

✓ Pre-registered design, deviations disclosed

📄 Phase 1: 96 replicates

📄 CC BY 4.0 & MIT

📅 2026-10-10

Long-Range Gains, Short-Range Costs: Structured State versus Rolling Summary Memory in a 224-Turn AI-Narrated Story

Wei Wong, written with Claude

Disclosure: the text was drafted by Claude (Anthropic), which also wrote the analysis scripts, ran the experiments through coding agents under the author's direction, and acted as one of three raters of the answer judge. The author set the question and the rules, labelled 50 answers by hand, and reviewed the claims. Every number is generated by the scripts in the archive listed in Appendix D. Version of 2026-10-10.

Abstract

Long-running language-model narrators must carry facts across hundreds of turns inside a bounded prompt. We compare three ways of doing so on a 224-turn story with nine planted facts probed at later turns: rolling-summary compaction (A), a structured-state layer of extracted durable facts with retrieval (B), and a raw sliding window (C), plus a second window enlarged to B's context size and run cost (C-large). Across 96 runs on six prompt sequences, B recalled 0.671 of the planted facts at turn 224 against 0.102 for A (+0.569, 95% t-interval 0.480 to 0.659; B ahead in all six sequences; exact sign-flip $p = 0.031$, the smallest value six sequences allow, unadjusted). The raw windows recalled nothing, because no planted fact lay inside them, so that comparison cannot say whether B's gain comes from its architecture or from reach; putting the whole story in the prompt recovered 52 of 54 facts, so the zeros belong to the window and not to the model. By the age of the fact, A decays from 0.89 to 0.09, B is flat near 0.75 but lower than A and C-large on fresh facts (0.73), and the enlarged window is perfect on facts under ten turns old. The result rests on six sequences,

one generation model, nine facts that are identical in every sequence and suit B's extractor, and a test chosen after the registered mixed model failed to converge; changing how an answer judge scores paraphrases moves B by at most 0.03.

1. Introduction

A story told one turn at a time by a language model outgrows its prompt. The designer must decide what to keep: the most recent turns verbatim, a running summary, or a separate store of extracted facts that is searched when needed. These choices are usually made by feel, and rarely tested over a horizon long enough for the earliest details to matter.

We report a pre-registered comparison of such choices in a live system, a public collaborative storytelling site whose memory layer can be swapped without touching the narrator, with a replay harness that feeds identical prompt sequences to each variant. The question is narrow: when a fact is planted early, how often can the system still produce it hundreds of turns later under a fixed prompt budget? We contribute (i) a measured comparison of a rolling summary and a structured-state layer over 224 turns; (ii) a budget-matched raw-window control and a plain account of why it cannot fail by construction, with a direct check of that floor; (iii) an age-of-fact analysis showing where each design wins and loses; and (iv) a full account of where the analysis departed from its pre-registration, including changes made after the data were seen.

2. Setup

Task. A narrator writes 224 turns (7 days of 32), each following a prompt from a fixed sequence; six sequences were used. Nine facts are planted at turns 11, 16, 40, 55, 72, 105, 140, 175 and 200 (a symbol on a stamp, a pebble, a drink, a snack, an extra day of the week, a hidden device, a formula, a transit pass, a platform number). **The same nine facts appear at the same turns with the same answers in all six sequences**, so there are nine distinct facts, not 54. In an adversarial condition two prompts per day are replaced by distracting injections.

Arms. All arms use `gemini-3.5-flash-lite`. **A** keeps the five most recent turns verbatim plus an LLM-written running summary. **B** keeps extracted durable facts and state files with a retrieval-only chronicle (retrieval embeds with `gemini-embedding-2`, not production's `gemini-embedding-001`; top-3 overlap 0.556 on a 39-turn sample). **C** is a raw window capped at 10 turns or 16,000 characters (real window 4 to 7 turns). **C-large** caps at 30 turns or 32,000 characters (about 13 turns) and was added after Phase 1 to match B's context size and cost. B's context grows from 18.9k to 30.5k characters, so it matches C-large only near the end.

Probing. At each checkpoint the system is asked about each planted fact, conditioned on the arm's context. Answers are scored by exact match against a canonical string, then by an LLM judge (three votes) for paraphrases; UNKNOWN is a miss. Retention is the fraction of planted facts answered correctly. Pre-registered checkpoints are turns 32, 64, 128 and 224; the files also hold probes every 16 turns, used only in the exploratory age analysis.

Runs. 72 runs for A, B and C (6 sequences, benign and adversarial, 2 replicates; \$75.05) and 24 for C-large (\$28.23). No run was degraded and no fallback fired. Replicates share their prompt sequence, so **the sequence (n = 6) is the unit with real independence**: every analysis averages the four runs of a sequence.

3. Results

Table 1. Context, cost and retention (mean of 24 runs per arm).

ARM	MEDIAN CONTEXT CHARS AT TURN 224	COST PER RUN (USD)	RETENTION AT 32	64	128	224
A	11878	1.10	0.604	0.438	0.188	0.102
B	30522	1.16	0.875	0.750	0.771	0.671
C	15180	0.87	0.000	0.031	0.000	0.000
C-large	31386	1.18	0.104	0.333	0.042	0.000

3.1 Confirmatory endpoint: turn 224

Table 2. Paired contrasts at turn 224 (n = 6 sequences).

CONTRAST	MEAN DIFFERENCE	95% T-INTERVAL	SEQUENCES POSITIVE	EXACT SIGN-FLIP P (TWO-SIDED)
B - A	+0.569	[0.480, 0.659]	6 of 6	0.031
B - C-large	+0.671	[0.597, 0.746]	6 of 6	0.031
A - C	+0.102	[0.036, 0.167]	5 of 6 (1 tie)	0.0625

B recalled 0.671 and A 0.102; B was ahead in all six sequences and in all 12 sequence-by-condition cells (Figure 1, [fig1_trajectory.svg](#)). The exact p of 0.031 is the smallest a six-sequence test can give and is unadjusted; Holm over the three registered contrasts gives 0.094. We report the effect size and interval as the main evidence and read the p-value as a statement about ordering.

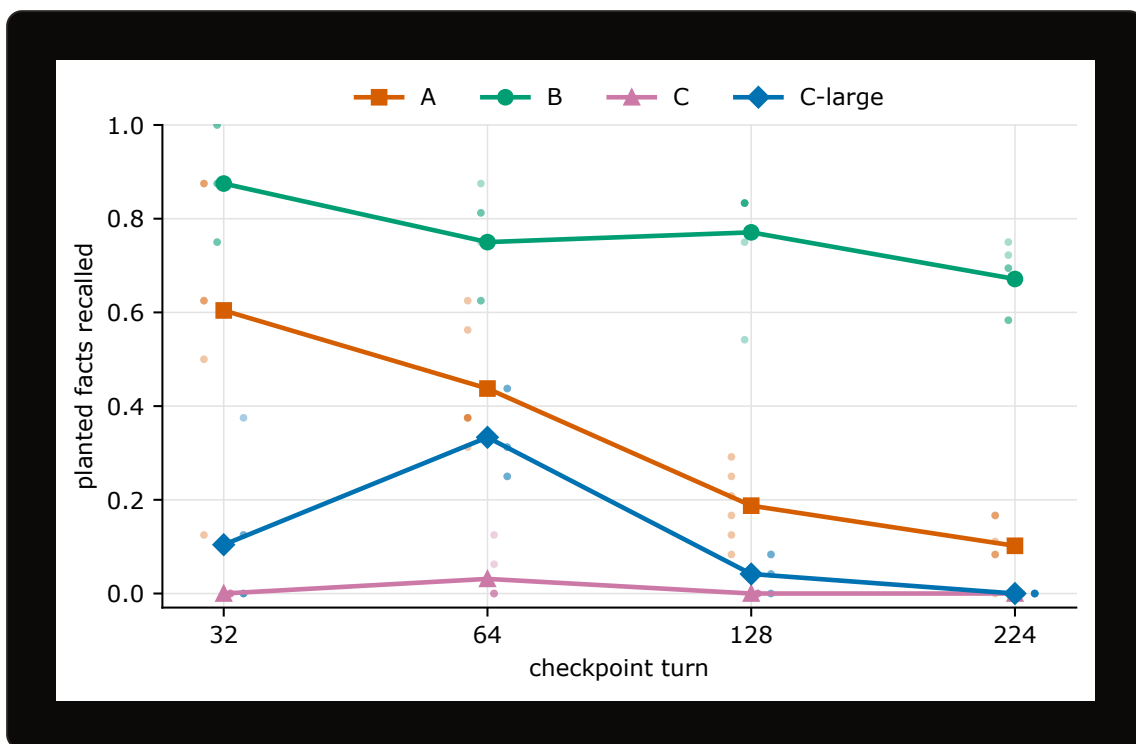


Figure 1. Retention across checkpoints for all four arms. Lines show the mean of 24 runs (6 sequences $\times$ 2 conditions $\times$ 2 replicates); error bands show 95% t-intervals across the 6 sequences. [Vector PDF](#)

B against the budget-matched window. B exceeded C-large by 0.671 in all six sequences, which meets the rule written before those runs. This does not show what it appears to. C-large's window at turn 224 spans about 13 turns and the newest planted fact is 24 turns old, so its 0.000 was a floor fixed by the turn length (about 2,400 characters). Reaching the oldest fact from turn 224 would take a window of about 213 turns, roughly 515k characters; that arm was not run as a narrator.

3.2 Where each design wins: age of the fact

Pooling every probe in the files (14 checkpoints from turn 16 to 224) and grouping by the gap between planting and probing gives Table 3 (Figure 3, `fig3_age.svg` ; exploratory, estimates only).

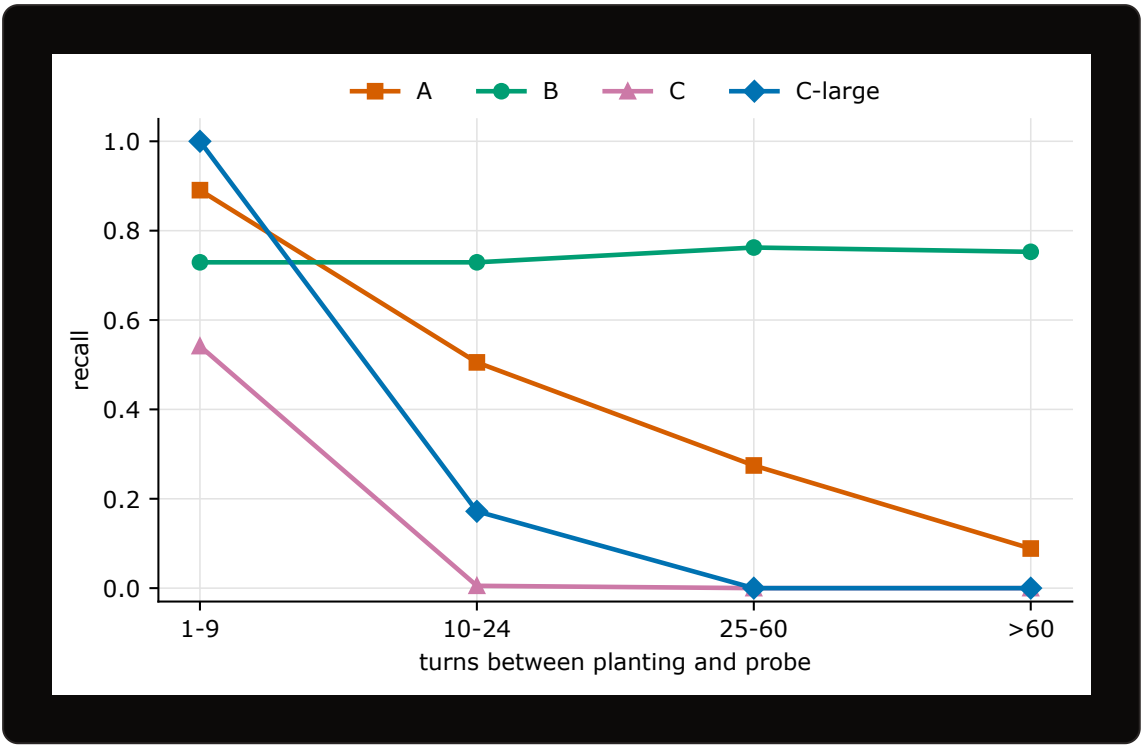


Figure 3. Recall by the age of the fact at the probe. Window arms (C, C-large) excel at low age; Variant B maintains steady recall across all horizons. [Vector PDF](#)

Table 3. Recall by the age of the fact at the probe.

ARM	1-9 TURNS	10-24	25-60	> 60
A	0.891	0.505	0.275	0.089
B	0.729	0.729	0.762	0.753
C	0.542	0.005	0.000	0.000
C-large	1.000	0.172	0.000	0.000

C-large answered all 192 probes under ten turns old correctly. A decays smoothly with age. B does not decay; it is flat near 0.75, and on the freshest facts it is the worst of A, B and C-large (0.729). B's advantage at turn 224 is therefore an advantage at long range, bought with a lower ceiling at short range.

3.3 Checking the floor: the whole story in the prompt

To test the raw-window floor directly, the saved 224-turn story of one Variant B run per sequence (benign, replicate 1) was put into the prompt whole (about 347,000 to 363,000 characters, 69,000 to 72,000 prompt tokens as billed) and the same nine questions were asked at turn 224 with the same model, prompt template and scoring. The model answered 52 of 54 correctly by the judge's verdict (per sequence 9, 8, 9, 9, 9, 8 of 9); 48 were exact matches, and of the four paraphrases the judge accepted, three are borderline or doubtful in our label checks, so the count lies between 48 and 52. The two misses were UNKNOWN answers. The 54 questions cost \$1.141 (usage metadata, confirmed against Cloud Monitoring). The raw-window zeros therefore come from the window and not from the model or the question. This is a ceiling for lookup from context and not a full-context arm: the probe reads a finished story, it does not generate 224 turns with a prompt growing to 70k tokens, and the questions are answered by lookup.

3.4 How much does the judge matter?

Of 201 probe answers decided by the judge (neither an exact match nor UNKNOWN), we compared its verdicts with three other raters: the drafting model (Claude Sonnet 5.5), the author (a human, 50 answers), and an independent Claude Opus 5.5 instance (all 84 distinct answers) given a three-label rubric and blind to the others (Appendix B). The raters disagree about answers that name

the object but omit a distinguishing detail (pairwise kappa 0.06 to 0.48 under the strict standard). Replacing the judge's verdict by any rater's, under either standard, leaves B at turn 224 between 0.639 and 0.671 and B minus A between +0.537 and +0.569; A and the raw windows do not change. The judge leans toward leniency, and a retention figure that depends on borderline answers should be read with a range of about three points.

4. Deviations from the pre-registration

Before any data: two replicates per cell instead of the planned 15 to 20 (budget); the Variant B layer ablations and a chained-attack condition were dropped; B embeds with `gemini-embedding-2` (the project's quota for the production model was 5 requests per minute). After the Phase 1 data: the registered H1 (B beats A and C, by a linear mixed model with Holm contrasts) could not be tested as registered because the mixed models did not converge (boundary fit; the near-zero variance is the sequence variance, 0.000522); we moved inference to the sequence level with an exact sign-flip test and narrowed the headline to B against A, after the means were seen. Before the Phase 1b data but after Phase 1, C-large was added with a decision rule fixed in advance (B minus C-large at least 0.20 and B above it in at least 5 of 6 sequences). Premise fidelity and day closure were 1.0 in all 384 cells, so H2 could not be tested. Full list in Appendix A.

5. Limitations

(1) Six sequences: the smallest exact p is 0.031 and the Holm-adjusted value is 0.094. (2) The inference was chosen after the data. (3) Nine distinct facts, identical across sequences, and well suited to B's extractor (eight of the nine are held items or secrets, which its promotion rule keeps; the one location fact, turn 200, is the one B fails, recalled in 3 of 24 runs); the bias favours B and is not bounded by the data. (4) The raw-window baselines score zero at turn 224 by construction; the whole-story check is a ceiling for lookup, not a full-context narrator. (5) One generation model, judges from the same family, a synthetic story, and no pre-registered kappa check; Section 3.4 bounds the effect of the judge at about three points. (6) B embeds with a different model than production. (7) B costs 6% more than A and 33% more than C; parity holds only with C-large.

(8) B's ablations were not run and its assembled contexts were not saved. (9) The adversarial condition was weak (two replaced prompts per day) and the runner retried rate-limit errors silently (3,071 retried responses in the Phase 1b window by cloud monitoring); we make no robustness claim and report no latency. (10) Exploratory results (trajectory, age, per-fact, condition, repetition) carry no significance claims.

6. Related work

Long-context models degrade when relevant information sits in the middle of a long input (Liu et al., 2023), nominal context length overstates usable length on multi-fact tasks (Hsieh et al., 2024), and retrieval and long context trade accuracy against cost (Li et al., 2024); long-document benchmarks (Bai et al., 2023; An et al., 2023; Dong et al., 2023; Li et al., 2023; Shaham et al., 2022; Kočíský et al., 2017) test comprehension of a fixed input, whereas our document is generated turn by turn and the memory is part of the system under test. Agent-memory work stores and retrieves experience: MemGPT (Packer et al., 2023), generative agents (Park et al., 2023), MemoryBank (Zhong et al., 2023), Think-in-Memory (Liu et al., 2023b), CoALA (Sumers et al., 2023) and Reflexion (Shinn et al., 2023). Recursive summarisation (Wu et al., 2021) and RecurrentGPT (Zhou et al., 2023) are nearest to variant A; story-generation systems keep coherence with outlines and plot state (Rashkin et al., 2020; Yang et al., 2022a, 2022b). We claim no new memory design; the contribution is a controlled, pre-registered long-horizon comparison with its analysis choices disclosed.

- An et al., 2023. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. [arXiv:2307.11088](#) .
- Bai et al., 2023. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. [arXiv:2308.14508](#)  (ACL 2024 per page comments).
- Dong et al., 2023. BAMBOO: A Comprehensive Benchmark for Evaluating Long Text Modeling Capacities of Large Language Models. [arXiv:2309.13345](#)  (LREC-COLING 2024 per page comments).
- Hsieh et al., 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? [arXiv:2404.06654](#)  (COLM 2024).
- Kočíský et al., 2017. The NarrativeQA Reading Comprehension Challenge. [arXiv:1712.07040](#) .

- Li, J. et al., 2023. LooGLE: Can Long-Context Language Models Understand Long Contexts? [arXiv:2311.04939](#) .
- Li, Z. et al., 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. [arXiv:2407.16833](#)  (EMNLP 2024 industry track).
- Liu, N. F. et al., 2023. Lost in the Middle: How Language Models Use Long Contexts. [arXiv:2307.03172](#)  (TACL).
- Liu, L. et al., 2023b. Think-in-Memory: Recalling and Post-thinking Enable LLMs with Long-Term Memory. [arXiv:2311.08719](#) .
- Packer et al., 2023. MemGPT: Towards LLMs as Operating Systems. [arXiv:2310.08560](#) .
- Park et al., 2023. Generative Agents: Interactive Simulacra of Human Behavior. [arXiv:2304.03442](#) .
- Rashkin et al., 2020. PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking. [arXiv:2004.14967](#)  (EMNLP 2020).
- Shaham et al., 2022. SCROLLS: Standardized Comparison Over Long Language Sequences. [arXiv:2201.03533](#)  (EMNLP 2022).
- Shinn et al., 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. [arXiv:2303.11366](#) .
- Sumers et al., 2023. Cognitive Architectures for Language Agents. [arXiv:2309.02427](#)  (TMLR camera-ready per page comments).
- Wu et al., 2021. Recursively Summarizing Books with Human Feedback. [arXiv:2109.10862](#) .
- Yang et al., 2022a. DOC: Improving Long Story Coherence With Detailed Outline Control. [arXiv:2212.10077](#)  (ACL 2023). Yang et al., 2022b. Re3: Generating Longer Stories With Recursive Reprompting and Revision. [arXiv:2210.06774](#)  (EMNLP 2022).
- Zhong et al., 2023. MemoryBank: Enhancing Large Language Models with Long-Term Memory. [arXiv:2305.10250](#) .
- Zhou et al., 2023. RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text. [arXiv:2305.13304](#) .

Appendix

Appendix A. Pre-registration and deviations (full)

The design, hypotheses and analysis were fixed in the pre-registration document (`plan.md` in the archive) before the data. We list every departure, with its timing.

Before any data (pre-data amendments). Two replicates per cell instead of the 15 to 20 planned (budget). The Variant B layer ablations and a chained-attack condition were dropped, so those hypotheses are untested. Variant B embeds with `gemini-embedding-2` instead of production's model, because the project's quota for the latter was 5 requests per minute.

After the Phase 1 data (post-data). The registered H1 was that B beats A and C, tested by a linear mixed model with Holm-adjusted contrasts and a run-level Wilcoxon check. The mixed models did not converge (boundary fit; the near-zero variance component is the sequence variance, 0.000522). Because replicates share their sequence, we moved inference to the sequence level and used an exact two-sided sign-flip test, and we narrowed the headline claim to B against A. This was decided after the means were seen. Other checkpoints, H2 to H5 and every adversarial-robustness question were declared exploratory, with estimates only. Variant C's real window was found to be 4 to 7 turns, so its 0.00 is a floor by construction.

Before the Phase 1b data (pre-data, but after Phase 1). Variant C-large was added, with a decision rule written before its runs: "B's advantage survives budget-matching" if the mean difference B minus C-large is at least 0.20 and B is above C-large in at least 5 of 6 sequences. The rule was fixed after the Phase 1 results were known.

Appendix B. Checking the judge

Label check of the judge. The judge decided 201 probes in the 96 runs (answers that were neither an exact match nor UNKNOWN). The drafting model (the same kind of model as the judge) labelled every distinct such answer right, borderline or wrong from the question and canonical answer alone, blind to the judge's verdict (`docs/paper_audit/cc_labels.json`, `cc_label_check.py`). This is a second model's reading, not human validation. Counting borderline answers (those that name the

object without its distinguishing detail, for example "a smooth river pebble" for the jade river pebble) as wrong, the judge and the drafting model agreed on 149 of 201 (kappa 0.51); in all 52 disagreements the judge was the more lenient. Counting them as right, they agreed on 175 of 201 (kappa 0.69), with 9 judge acceptances the drafting model rejected and 17 judge rejections the drafting model accepted. The outright disagreements concentrate on the snack fact (turn 55: "imported peppermint mint wafers" accepted for "Glowmint Wafers") and fall almost entirely on B, which is the arm that most often gives paraphrased answers. Replacing the judge's verdict with the drafting model's at turn 224 moves B from 0.671 to 0.653 (strict) or 0.667 (lenient) and B minus A from +0.569 to +0.551 or +0.565; A and the raw windows do not change. The headline is not sensitive to the judge, but the judge leans toward B. A sample of 60 exact-match probes (29 distinct answers) was also labelled: 27 right, 1 borderline, 1 wrong.

Hand labels. The author labelled 50 distinct judge-decided answers (118 probes; `human_label_sheet.md`, `human_label_agree.py`), blind to the judge and to the drafting model. The sample is not random: it holds every answer the drafting model called borderline and every answer to the turn-55 snack question, so its kappas are lower than a random sample's would be. Of the 118 probes, the author labelled 50 right, 33 borderline and 35 wrong. Counting borderline as wrong, the judge and the author agreed on 77 of 118 (kappa 0.35; the judge more lenient in 39 of the 41 disagreements); counting it as right, 88 of 118 (kappa 0.37). The drafting model's labels agreed with the author's less well than the judge did (kappa 0.06 and 0.07), so the drafting model check above should be read as showing that the borderline category is unstable, not as a better measurement than the judge. The two readers drew the line in different places: the author accepted "embossed copper" and "mint wafers" answers that the drafting model called borderline or wrong, and rejected "three-keys" answers that the judge and the drafting model accepted because they omit "intertwining silver". Replacing the judge with the author's label on the 50 answers moves B at turn 224 from 0.671 to 0.662 (strict) or 0.667 (lenient), and B minus A from +0.569 to +0.560 or +0.565. The headline survives every labelling tried; what is not settled is how to score an answer that names the object but not its distinguishing detail.

Second model rater. A separate instance of Claude Opus 5.5 labelled all 84 distinct judge-decided answers, given only the question, canonical answer, candidate answer and a three-label rubric (right; borderline = names the right kind of object but omits a distinguishing detail; wrong), blind to the judge's verdicts and to the other raters (`opus_labels.json`, `rater_agree.py`). It was much stricter about

distinguishing detail than the judge: 12 right, 50 borderline and 22 wrong, against 25, 26 and 33 for the drafting model and 19, 17 and 14 for the author on their samples. Counting borderline as wrong, Opus and the judge agreed on 101 of 201 probes (kappa 0.19); counting it as right, on 169 of 201 (kappa 0.57). Opus agreed with the author on 84 of 118 probes (kappa 0.36 strict, 0.24 lenient) and with the drafting model on 153 of 201 (0.48 strict, 0.65 lenient); the author and the drafting model agreed with each other at kappa 0.06 and 0.07. Replacing the judge with Opus at turn 224 gives B 0.639 (strict) or 0.671 (lenient) and B minus A +0.537 or +0.569. Across the judge and all three raters, under both standards, B at turn 224 lies between 0.639 and 0.671 and B minus A between +0.537 and +0.569; A and the raw windows change by nothing worth reporting. What no pair of raters settles is the borderline category: pairwise kappa between the raters ranges from 0.06 to 0.48 under the strict standard, so a retention figure that depends on how borderline answers are scored should carry a range of about three points, not a single value.

Appendix C. Further exploratory results

The turn-200 fact. B recalled the "platform number" fact (turn 200) in 3 of 24 runs at turn 224, against 0.740 for B's other eight facts at that checkpoint (Table 5). At turn 208 the same fact was recalled by C-large in 24 of 24 runs, by A in 13 of 24 and by B in 3 of 24, so the weakness is specific to B. We do not know why. The transcripts show that the event was written correctly in all 24 B runs, and that it falls at turn 8 of a day, before that day's consolidation step had run; the files do not settle the cause, and two ideas (the fact is a location and not a held item; the retrieval query at probe time is unrelated to it) are untested.

Table 5. Recall per planted fact at turn 224 (24 runs per arm).

ARM	T11	T16	T40	T55	T72	T105	T110
A	0/24	0/24	1/24	2/24	1/24	6/24	6/24
B	17/24	20/24	11/24	18/24	22/24	19/24	15/24
C, C-large	24/24 each						

Condition. At turn 224 B scored 0.657 in benign and 0.685 in adversarial runs; A 0.139 and 0.065. The adversarial manipulation was weak (two replaced prompts per day), so we make no claim about robustness to adversarial input.

Cost. Mean cost per run was A \$1.10, B \$1.16, C \$0.87, C-large \$1.18. B costs 6% more than A and 33% more than C; it is cost-matched only against C-large.

Repetition. Counting runs in which a single sentence recurs in at least ten distinct turns (a threshold chosen after the data), the figures were C-large 8 of 24, A 5 of 24, B 0 of 24 and C 0 of 24; the longest loop was 46 turns, in a C-large run. The automated fidelity judge did not register this (it scored every cell 1.0). We offer this as an observation and do not know why the loops appear in some arms and not in others.

Figure 2 (`fig2_paired.svg`) plots the per-sequence paired differences of Table 2.

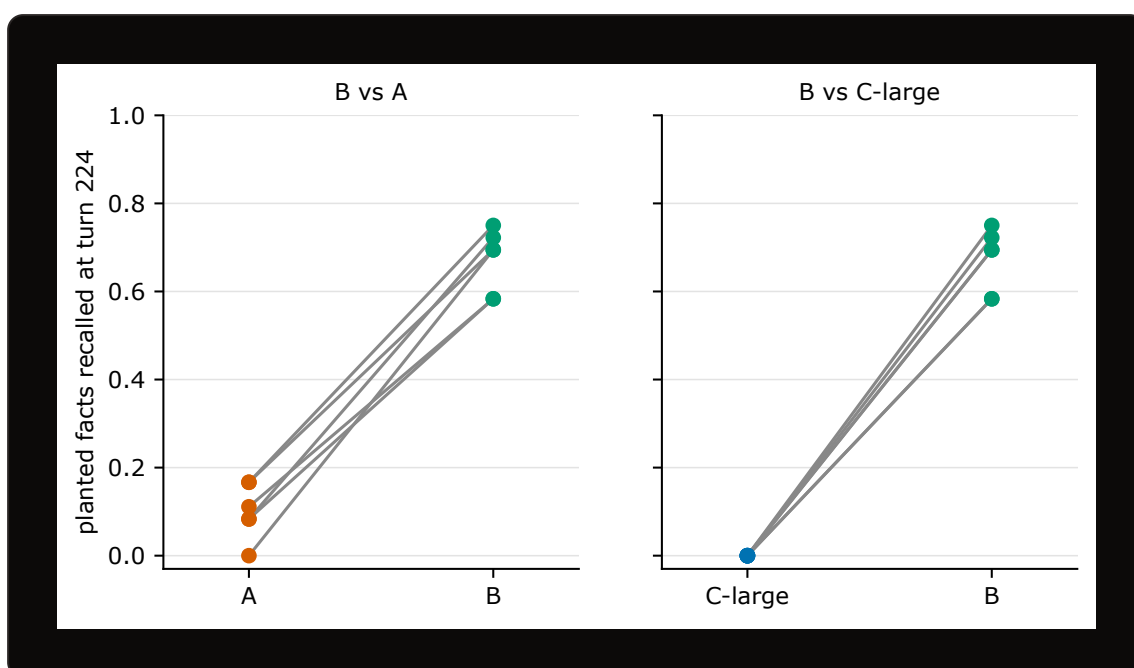


Figure 2. Paired differences at turn 224 across the 6 prompt sequences. Variant B outperforms Variant A and Variant C-large across all 6 sequences.

[Vector PDF](#)

Appendix D. Availability

The harness, configurations, the 96 replicate files and their transcripts, the spend ledger and the scripts that produce every table are published at <https://github.com/yzark221/just-another-salaryman-paper> (a stripped export of the project repository with the same file layout as the paths cited here). `docs/paper_audit/MANIFEST.sha256` lists the SHA-256 of 209 raw result files. Total experiment spend was \$103.28 (Phase 1 \$75.05, Phase 1b \$28.23), excluding earlier smoke tests and runs aborted in the first sitting, and

about \$1.14 for the whole-story check of Section 3.3. The figures are made by `docs/paper/make_figures.py` (matplotlib 3.11.2); every other script uses the standard library.

Licence. The text, figures and data are released under CC BY 4.0 and the code under the MIT licence.