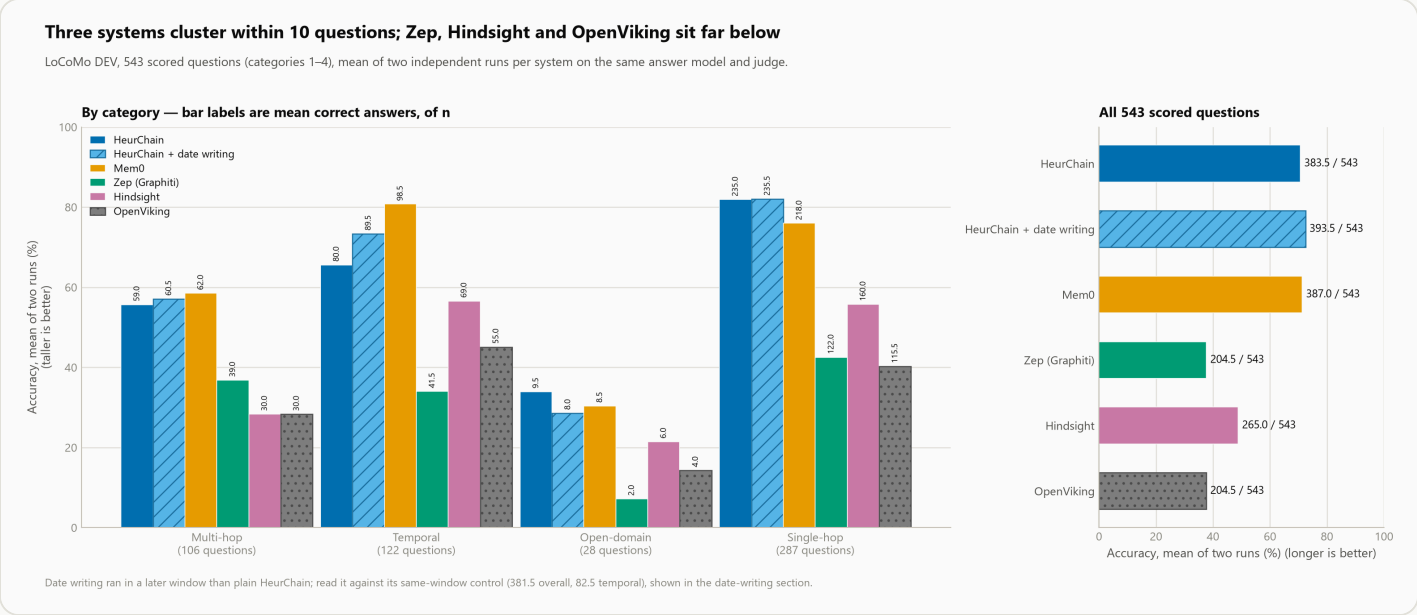
Six systems on LoCoMo DEV

543 scored questions across four conversations, two independent runs per system in the six-system comparison, on the same answer model and judge. HeurChain + date writing posts the highest mean, but the lead over Mem0 is 6.5 questions; the clearer story is the category shape, the abstention rates, and how much run-to-run noise there is. A matched same-window comparison with Mem0 follows the six-system sections.

- HeurChain
  383.5 /543
  70.6% · two-run mean
- HeurChain + date writing
  393.5 /543
  72.5% · two-run mean
- Mem0
  387.0 /543
  71.3% · two-run mean
- Zep (Graphiti)
  204.5 /543
  37.7% · two-run mean
- Hindsight
  265.0 /543
  48.8% · two-run mean
- OpenViking
  204.5 /543
  37.7% · two-run mean

Headline



Mean correct answers by category and overall, two runs per system.

▶ Table view — the same numbers as the image

Summary table

Date writing leads by 6.5 questions; the category profile differs sharply by system

Correct answers out of 543 scored questions (LoCoMo DEV, categories 1–4). Per-run numbers come from each run; every other cell is the mean of the two runs. Cost is each run's recorded spend.

All scored questions — per run, mean, and cost

Run	Correct / 543	Accuracy	Cost per run (USD)
HeurChain — run a	386 / 543	71.1%	$0.4396
HeurChain — run b	381 / 543	70.2%	$0.0717
HeurChain — mean of runs	**383.5 / 543**	**70.6%**	**$0.2556**
HeurChain + date writing — run a	391 / 543	72.0%	$0.2304
HeurChain + date writing — run b	396 / 543	72.9%	$0.2807
HeurChain + date writing — mean of runs	**393.5 / 543**	**72.5%**	**$0.2556**
Mem0 — run a	386 / 543	71.1%	$0.4291
Mem0 — run b	388 / 543	71.5%	$0.4622
Mem0 — mean of runs	**387.0 / 543**	**71.3%**	**$0.4456**
Zep (Graphiti) — run a	205 / 543	37.8%	$0.4767
Zep (Graphiti) — run b	204 / 543	37.6%	$0.4770
Zep (Graphiti) — mean of runs	**204.5 / 543**	**37.7%**	**$0.4768**
Hindsight — run a	262 / 543	48.3%	$0.0759
Hindsight — run b	268 / 543	49.4%	$0.5167
Hindsight — mean of runs	**265.0 / 543**	**48.8%**	**$0.2963**
OpenViking — run a	208 / 543	38.3%	$0.3190
OpenViking — run b	201 / 543	37.0%	$0.0615
OpenViking — mean of runs	**204.5 / 543**	**37.7%**	**$0.1903**

By category — mean of two runs, with accuracy

Category	n	HeurChain	HeurChain + date writing	Mem0	Zep (Graphiti)	Hindsight
1 Multi-hop	106	59.0 (55.7%)	60.5 (57.1%)	62.0 (58.5%)	39.0 (36.8%)	30.0 (28.3%)
2 Temporal	122	80.0 (65.6%)	89.5 (73.4%)	98.5 (80.7%)	41.5 (34.0%)	69.0 (56.6%)
3 Open-domain	28	9.5 (33.9%)	8.0 (28.6%)	8.5 (30.4%)	2.0 (7.1%)	6.0 (21.4%)
4 Single-hop	287	235.0 (81.9%)	235.5 (82.1%)	218.0 (76.0%)	122.0 (42.5%)	160.0 (55.7%)
All scored	**543**	**383.5 (70.6%)**	**393.5 (72.5%)**	**387.0 (71.3%)**	**204.5 (37.7%)**	**265.0 (48.8%)**

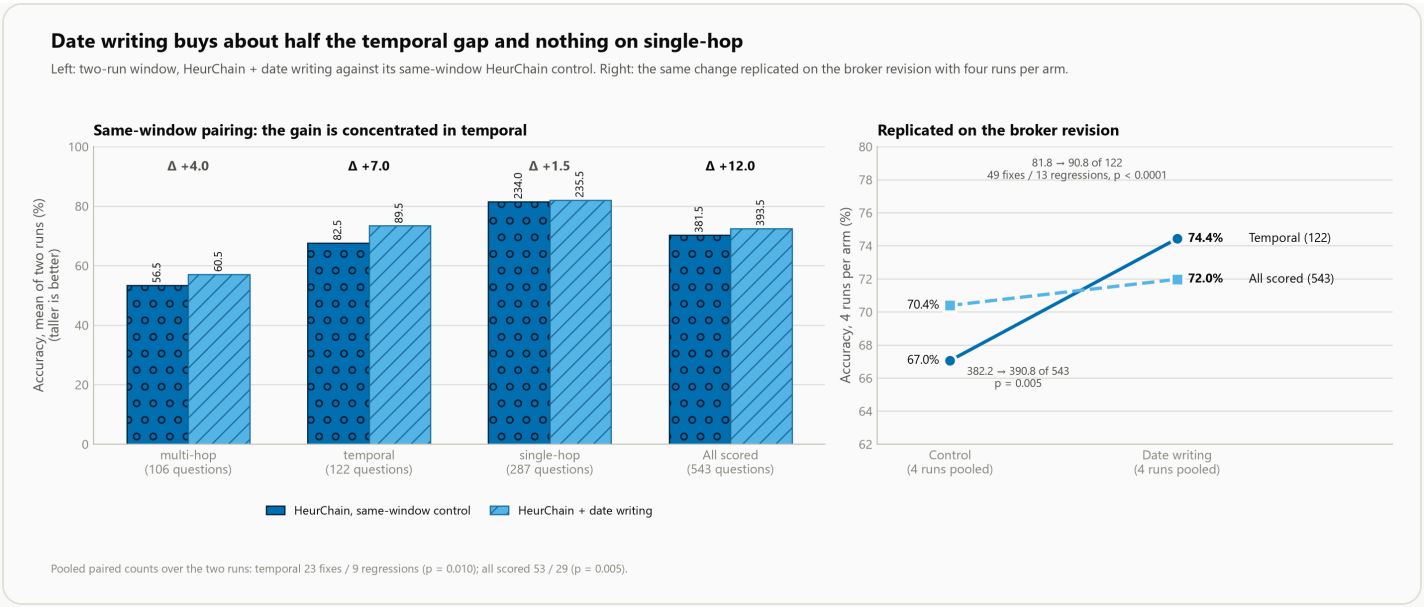
By conversation — mean of two runs, with accuracy

Conversation	n	HeurChain	HeurChain + date writing	Mem0	Zep (Graphiti)	Hindsight
conv-26	152	107.5 (70.7%)	110.5 (72.7%)	103.0 (67.8%)	63.0 (41.4%)	73.0 (48.0%)
conv-30	81	60.5 (74.7%)	64.5 (79.6%)	57.0 (70.4%)	27.5 (34.0%)	33.5 (41.4%)
conv-41	152	111.0 (73.0%)	111.0 (73.0%)	116.0 (76.3%)	47.5 (31.2%)	86.5 (56.9%)
conv-50	158	104.5 (66.1%)	107.5 (68.0%)	111.0 (70.3%)	66.5 (42.1%)	72.0 (45.6%)
All scored	**543**	**383.5 (70.6%)**	**393.5 (72.5%)**	**387.0 (71.3%)**	**204.5 (37.7%)**	**265.0 (48.8%)**

Means average the two independent runs, so mean cells can be fractional. Accuracy = correct / n.
Dollar cost per run is dominated by prompt-caching, not by retrieval: HeurChain run a paid $0.4396 and run b $0.0717 for the same prompts.
Compare systems on token volume, not on these dollar figures. Only four conversations: treat per-conversation cells as descriptive.

Per-run totals, means and cost, plus the same counts split by category and conversation.

Date writing

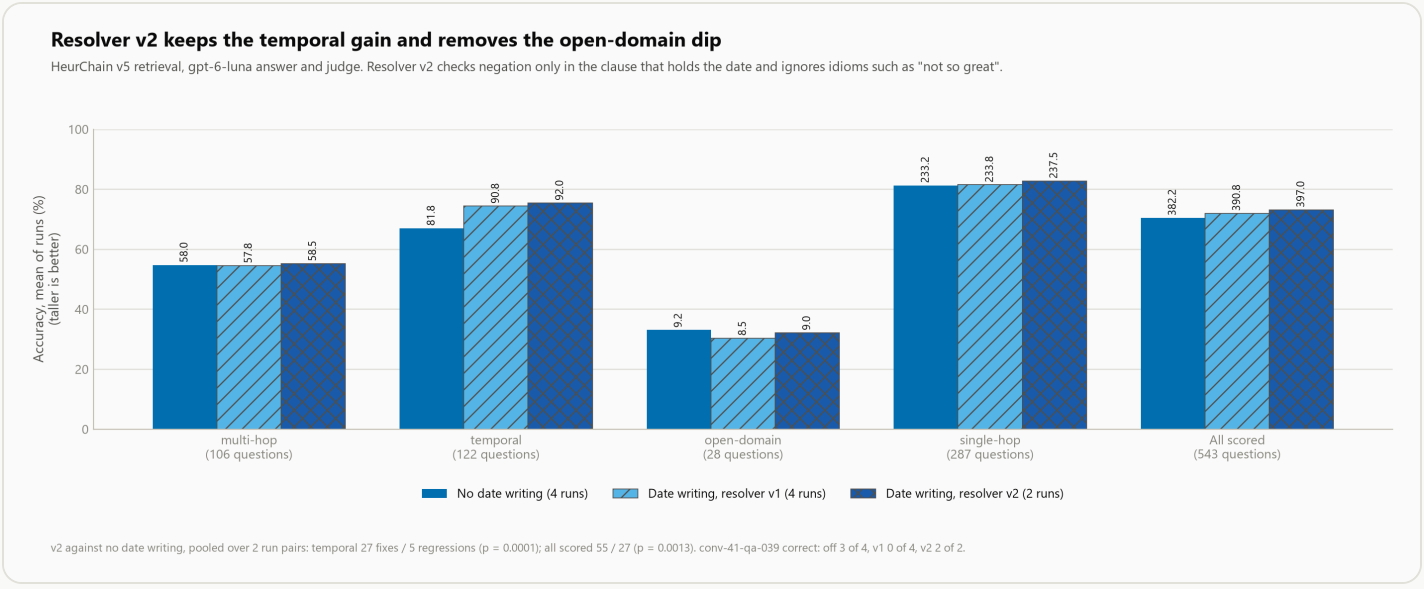


HeurChain + date writing against its same-window control (two runs), and the replication on the broker revision (four runs per arm).

HeurChain + date writing ran in a later window than plain HeurChain, so the effect is read against its same-window control, not against the plain column. Over the two runs the paired result is 23 fixes and 9 regressions on temporal questions of 122 (p = 0.010), and 53 fixes and 29 regressions overall (p = 0.005). On the broker revision with four runs per arm the gain replicated: temporal 81.8 → 90.8 of 122 (49/13, p < 0.0001), all scored 382.2 → 390.8 of 543 (p = 0.005), multi-hop 58.0 → 57.8 and single-hop 233.2 → 233.8, both unchanged.

▶ Table view — the same numbers as the image

Latest build: resolver v2

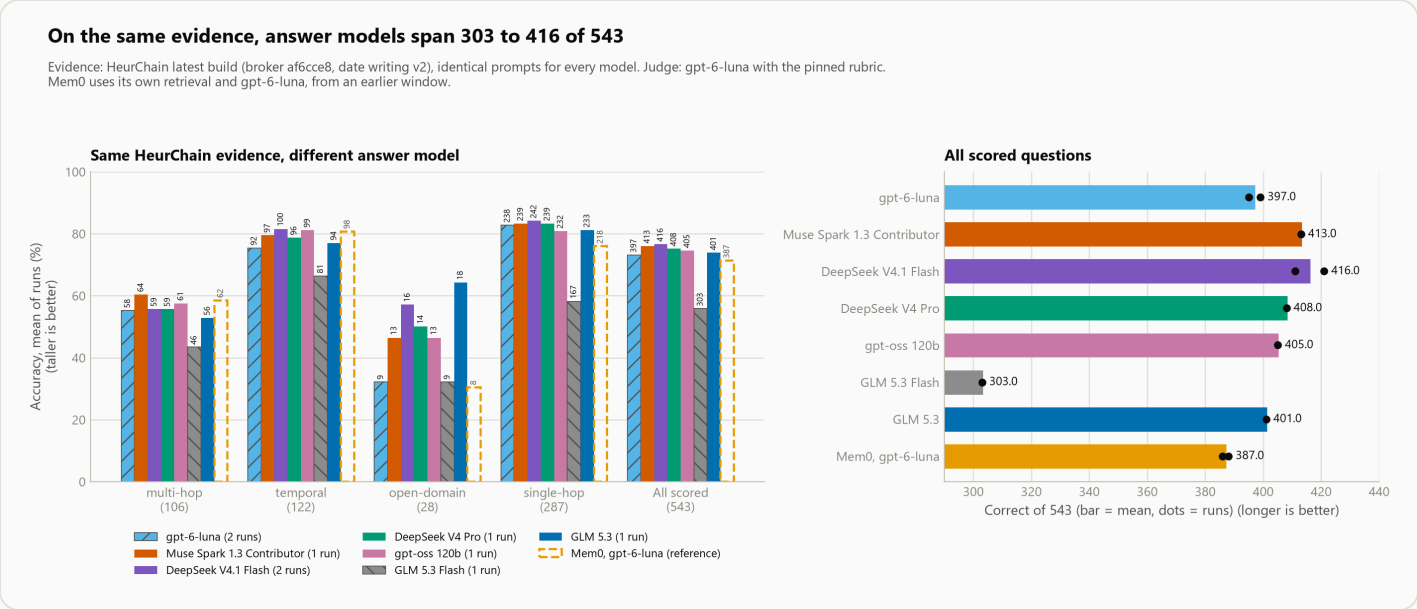


No date writing, resolver v1 and resolver v2 on the same HeurChain retrieval, gpt-6-luna answer and judge.

Resolver v2 fixes a precision bug in date writing: v1 skipped any sentence containing a negating word, so "something not so great happened yesterday" lost its date while a distractor event kept one, and conv-41-qa-039 went from right to UNKNOWN. v2 checks negation only in the date's own clause and ignores negator idioms. Only 7 of 1,851 evidence units change, so most of the v2-over-v1 overall gap is run-to-run variance; the fix's own effect is about +2 per run on the 29 scored questions whose prompts changed. Cells rebuilt from the latest broker build are identical to the v2 cells, so the v2 column is the latest build.

▶ Table view — the same numbers as the image

Answer models on the latest build



Answer models on the latest HeurChain build: by category (left) and overall with per-run dots (right); Mem0 with gpt-6-luna as a dashed reference.

Every model sees byte-identical prompts built from the latest HeurChain evidence and is judged by gpt-6-luna with the pinned rubric, so differences are the answer model's. Paired counts are question-by-question against the two gpt-6-luna runs. Muse has one run; its answer cost uses a third-party per-token price that Meta has not confirmed. Ollama Cloud is billed by subscription, so Ollama answers show $0 and the cost column is the judge only. DeepSeek V4 Pro, gpt-oss 120b, GLM 5.3 Flash and GLM 5.3 have one baseline run each; second runs were not made. GLM 5.3 Flash often ignores the requested JSON answer envelope: 176 of its answers are blocked for format and count as wrong under the protocol, so its row measures instruction-following as much as reading. GLM 5.3 has the same habit to a lesser degree (41 blocked). The Mem0 row is a reference only: Mem0 used its own retrieval and gpt-6-luna, in an earlier window.

gpt-oss 120b against Mem0's published claims: the Mem0 paper (arXiv 2504.19413) reports an LLM-judge score of 66.9% for Mem0, 68.4% for Mem0 with graph memory and 72.9% for full context, all with gpt-4o-mini answering, on all ten LoCoMo conversations. Here gpt-oss 120b, an open-weight model of similar class, answers from HeurChain evidence and scores 405 of 543 (74.6%), and Mem0 with gpt-6-luna scores 387 (71.3%). The comparison is directional only: this is the four-conversation DEV split, the judge is gpt-6-luna with a pinned rubric rather than the paper's gpt-4o-mini judge, and the answer model differs.

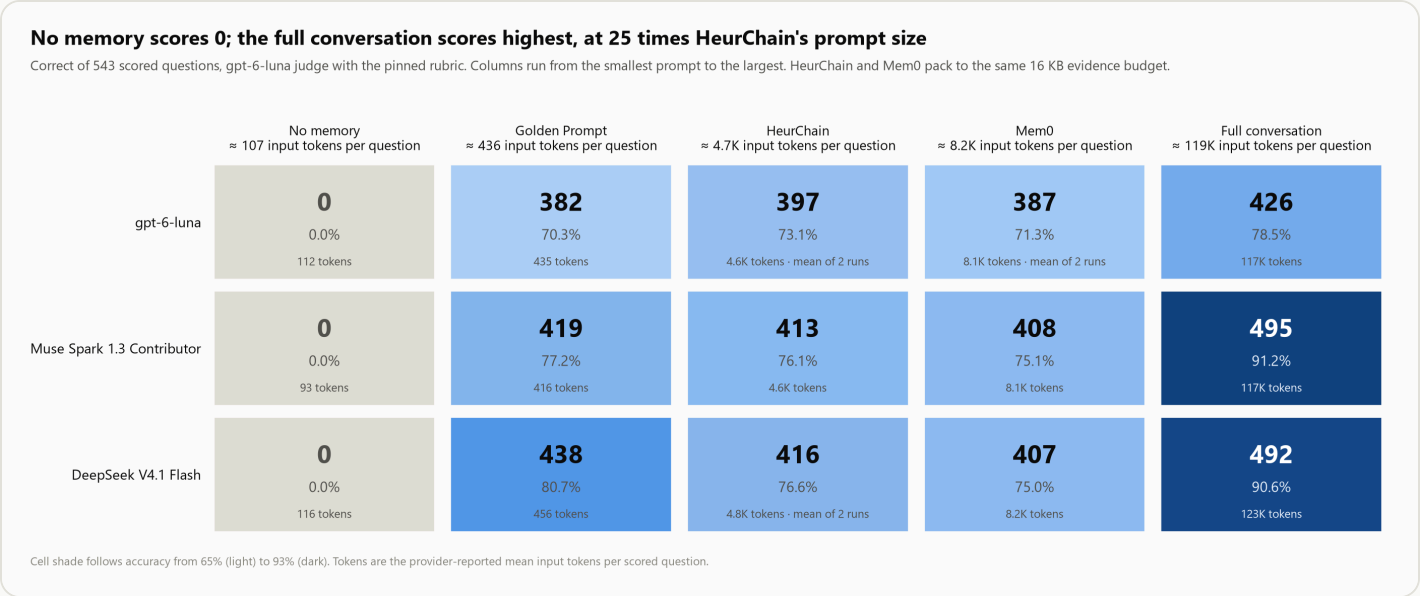
▼ Table view

Answer model	Correct per run	Multi-hop	Temporal	Open-domain	Single-hop	All	Blocked	Cost per run (USD)	Paired vs gpt-6-luna, all (fixes / regressions)
gpt-6-luna (OpenAI Responses)	395, 399	58.5	92.0	9.0	237.5	397.0	0	$0.256	—
Muse Spark 1.3 Contributor (Meta Model API)	413	64.0	97.0	13.0	239.0	413.0	0	$0.434	74 / 42 (p = 0.0019)
DeepSeek V4.1 Flash (Ollama Cloud)	411, 421	59.0	99.5	16.0	241.5	416.0	5	$0.031	77 / 39 (p = 0.0003)
DeepSeek V4 Pro (Ollama Cloud)	408	59.0	96.0	14.0	239.0	408.0	0	$0.007	62 / 40 (p = 0.0185)
gpt-oss 120b (Ollama Cloud)	405	61.0	99.0	13.0	232.0	405.0	3	$0.001	77 / 61 (p = 0.1007)
GLM 5.3 Flash (Ollama Cloud)	303	46.0	81.0	9.0	167.0	303.0	176	$0.005	75 / 263 (p = 1.0000)
GLM 5.3 (Ollama Cloud)	401	56.0	94.0	18.0	233.0	401.0	41	$0.003	74 / 66 (p = 0.2771)
Mem0 with gpt-6-luna (reference, own retrieval)	386, 388	62.0	98.5	8.5	218.0	387.0	—	—	—

Memory grid: evidence × answer model

Five kinds of evidence, from none to the whole conversation, given to three answer models: gpt-6-luna, Muse Spark 1.3 Contributor and DeepSeek V4.1 Flash. The grid separates what the memory

system delivers from what the answer model does with it.



Correct answers for every evidence arm with each answer model, columns ordered by prompt size.

What the grid shows

1. **Every point comes from the evidence.** With no memory all three models score 0 of 543, so the answers are not in the models' own knowledge and every correct answer in the grid depends on what the prompt carries.

2. **The full conversation is the accuracy ceiling, at about 25 times HeurChain's prompt.** gpt-6-luna scores 426, Muse 495 and DeepSeek V4.1 Flash 492 with the full conversation, against 397, 413 and 416 with HeurChain (per-question $p = 0.0025$, $p < 0.0001$ and $p < 0.0001$). It sends about 119,000 input tokens per question, against about 4,700 for HeurChain and 8,200 for Mem0.

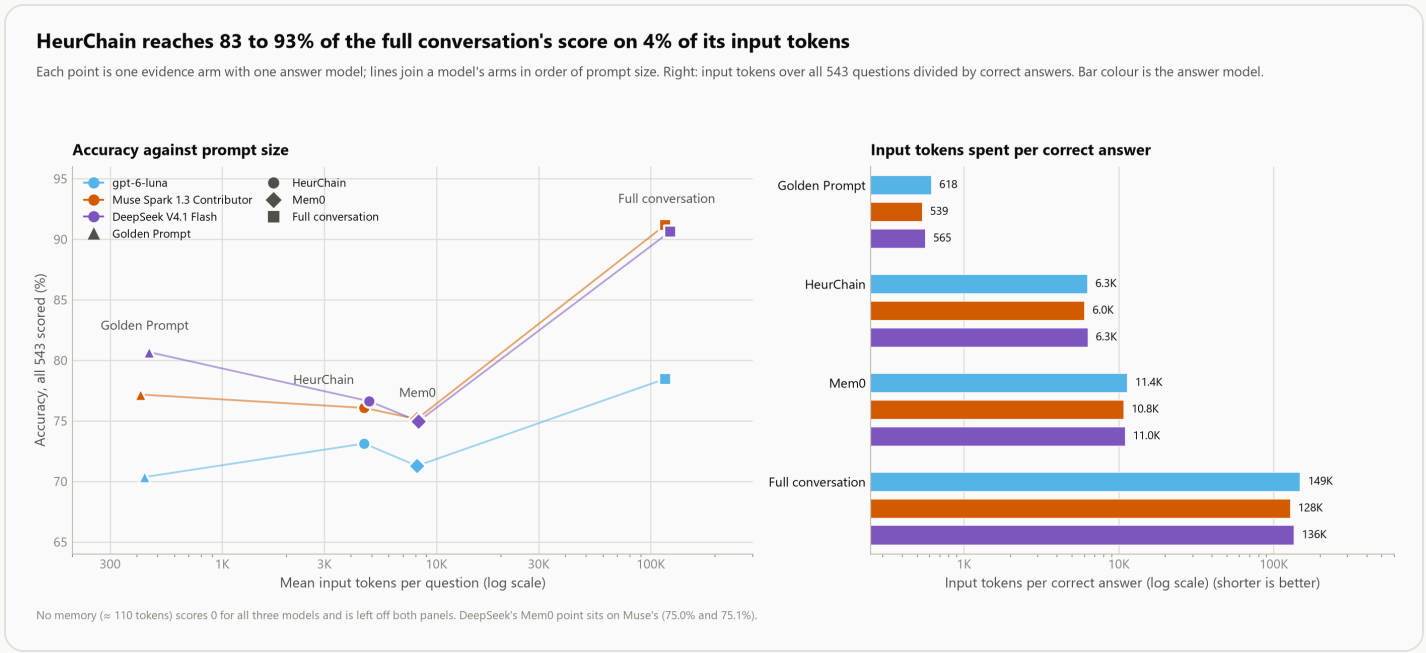
3. **HeurChain and Mem0 are level overall; they win different questions.** HeurChain leads by 10, 5 and 9 questions with gpt-6-luna, Muse and DeepSeek, but no overall difference is significant per question ($p = 0.1007$, 0.3581, 0.2813). HeurChain is ahead on single-hop by 12.5 to 19.5 questions ($p = 0.0010$ with gpt-6-luna); Mem0 is a little ahead on temporal and multi-hop. HeurChain sends 43% fewer input tokens for the same 16 KB budget.

4. **Multi-hop is where retrieval loses most.** Given only the Golden Prompt, the models answer 79 to 82 of 106 multi-hop questions; with HeurChain they answer 58.5 to 64 (per-question $p \le 0.0020$ for each model). Retrieval does not deliver every hop, and that is HeurChain's largest single gap.

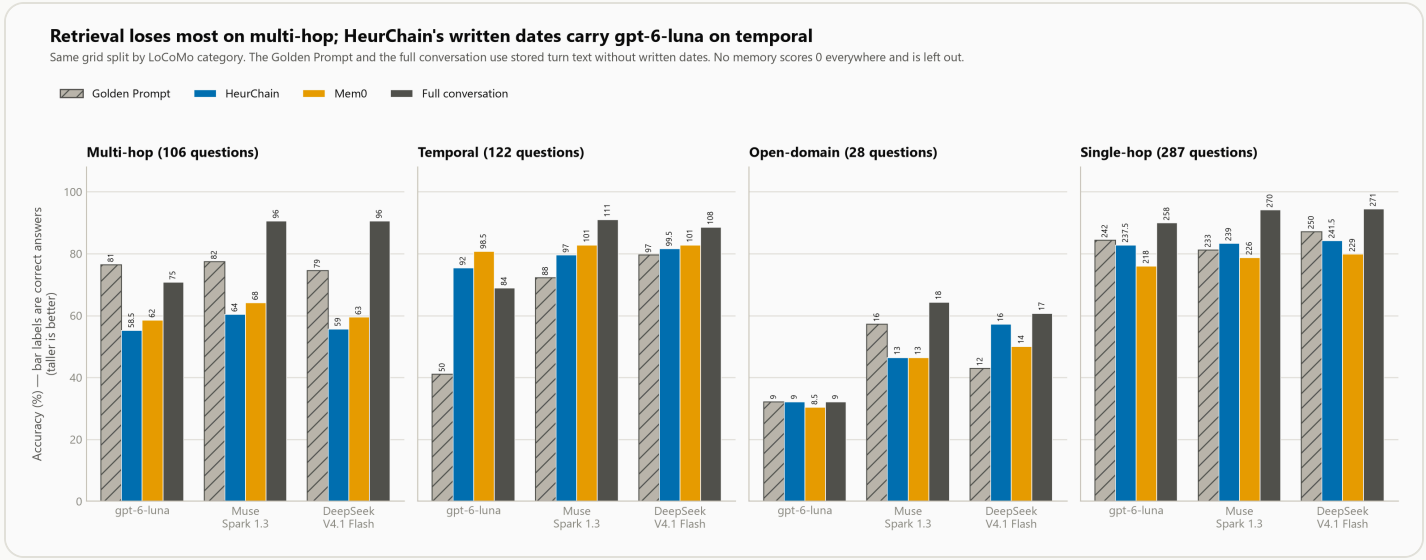
5. **Written dates matter more than the right turns for gpt-6-luna.** With the Golden Prompt gpt-6-luna answers 50 of 122 temporal questions, because it repeats relative phrases such as "Yesterday" or "Last week" instead of resolving them. With HeurChain's written dates it answers

92 ($p < 0.0001$). Muse and DeepSeek resolve the dates themselves (88 and 97 with the Golden Prompt).

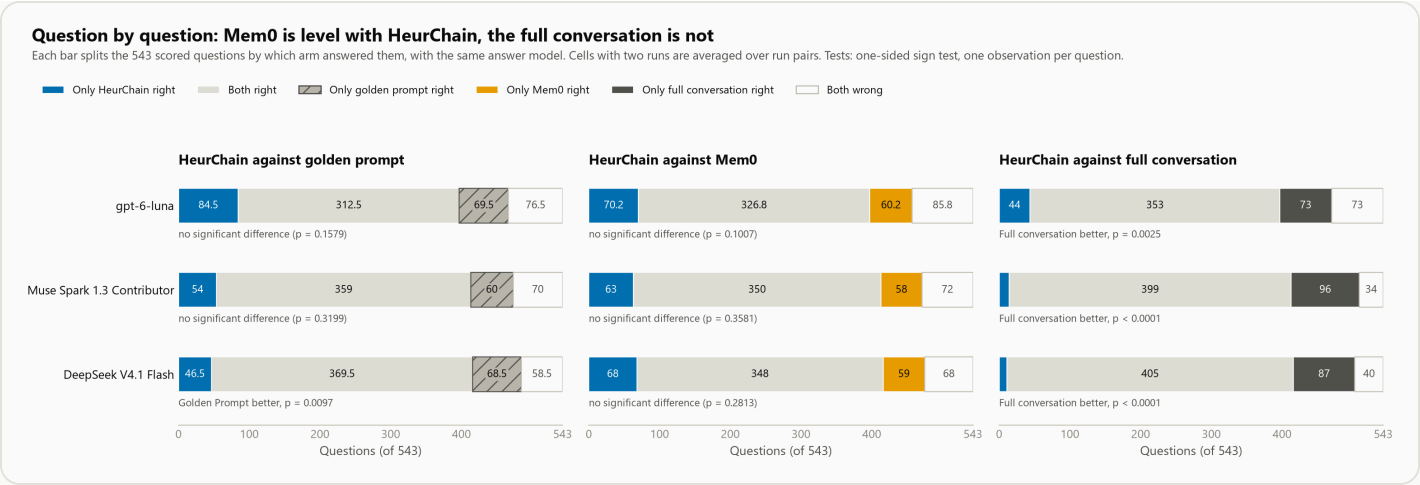
6. **DeepSeek V4.1 Flash gains most from better evidence.** The Golden Prompt lifts it from 416 to 438 ($p = 0.0097$), on about 440 input tokens per question. With the full conversation it reaches 492, close to Muse's 495.



Left: accuracy against mean input tokens per question. Right: input tokens spent per correct answer.



The grid by LoCoMo category, with the four evidence arms side by side for each answer model.



Question-level overlap between HeurChain and each other arm, with the same answer model, and the per-question sign test.

Each cell gives one evidence arm to one answer model, with byte-identical prompts per arm and the gpt-6-luna judge with the pinned rubric throughout. No memory sends the question alone. Golden Prompt sends only the turns LoCoMo marks as evidence for that question, in dialogue order: an upper bound for retrieval precision, not a system. HeurChain is the latest build (date writing v2) and Mem0 is Mem0's own retrieval; both are packed to the same 16 KB evidence budget. Full conversation sends every turn of the question's conversation, about 366 KB of prompt. The Golden Prompt and full-conversation cells use the v5 stored turn text, which has no written dates. HeurChain cells for gpt-6-luna and DeepSeek V4.1 Flash average two runs, and gpt-6-luna's Mem0 cell averages the two Mem0 runs from the earlier window; every other cell is one run, and repeated runs of one model differ by 4 to 10 questions. Tokens are provider-reported, so tokenizers differ slightly by model. Tests are one-sided sign tests with one observation per question: each arm's mean correctness over its runs.

▼ Table view: every grid cell

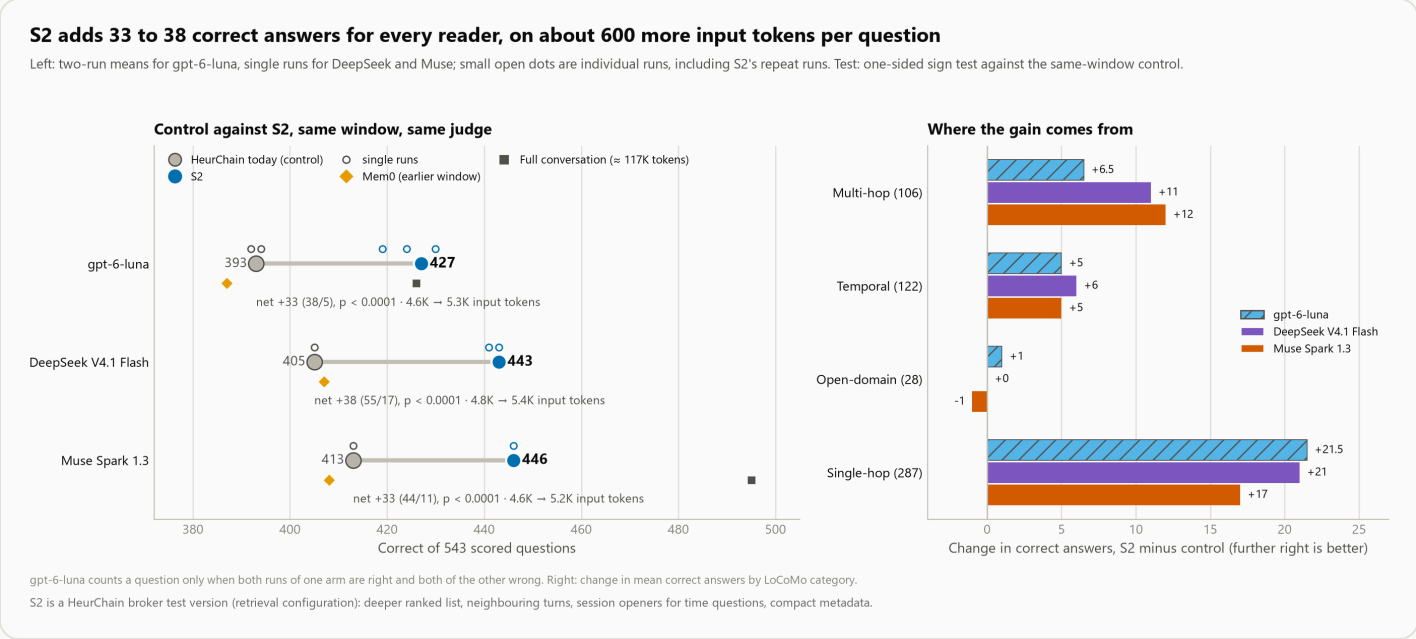
Answer model	Evidence	Correct per run	Mean	Accuracy	Multi-hop (106)	Temporal (122)	Open-domain (28)	Single-hop (287)	Input tokens per question	Input tokens per correct answer
gpt-6-luna	No memory	0	0	0.0%	0	0	0	0	112	—
gpt-6-luna	Golden Prompt	382	382	70.3%	81	50	9	242	435	618
gpt-6-luna	HeurChain	395, 399	397	73.1%	58.5	92	9	237.5	4,604	6,297
gpt-6-luna	Mem0	386, 388	387	71.3%	62	98.5	8.5	218	8,106	11,374

Answer model	Evidence	Correct per run	Mean	Accuracy	Multi-hop (106)	Temporal (122)	Open-domain (28)	Single-hop (287)	Input tokens per question	Input tokens per correct answer
gpt-6-luna	Full conversation	426	426	78.5%	75	84	9	258	116,670	148,713
Muse Spark 1.3 Contributor	No memory	0	0	0.0%	0	0	0	0	93	—
Muse Spark 1.3 Contributor	Golden Prompt	419	419	77.2%	82	88	16	233	416	539
Muse Spark 1.3 Contributor	HeurChain	413	413	76.1%	64	97	13	239	4,589	6,033
Muse Spark 1.3 Contributor	Mem0	408	408	75.1%	68	101	13	226	8,098	10,777
Muse Spark 1.3 Contributor	Full conversation	495	495	91.2%	96	111	18	270	116,808	128,135
DeepSeek V4.1 Flash	No memory	0	0	0.0%	0	0	0	0	116	—
DeepSeek V4.1 Flash	Golden Prompt	438	438	80.7%	79	97	12	250	456	565
DeepSeek V4.1 Flash	HeurChain	411, 421	416	76.6%	59	99.5	16	241.5	4,850	6,331
DeepSeek V4.1 Flash	Mem0	407	407	75.0%	63	101	14	229	8,250	11,007
DeepSeek V4.1 Flash	Full conversation	492	492	90.6%	96	108	17	271	122,953	135,698

▶ Table view: question overlap with HeurChain

Batch 1: retrieval levers and the refinement loop

Batch 1 tested retrieval levers proposed by the multi-hop and temporal ideation. It ran one date-resolver revision, then a bounded refinement loop around the winner. Every arm is a set of frozen cells for the same 701 DEV questions, judged by gpt-6-luna with the pinned rubric. Every rule was written down before its stage ran.



Left: today's HeurChain cells against S2 for each reader, with Mem0 and the full conversation for reference. Right: the change by LoCoMo category.

What batch 1 shows

1. **S2 adds 33 to 38 correct answers with all three readers.** gpt-6-luna goes from 393 to 427 (two-run means; both runs agree on 38 gains and 5 losses, $p < 0.0001$), DeepSeek V4.1 Flash from 405 to 443 (55/17) and Muse Spark 1.3 from 413 to 446 (44/11). S2 sends about 600 more input tokens per question than today's cells (4,604 → 5,273 with gpt-6-luna).

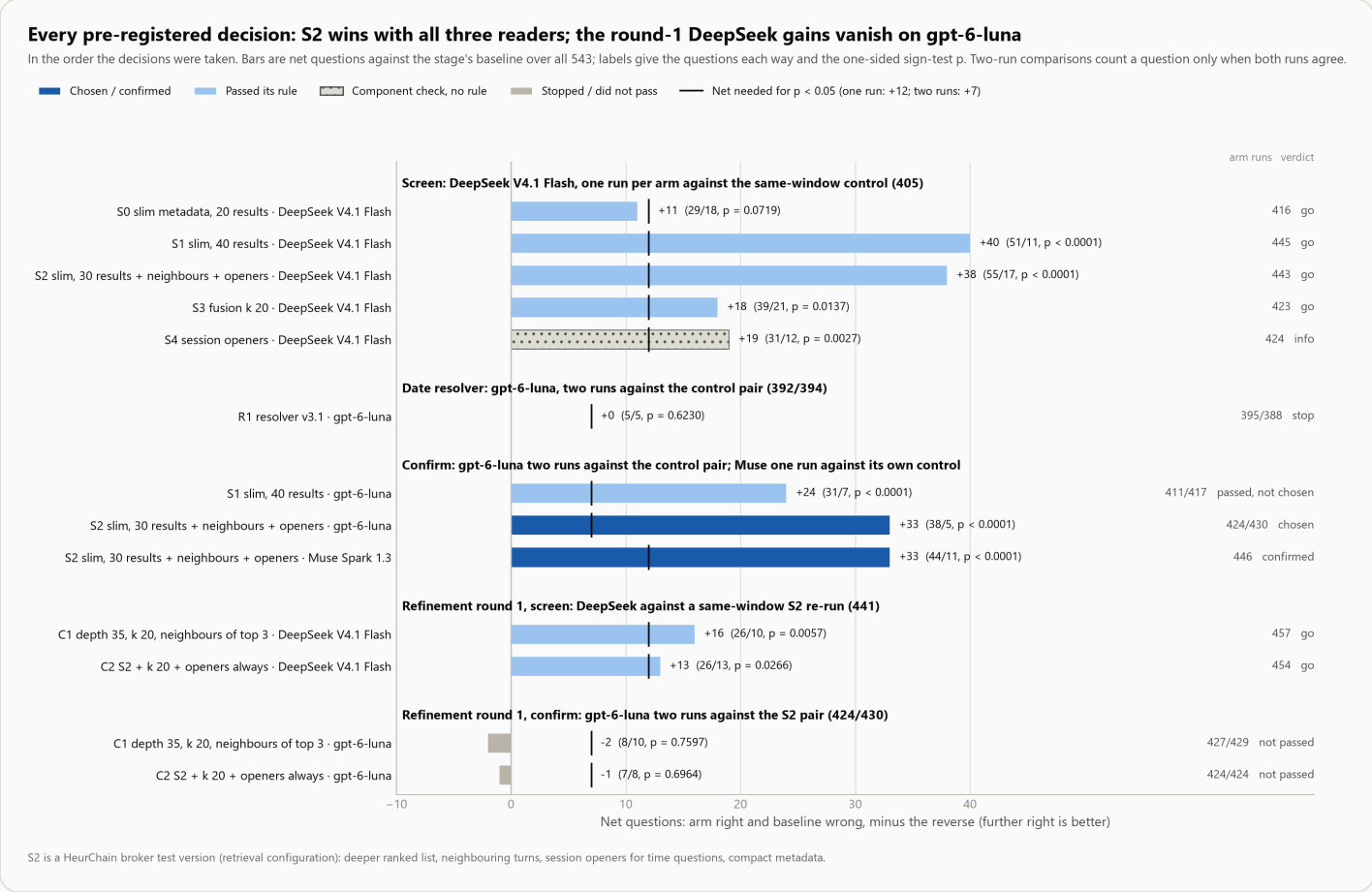
2. **On gpt-6-luna, S2 is level with reading the whole conversation.** S2's three gpt-6-luna runs score 424, 430, 419 against 426 for the full conversation, on 4.5% of its input tokens, and against 387 for Mem0 (earlier window). Muse is different: with the whole conversation it reaches 495, well above S2's 446.

3. **The gain is mostly single-hop and multi-hop.** Across the three readers S2 changes single-hop by +17 to +21.5, multi-hop by +6.5 to +12, temporal by +5 to +6 and open-domain by -1 to +1. On the questions where S2 adds a labelled turn, gpt-6-luna's both-runs-agree count is 25 gains to 0 losses.

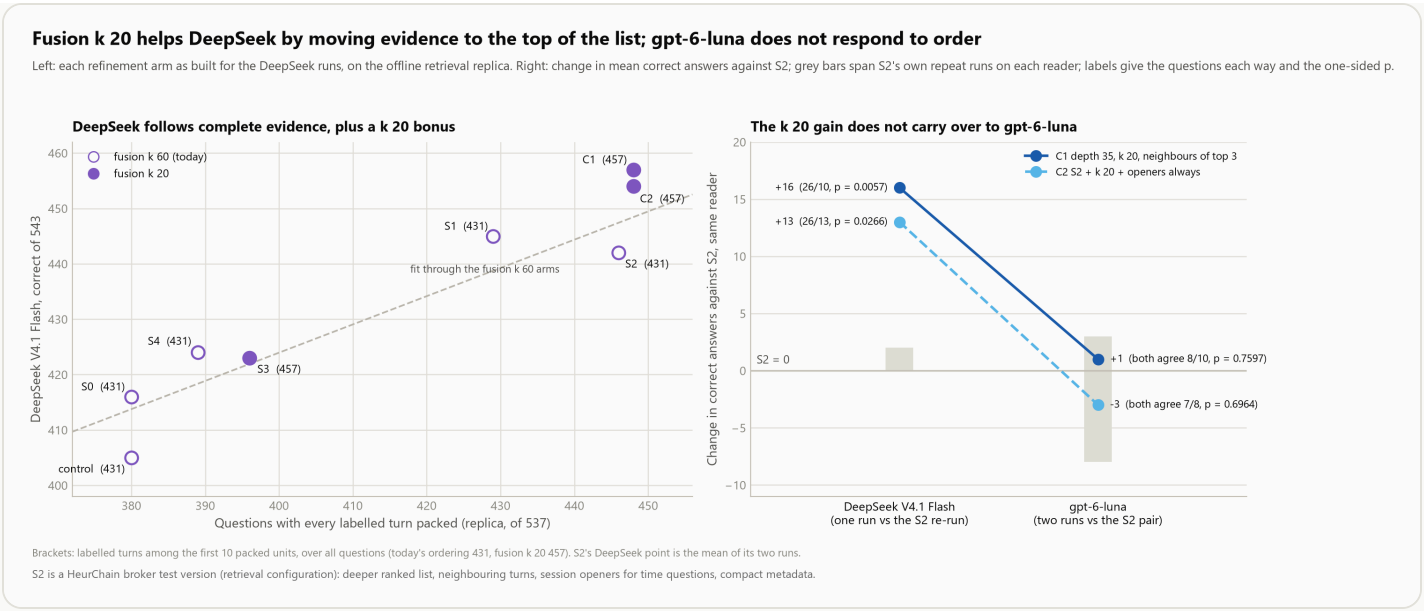
4. **Date resolver v3.1 changed nothing.** R1 scored 395/388 against the control pair 392/394: both runs agree on 5 gains and 5 losses, and its pre-registered test on the 213 prompts it changed came out at net −4. It was stopped.

5. **The refinement candidates gained on DeepSeek and lost it on gpt-6-luna.** Against a same-window S2 re-run (441), C1 scored 457 (+16, p = 0.0057) and C2 454 (+13, p = 0.0266) on DeepSeek. On gpt-6-luna both runs agree on a net of -2 and -1 against the S2 pair. On the replica, fusion k 20 puts 26 more labelled turns among the first 10 packed units (457 against 431) while complete evidence barely moves (448 against 446). DeepSeek rewards evidence near the top; gpt-6-luna reads the whole 40-unit context. The loop stopped as pre-registered and S2 stands.

6. **Batch 1 cost less than its cap.** Measured spend was $3.99 for batch 1 (cap $5) and $1.75 for refinement round 1 (cap $10); DeepSeek runs cost only the judge, about $0.03 each. Cell builds, about $0.20, are not included. The holdout release below is costed separately.



Every pre-registered test in batch 1 and refinement round 1, in the order the decisions were taken.



Left: DeepSeek score against complete evidence on the replica, fusion k 60 and k 20 arms. Right: the round-1 candidates against S2 on each reader.

Arms S0 to S4, C1 and C2 were built offline from the retrieval replica, which reproduces the broker's v2 retrieval and its 14,020 rendered units. The broker itself was not changed. R1 was regenerated by the broker from a local branch (resolver v3.1 merged with read-time rendering). S2 keeps the ranked list to 30, cuts unit metadata to the sample id, and appends the turns either side of the top 5 within a session. For questions about time it also appends the first two turns of the first three sessions. Every arm packs to the same 16 KB budget. DeepSeek screens because it costs only the judge and responds most to better evidence; gpt-6-luna confirms because it is the pinned reader. One run against one needs a net of about +12 for $p < 0.05$ on all 543. Two runs against two, counting only questions where both runs agree, need about +7. A third gpt-6-luna S2 run (419) checked for drift and fell inside its pre-set range. All batch-1 and refinement runs used the DEV split; the full write-up, pre-registrations and evaluators are published with the research notes.

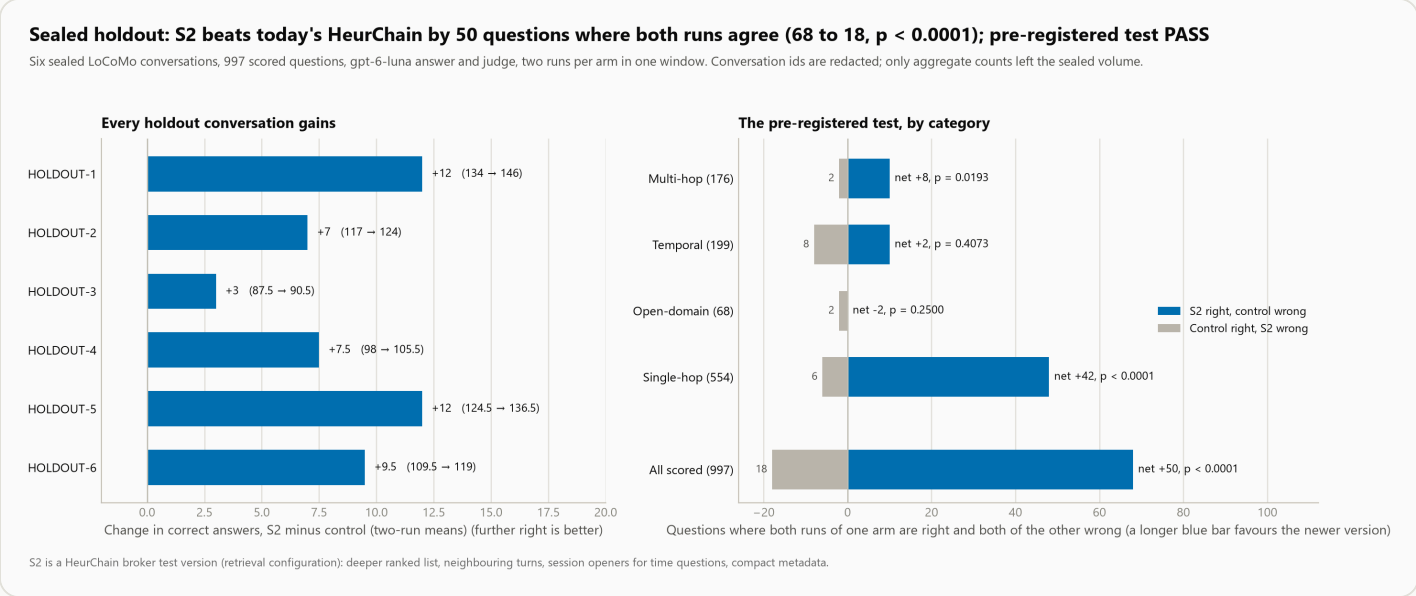
▼ Table view: S2 against today's cells, by reader

Reader	Control runs	S2 runs (incl. repeats)	Change, all	Multi-hop	Temporal	Open-domain	Single-hop	Input tokens per question
gpt-6-luna	392, 394	424, 430, 419	+34	+6.5	+5	+1	+21.5	4,604 → 5,273
DeepSeek V4.1 Flash	405	443, 441	+38	+11	+6	+0	+21	4,850 → 5,401
Muse Spark 1.3	413	446	+33	+12	+5	-1	+17	4,589 → 5,221

▶ Table view: every pre-registered test

▶ Table view: spend

Sealed holdout release



Sealed holdout release: change per redacted conversation (left) and the pre-registered both-runs-agree test by category (right).

1. **The pre-registered holdout test passed.** Over the 997 scored holdout questions S2 averages 721.5 correct against 670.5 for today's HeurChain (717, 726 against 681, 660). Where both runs of an arm agree, S2 wins 68 questions and loses 18: net +50, $p < 0.0001$, against a pre-registered bar of +10 at $p < 0.05$. No category is significantly worse, so the guardrail held.

2. **The gain is single-hop and multi-hop again; temporal is flat.** Single-hop nets +42 (48/6, $p < 0.0001$) and multi-hop +8 ($p = 0.0193$). Temporal (+2) and open-domain (-2) are within noise, as they were on DEV.

3. **Every holdout conversation gains.** Two-run means rise by +3 to +12 questions per conversation, so the result is not carried by one conversation.

4. **The cost matches DEV.** S2 sends 5,219 input tokens per question against 4,586 for today's cells (14% more). The four runs cost $3.40 measured, plus a few cents for the first minutes of one run that was pre-empted and resumed, and about $1 of compute.

5. **Runs of the same arm drift, which the design absorbs.** The two control runs score 681 and 660, further apart than on DEV. Counting only questions where both runs agree, and interleaving the arms in one window, keeps that drift out of the comparison.

One supervised release under the reproducible-bench holdout tier, run once and pre-registered before any holdout data was read. Arms: S2 built into the broker (neighbours, session_head time_question and metadata_keys, all off by default) against the same broker with those options off, which reproduces today's v2 cells exactly on DEV. Embeddings were recorded once from the pinned DEV embedder and replayed strictly; inputs were hash-checked in a separate environment, and the one-shot approval was consumed once for the four runs. One pre-registered precondition was amended with the owner's confirmation before any holdout data was used: broker S2 reproduced the replica S2 prompts on 699 of 701 DEV questions, against a bar of 700, with both differences in the replica's ranking.

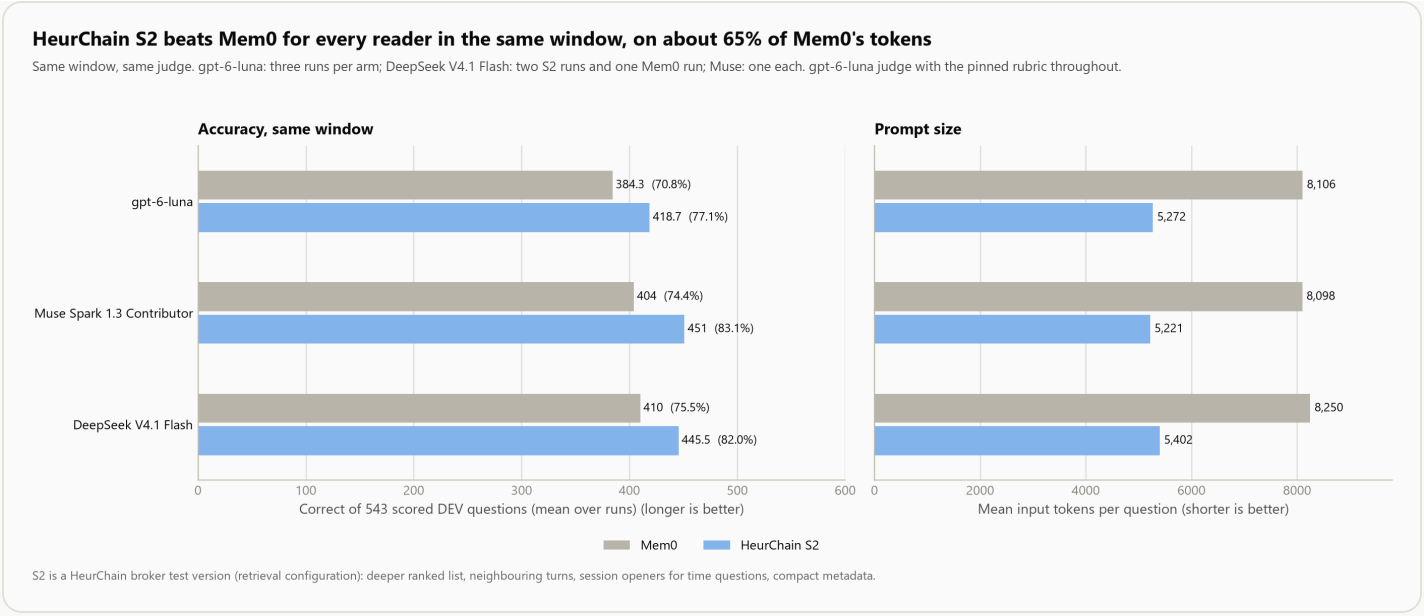
▼ Table view: the pre-registered test

Category	Questions	S2 better / control better (both runs agree)	Net	p (S2 better)	p (control better)
multi-hop	176	10 / 2	+8	p = 0.0193	p = 0.9968
temporal	199	10 / 8	+2	p = 0.4073	p = 0.7597
open-domain	68	0 / 2	-2	p = 1.0000	p = 0.2500
single-hop	554	48 / 6	+42	p < 0.0001	p = 1.0000
all scored	997	68 / 18	+50	p < 0.0001	p = 1.0000

▶ Table view: holdout runs

Same-window comparison with Mem0

Mem0 and HeurChain S2 were run in the same window: the same questions, the same answer model for each reader, the same judge and rubric, and the same evidence budget. The chart gives accuracy and prompt size; the tables give runs by category and the paired tests.



Mem0 and HeurChain S2 in the same window: correct answers (left) and prompt size (right) for three readers.

What the same window shows

1. **HeurChain S2 is ahead for every reader, on about two-thirds of the tokens.** S2 scores 418.7 against 384.3 with gpt-6-luna, 451 against 404 with Muse Spark 1.3 Contributor and 445.5 against 410 with DeepSeek V4.1 Flash, correct of 543 scored DEV questions. The paired all-question tests are $p = 0.0038$, $p < 0.0001$ and $p = 0.0013$. S2 sends about 65% of Mem0's input tokens in the same window.

2. **The lead is concentrated in single-hop questions.** With gpt-6-luna S2 is right where Mem0 is wrong on 57 single-hop questions and wrong where Mem0 is right on 14 ($p < 0.0001$); DeepSeek 45 to 12 ($p < 0.0001$), Muse 41 to 7 ($p < 0.0001$). Single-hop carries most of the aggregate difference for all three readers.

3. **Temporal questions remain Mem0's strongest category.** With gpt-6-luna Mem0 scores 99.7 on temporal against S2's 92.7 (10 to 22, $p = 0.9900$). With DeepSeek the two are level (104 against 103.5, 11 to 13, $p = 0.7294$), and with Muse S2 is ahead by three (13 to 10, $p = 0.3388$). S2 does not take the category.

4. **Multi-hop and open-domain do not separate the two systems.** Multi-hop paired counts are 23 to 21 with gpt-6-luna ($p = 0.4402$), 18 to 17 with DeepSeek ($p = 0.5000$) and 21 to 14 with Muse ($p = 0.1553$). Open-domain holds only 28 questions, so its counts are small and descriptive.

5. **These numbers come from a later window than the six-system table.** The six-system table's gpt-6-luna Mem0 mean is 387.0 from the original window; in this matched window it is 384.3. S2's mean here matches its batch-1 value of about 419. Read the same-window comparison against the same-window numbers, not against the older Mem0 column.

Runs by category

Correct of each category's questions, mean over runs.

▼ Table view: runs by category

Reader	Memory	Runs	Correct per run	Multi-hop	Temporal	Open-domain	Single-hop	All	Input tokens/question
gpt-6-luna	Mem0	3	385 / 378 / 390	59.7	99.7	8	217	384.3	8,106
gpt-6-luna	HeurChain S2	3	425 / 412 / 419	62	92.7	9	255	418.7	5,272
Muse Spark 1.3 Contributor	Mem0	1	404	67	99	11	227	404	8,098
Muse Spark 1.3 Contributor	HeurChain S2	1	451	74	102	14	261	451	5,221
DeepSeek V4.1 Flash	Mem0	1	410	67	104	13	226	410	8,250
DeepSeek V4.1 Flash	HeurChain S2	2	445 / 446	69.5	103.5	15.5	257	445.5	5,402

Paired tests

S2 better / Mem0 better, one-sided sign test on each question's mean correctness over runs.

▼ Table view: paired tests

Reader	Comparison	Multi-hop	Temporal	Open-domain	Single-hop	All
gpt-6-luna	HeurChain S2 vs Mem0	23 / 21, p = 0.4402	10 / 22, p = 0.9900	4 / 3, p = 0.5000	57 / 14, p < 0.0001	94 / 60, p = 0.0038
Muse Spark 1.3 Contributor	HeurChain S2 vs Mem0	21 / 14, p = 0.1553	13 / 10, p = 0.3388	3 / 0, p = 0.1250	41 / 7, p < 0.0001	78 / 31, p < 0.0001
DeepSeek V4.1 Flash	HeurChain S2 vs Mem0	18 / 17, p = 0.5000	11 / 13, p = 0.7294	4 / 2, p = 0.3438	45 / 12, p < 0.0001	78 / 44, p = 0.0013

Same window, same judge and rubric, same evidence budget. gpt-6-luna ran three times per arm; DeepSeek V4.1 Flash twice on S2 and once on Mem0; Muse Spark 1.3 Contributor once per arm. Mem0 ran a few hours after the S2 runs. Paired tests are one-sided sign tests on each question's

mean correctness over runs, the test used in the memory grid; the counts are the per-question wins and losses and the p-values are recomputed from them.

Findings

1. **HeurChain + date writing posts the highest two-run mean, but the lead is small.** It scores 393.5 of 543 (72.5%), against Mem0 387.0 (71.3%) and plain HeurChain 383.5 (70.6%). That is 6.5 questions over Mem0, from a different run window, so it is a small lead, not a win. All three sit within 10 questions of each other.

2. **Mem0 still leads temporal questions, and date writing closes about half that gap.** Mem0 scores 98.5 of 122 temporal questions against 89.5 with date writing and 80.0 for plain HeurChain. The date-writing column's own same-window control scored 82.5, so date writing added 7.0 questions per run (9.0 over four runs on the broker revision), roughly half of the 16-question gap between that control and Mem0.

3. **HeurChain leads single-hop, and date writing leaves single-hop and multi-hop essentially unchanged.** HeurChain scores 235.0 single-hop and 235.5 with date writing, against Mem0's 218.0 of 287 — a 17-question HeurChain lead. Multi-hop moves 56.5 to 60.5 in this window (+4.0) and single-hop by 1.5; the four-run replication shows no multi-hop change (58.0 without date writing, 57.8 with it) and no single-hop change (233.2 without, 233.8 with).

4. **Zep and Hindsight score far lower under these matched settings, because they abstain far more.** Zep reaches 204.5 of 543 and Hindsight 265.0, against 383.5–393.5 for the top three. Over all scored questions Zep answers UNKNOWN 47.9% of the time and Hindsight 37.6%, against 16.9% for HeurChain and 18.0% for Mem0 (both runs pooled). Neither is a fair reading of Zep or Hindsight as published.

5. **OpenViking, run through its own LoCoMo memory workflow, ties Zep at 204.5 and abstains on half the questions.** It answers UNKNOWN 52.1% of the time (both runs pooled). Its extracted memories are accurate, with dated event summaries that quote the original lines, but they cover only part of each conversation: across all 701 questions retrieval reached 51 to 117 distinct memory files per conversation, against 369 to 663 turns. Each event file also repeats its body twice, which uses up evidence space. The method was its documented single-search flow: an 8-message commit per chunk, then find over the conversation's memories, top 20, no rerank. That gives OpenViking a larger evidence allowance than its own default of 4,000 characters. OpenViking's published 80-83% comes from agent integrations that search repeatedly and can read archived sessions, with a Doubao model answering and judging, so it is not comparable.

6. **The two HeurChain arms agree far more with each other than any pair of systems.** HeurChain and date writing agree on 86.7% of scored questions; HeurChain and Mem0 on 68.9%; HeurChain and Hindsight 56.5%; HeurChain and Zep 50.1%. Zep agrees with every other system on only 48–50% of questions.

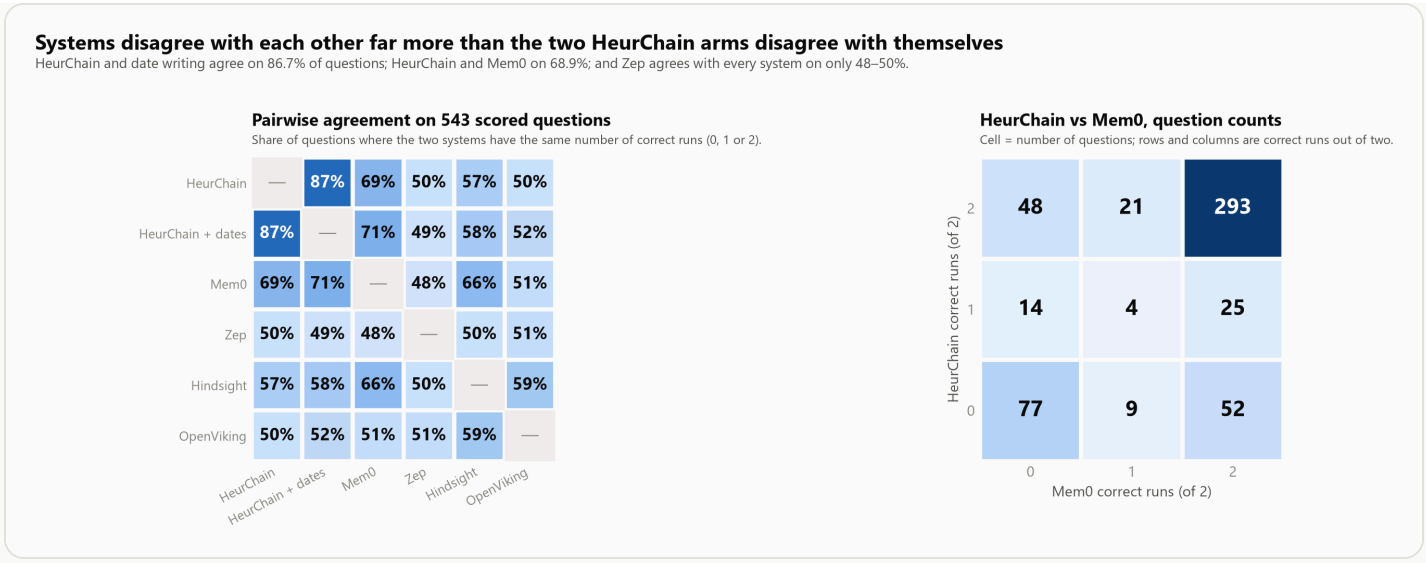
7. **Run-to-run noise is large next to the gaps between systems.** Flips between the two runs are 43 questions for HeurChain, 37 for date writing, 34 for Mem0, 43 for Zep and 22 for Hindsight — 4–8% of the scored set. The top three means span 10.0 questions, while each system's own two runs differ by 1 to 6 questions. On this DEV split the ranking of the top three is not separable from run noise.

8. **Abstention is concentrated where the systems are weakest.** On temporal questions Zep answers UNKNOWN 57% of the time and Hindsight 36%, against 12% for Mem0. On single-hop, Zep abstains 44% of the time; on open-domain every system abstains on more than 60% of answers, and Zep on 85.7%.

9. **Answer length and prompt size separate the systems as much as accuracy does.** Median answer length is 48 characters for HeurChain (44 with date writing), 30 for Mem0, 15 for Hindsight and 7 for Zep — the last two because an UNKNOWN answer is 7 characters. Median prompt size is 14.2 KB for both HeurChain arms, 16.0 KB for Mem0, 16.1 KB for Hindsight and 12.4 KB for Zep.

10. **Cost per run tracks prompt caching, not accuracy.** $0.2556 per HeurChain run, $0.2556 for date writing, $0.4456 for Mem0, $0.4768 for Zep and $0.2963 for Hindsight. The spread within a system can exceed the spread between systems: one HeurChain run cost $0.4396 and the other $0.0717 for the same prompts, while Hindsight's two runs were $0.0759 and $0.5167. Compare systems on token volume, not on these dollar figures.

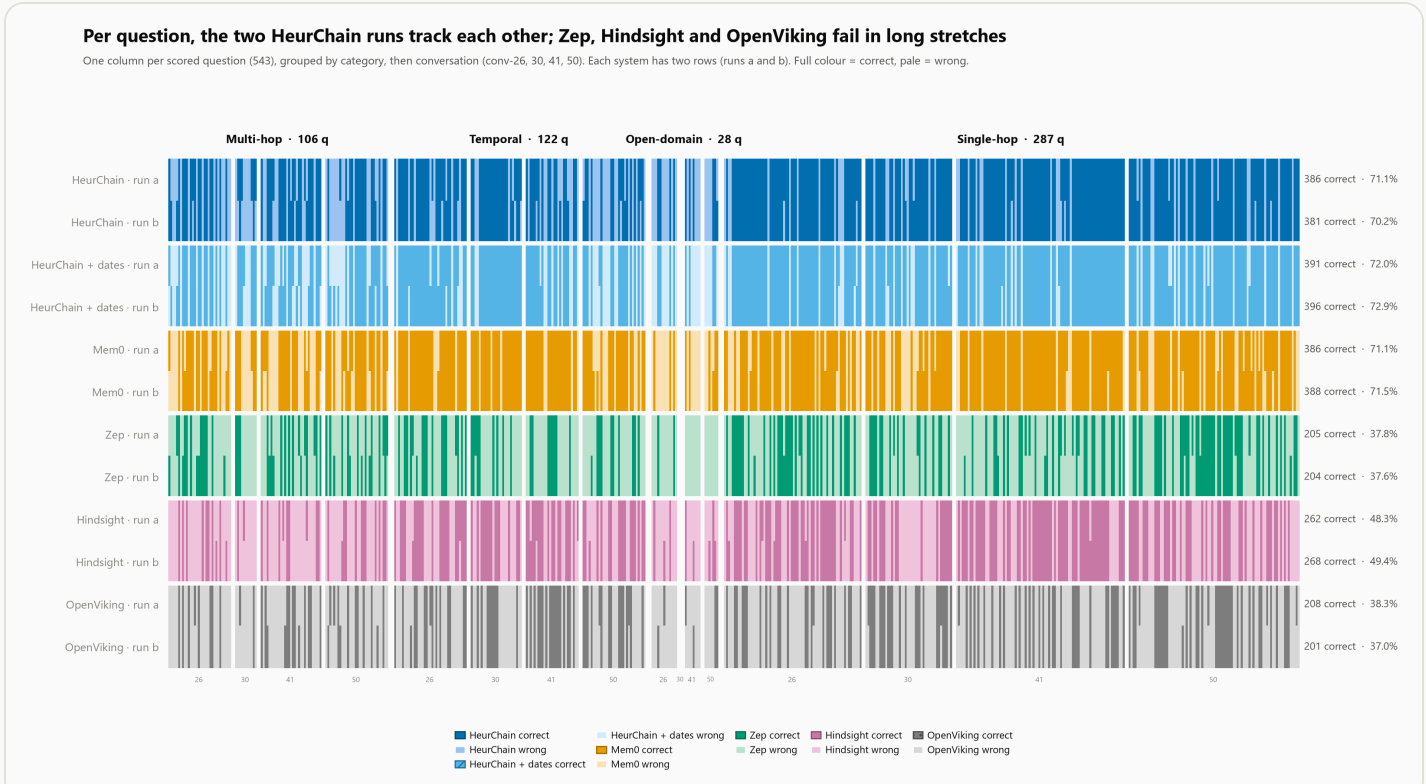
Agreement



Left: pairwise agreement between all six systems. Right: the original HeurChain vs Mem0 question-count grid.

The left grid is the share of the 543 scored questions on which two systems have the same number of correct runs (both 0, both 1, or both 2); the diagonal is therefore trivially 100% and is shown as a dash. The right grid is the original HeurChain vs Mem0 count matrix: 77 questions are wrong for both, 293 are right for both, and the diagonal sums to 374.

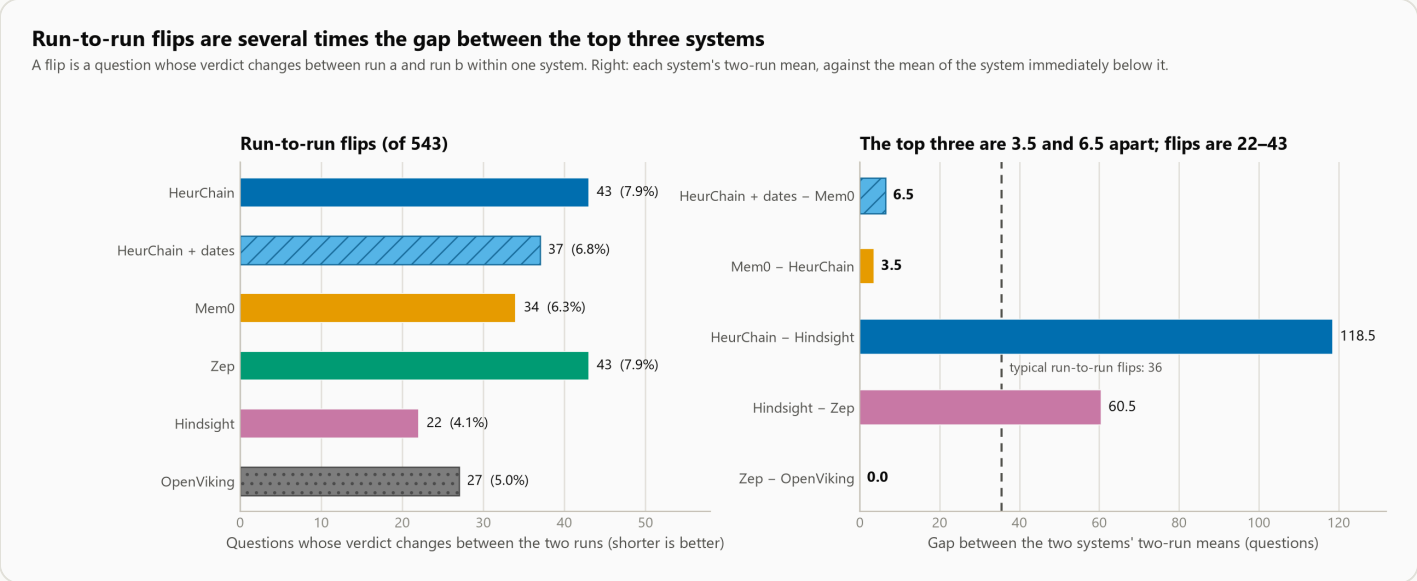
Question-by-question array



Every scored question as one thin column, two rows per system; full colour is correct, pale is wrong.

Columns are ordered by category (1–4), then by conversation (conv-26, 30, 41, 50), then by question ordinal. Category boundaries are wide surface gaps; conversation boundaries are narrower. Each row prints its correct count and accuracy at the right edge.

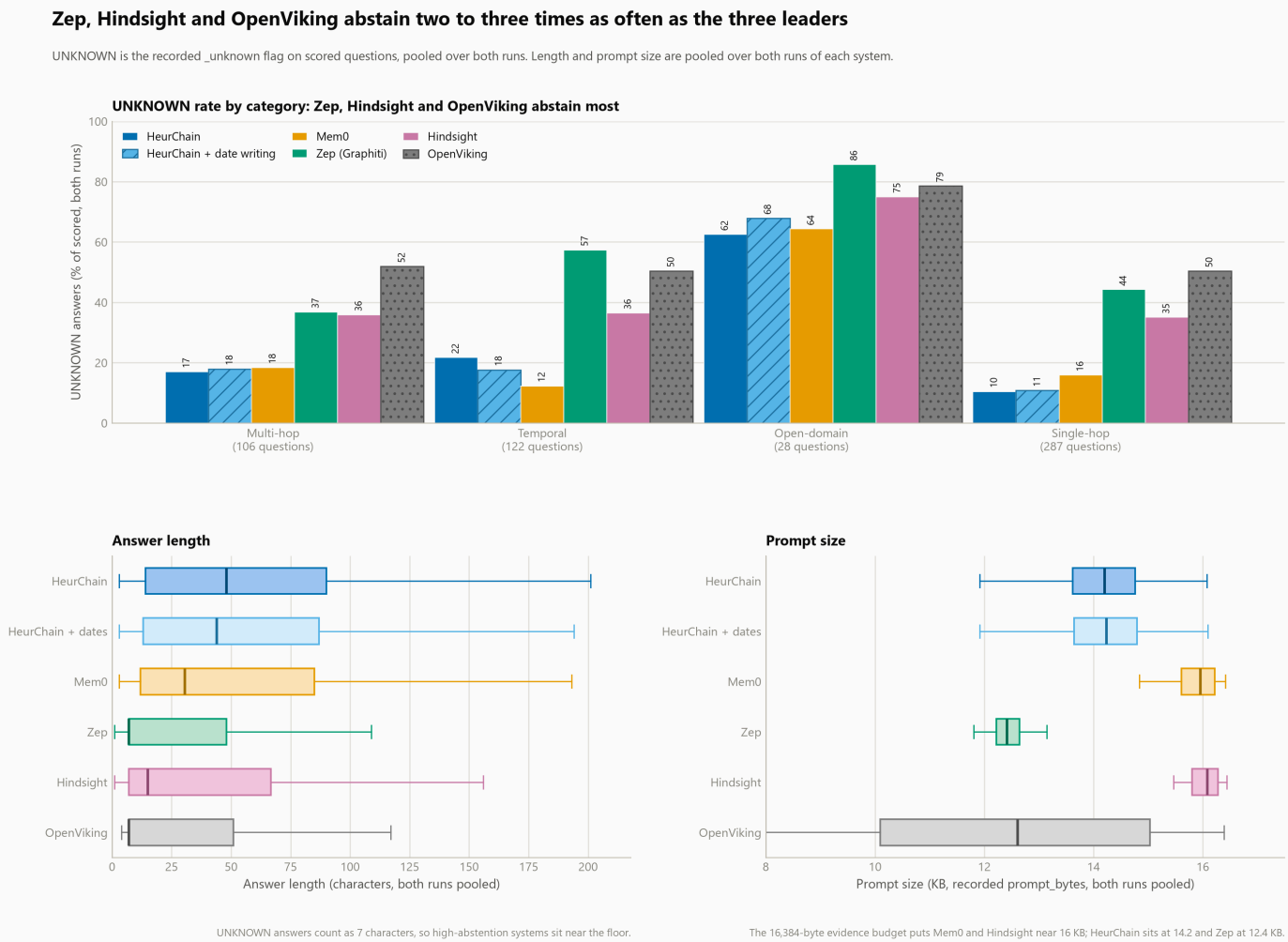
Stability



Run-to-run flips per system beside the gap between each system's two-run mean and the mean of the system below it.

A flip is a question whose verdict changes between run a and run b within one system. The right panel ranks the systems by two-run mean and shows each system's gap to the mean of the system immediately below it, against a dashed line at the median flip count (37).

Answer behaviour



UNKNOWN rate by category, answer length, and prompt-size distribution.

Length and UNKNOWN are pooled over both runs of each system. UNKNOWN is the recorded unknown flag on scored questions. Answer length is the recorded answer length in characters; an UNKNOWN answer contributes 7 characters. Prompt size uses the recorded prompt bytes. Outliers are hidden.

Limits

- DEV split only: four conversations (conv-26, conv-30, conv-41, conv-50) and 543 scored questions in categories 1–4. The 158 adversarial category-5 questions were answered but not judged and are excluded from every accuracy number.
- Zep used Graphiti 0.31.2 edge search only (top 20 fact edges), whereas Zep's own LoCoMo method also passes entity summaries and its own prompt. Hindsight's retain kept about 0.3 facts per turn. These are not fair readings of Zep or Hindsight as published.

- Plain HeurChain and the date-writing column ran in different windows; read the date-writing effect against its same-window control.
- No significance claims are made across conversations: there are only four, so per-conversation cells are descriptive.
- The p-values quoted for date writing are the supplied paired results for that comparison. The same-window p-values were recomputed from the per-question win/loss counts with the same one-sided sign test.
- The same-window S2 and Mem0 runs were not interleaved; Mem0 ran a few hours after S2, so a window effect cannot be ruled out.

Method notes

Accuracy = correct / n for a single run; a system mean is the average of its two runs. Costs are the recorded per-run values, and mean correct counts can be fractional because they average two runs.

Per-question answers, labels, token counts and prompt sizes come from the study's recorded per-question table; per-run totals and category and conversation means come from its per-run summary. Nothing was recomputed from a different source.

Zep's first run was preempted at 451 of 701 answers and rerun in full; the summary uses that rerun.

The same-window section uses matched runs and reports paired tests computed from the per-question counts, as described in its note.