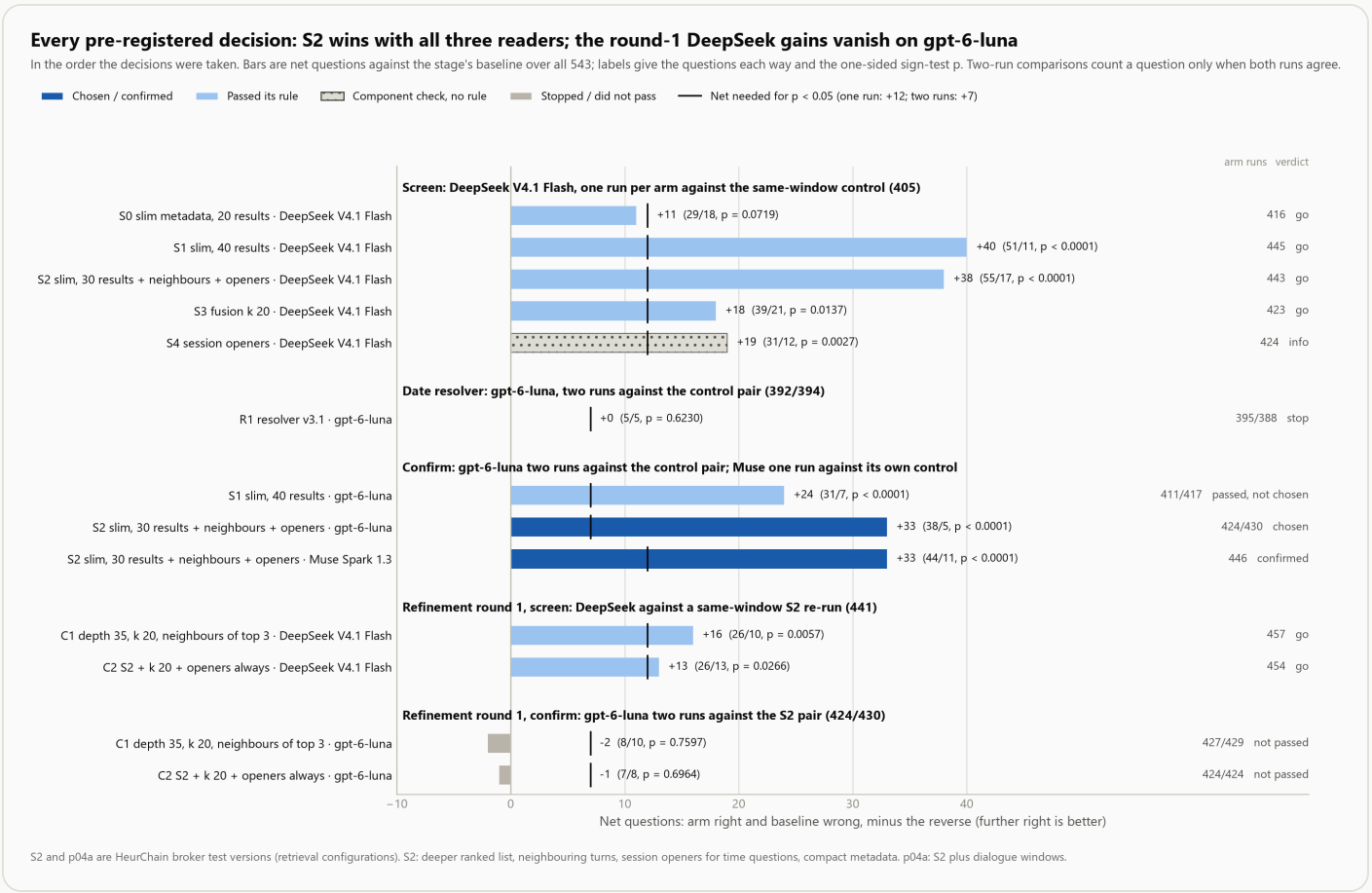


Previous HeurChain against S2 for each reader, with Mem0 and the full conversation for reference, and the change by category.

Reader	Multi-hop	Temporal	Open-domain	Single-hop	All
gpt-6-luna	+6.5	+5	+1	+21.5	+34

Reader	Multi-hop	Temporal	Open-domain	Single-hop	All
Muse Spark 1.3	+12	+5	-1	+17	+33
DeepSeek V4.1 Flash	+11	+6	+0	+21	+38

Every decision was pre-registered and taken in order. The chart below also shows the variants that did not pass: a date-resolver revision, and refinements that helped DeepSeek but not gpt-6-luna.

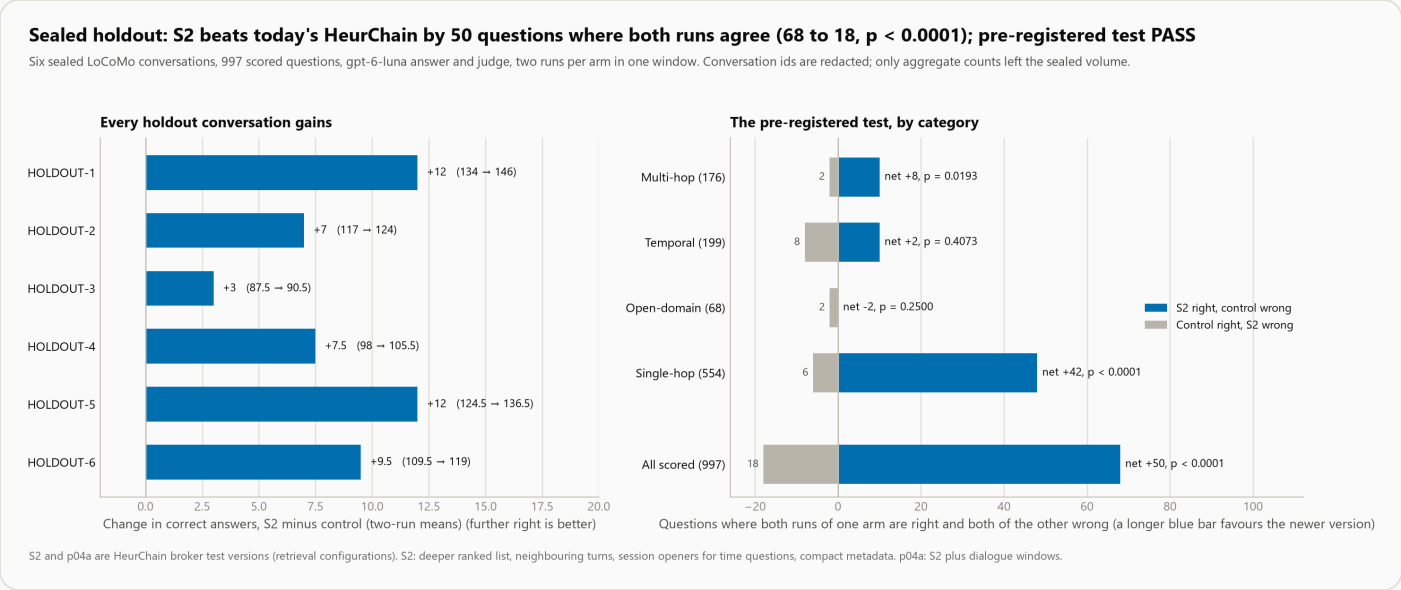


Every pre-registered decision on DEV, in order.

5. Holdout confirmation

Six conversations were sealed from the start. After the DEV work, S2 was built into the broker as options that are off by default. It was then run once on the holdout against the same broker with those options off: gpt-6-luna, two runs per arm, 997 scored questions. The pass rule was written down beforehand: a net of at least +10 at $p < 0.05$, with no category significantly worse.

S2 averaged 721.5 correct against 670.5 and won 68 questions to 18, net +50, p < 0.0001. All six conversations improved.



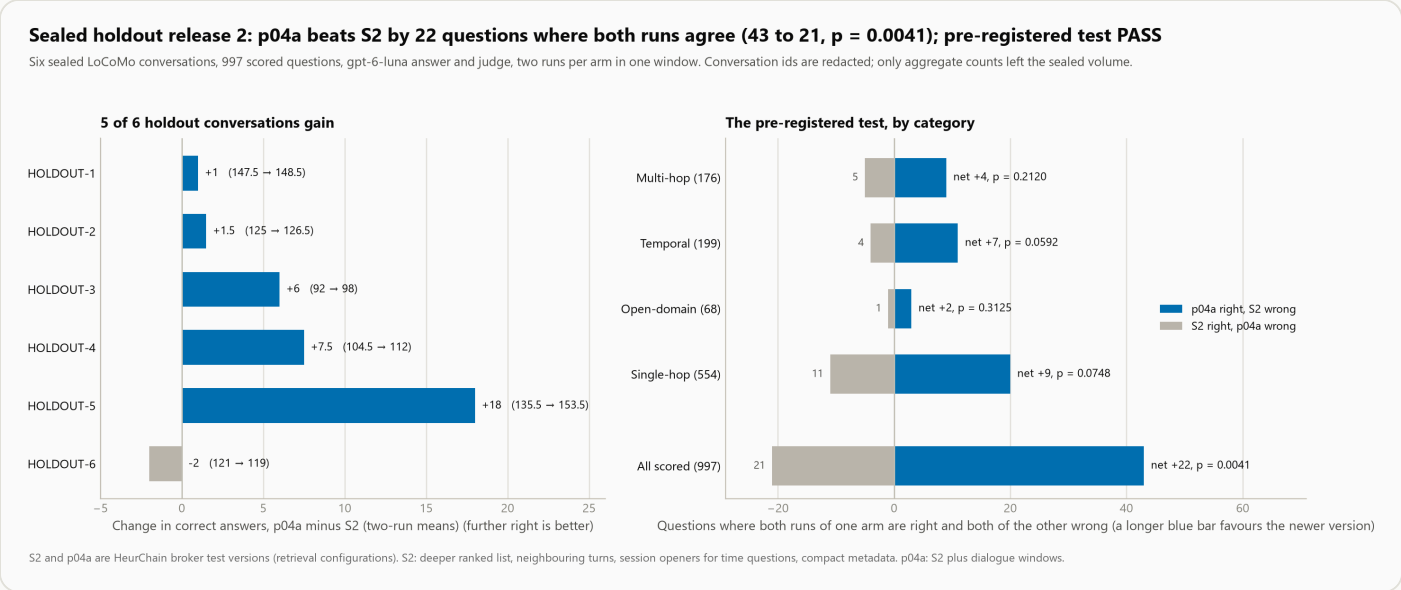
Sealed holdout: change per conversation (ids redacted) and the pre-registered test by category.

Category	Questions	S2 better / previous better	Net	p
multi-hop	176	10 / 2	+8	0.0193
temporal	199	10 / 8	+2	0.4073
open-domain	68	0 / 2	-2	1.0000
single-hop	554	48 / 6	+42	< 0.0001
all scored	997	68 / 18	+50	< 0.0001

5.2 Second holdout run: p04a against S2

After the recursive improvement campaign (section 6) promoted p04a on DEV, a second holdout run compared it with S2 on the same six conversations. It was pre-registered with a stricter bar than the first: a net of at least +10 at p < 0.01, and no category net -6 or worse at p < 0.05. S2's cells from the first run were reused unchanged, so S2's scores here (728, 723) can be compared with its first-run scores (717, 726). p04a averaged 757.5 correct against 725.5 and won 43 questions to 21, net +22, p = 0.004: PASS. Every category moved in p04a's favour, though none is significant on its own; on multi-hop plus temporal it won 20 to 9 (p = 0.03). Five of six conversations gained, one by 18 questions, and one lost 2. p04a sent 4,803 input tokens per

question against S2's 5,219. Because the decision to run a second check was made after the first result was known, this is reported as a second use of the holdout, not an untouched test.



Second sealed holdout run, p04a against S2: change per conversation (ids redacted) and the pre-registered test by category.

Category	Questions	p04a better / S2 better	Net	p
multi-hop	176	9 / 5	+4	p = 0.21
temporal	199	11 / 4	+7	p = 0.06
open-domain	68	3 / 1	+2	p = 0.31
single-hop	554	20 / 11	+9	p = 0.07
multi-hop + temporal	375	20 / 9	+11	p = 0.03
all scored	997	43 / 21	+22	p = 0.004

6. Recursive improvement on top of S2

A recursive improvement campaign then let a coding agent propose broker changes under fixed, pre-registered rules: a free evidence check, a DeepSeek screen, and promotion on three gpt-6-luna runs against the champion's three. The fitness was the gain on multi-hop plus temporal questions, with single-hop, open-domain and the total required not to fall. The first passing proposal, 'dialogue windows', folds each top result's neighbouring turns, including lower-ranked results that sit next to it, into one gap-free stretch of dialogue. On DeepSeek it

scored 473 against the champion's 445 and 446. On gpt-6-luna it scored 437, 437, 435 against 425, 412, 419 (multi-hop +8.3, temporal +4.3, single-hop +3.7, total +17.7, mean of three runs each). It met every promotion rule and became the new champion. On multi-hop and temporal questions it won 39 to 22 ($p = 0.02$), and it gained in all four DEV conversations. The gain was larger on the half used for tuning (+11.0) than on the half held back for validation (+1.7), so some of it may be fitted to the tuning conversations. It was therefore checked on the sealed holdout (section 5.2), where it passed. Note also that S2 itself scored lower in this window (425, 412, 419) than in its first (424, 430); same-window comparison is what makes the gain readable.

7. Limits

- One benchmark: LoCoMo's 10 conversations are synthetic two-person chats; results may not transfer to other memory workloads.
- The five-system comparison (section 3) and the memory grid's Mem0 column come from earlier windows. The head-to-head with Mem0 (section 3.1) was re-run in the same window as S2 and p04a; prefer those numbers. The same reader drifts by several questions between windows.
- p04a was found by a search that used the four DEV conversations, and its gain was larger on the tuning half than on the validation half. The holdout check guards against that, but it was the holdout's second use, chosen after release 1's aggregate result was known.
- Zep was read through Graphiti edge search only, and Hindsight retained about 0.3 facts per turn. These are not faithful readings of those systems as published.
- gpt-6-luna both answers and judges in its own rows. The judge is the same for every arm, so this does not favour any memory system, but absolute scores may differ under another judge.
- Muse numbers are single runs on DEV, as are DeepSeek's Mem0 run and its S2 numbers in section 4; both holdout checks used gpt-6-luna only.
- Small local models (below about 30B parameters) were not tested.