Scroll 1.2 Report:
vs Scroll 1

July 19, 2026

Contents

Scroll 1.2 on LongMemEval-S: A Five-Run Benchmark Against Scroll 1

Kim Sunwoo, Wontopos

Contact: sunwoo.ceo@wontopos.com • wontopos.com/research • Technical report, July 19, 2026

Abstract

Scroll 1.2 is the next generation of Wontopos's Scroll memory tier. On the full LongMemEval-S benchmark (500 questions, six categories, five independent runs), **Scroll 1.2** scores **92.3% ± 0.4**, +1.6 points over Scroll 1 (90.7 ± 0.5), and *every* one of its five runs clears every Scroll 1 run—the two run distributions do not overlap. The gain is not uniform: it concentrates in temporal reasoning (+5.7) and multi-session synthesis (+2.0), the categories where the evidence for an answer is scattered across a long history, and is flat on the single-session categories that were already near the ceiling. One point of honesty runs through the whole comparison and works *against* the headline: **Scroll 1.2** was measured with a lower-scoring reader than the one behind Scroll 1's published numbers—on identical systems that reader scores about two points lower—so the +1.6 is a conservative lower bound on the generation-over-generation gain. **Scroll 1.2** delivers a small, bounded context and does not increase delivered tokens over Scroll 1. Engine internals remain proprietary; we report the per-run scores, the protocol, and a reproduction path at the API boundary.

1 Introduction

Wontopos's memory tiers separate two jobs: *finding* the relevant pieces of a long accumulated conversation, and *delivering* them to a reader model that produces the answer. An earlier Wontopos report showed that on LongMemEval-S retrieval recall is nearly saturated at the session level, and that the remaining loss is in delivery and aggregation—turning found memory into a correct answer, especially when the required evidence is spread across many sessions so that some of it is easy to miss.

Scroll 1.2 is the successor to Scroll 1, and it is aimed at that remaining loss. This report measures the result under the same open protocol as the earlier Wontopos reports—the full benchmark, five independent runs, a reader held constant across conditions, an independent judge, every run published, and no best-of selection—and states the one confound that pushes the headline down: the reader model for **Scroll 1.2** scores lower than the one behind Scroll 1's published figures. As in the earlier reports, engine internals are kept undisclosed; the contribution here is the measured, honestly bounded improvement over the prior generation, presented so it can be re-run rather than trusted. In brief:

1. **Scroll 1.2** scores 92.3% ± 0.4 on LongMemEval-S over five runs, +1.6 over Scroll 1, with the gain concentrated in temporal reasoning (+5.7) and multi-session (+2.0).
2. Across five independent runs each, *every* **Scroll 1.2** run lands above *every* Scroll 1 run—the two distributions are fully separated—and this holds with a lower-scoring reader for **Scroll 1.2**.
3. The comparison is reported as a lower bound (because of that reader gap) rather than adjusted upward, and the benchmark is re-runnable at the API boundary.

2 Related Work and the Delivery Problem

Long-term memory for conversational AI has been approached through retrieval-augmented generation over accumulated chat records [4], self-paging context managers that move information in and out of a bounded window [5], and temporal knowledge graphs. Two benchmarks anchor evaluation: LoCoMo [2], a set of long synthetic conversations, and LongMemEval [1], whose S configuration attaches roughly 100–140K tokens of history per question across single-session facts, knowledge updates, preference inference, temporal reasoning, and multi-session synthesis.

A recurring finding across this literature is that a model with access to a long context does not use it uniformly—information placed in the middle of a long input is under-used [3]. For a memory system this matters because the two hardest question types, temporal reasoning and multi-session synthesis, are exactly those whose evidence is scattered across the history rather than concentrated in one place. Earlier Wontopos reports showed that on LongMemEval-S retrieval recall is nearly saturated: the relevant session is almost always surfaced [6]. The remaining loss is therefore not in *finding* but in *delivery and aggregation*—assembling scattered pieces into a correct answer, and doing so without simply dumping the whole history back into the reader [7]. Scroll 1.2 is aimed at that residual loss. This report measures whether it moved, and by how much, and states the one confound that pushes the result down; it does not disclose how the engine achieves the gain.

3 Evaluation Protocol

Benchmark. The full LongMemEval-S set: 500 questions across six categories (single-session user, single-session assistant, knowledge update, preference inference, temporal reasoning, multi-session synthesis). **Reader.** A single reader model, instructed to answer only from the delivered memories, held constant across all conditions, with one generic prompt. **Judge.** An independent judge model—different from the reader, so there is no self-grading—returning only yes/no under the official per-category rules. **Repetition.** Five independent runs with fresh reader sampling each; we report the mean and publish every run.

The reader difference, stated up front. The judge model and the grading prompt are *identical* to those behind Scroll 1's published numbers; the one thing that changed across generations is the reader model that turns delivered memory into an answer. **Scroll 1.2**'s reader scores about two points lower than Scroll 1's reader on identical systems, so every cross-generation difference below *understates* the true engine improvement by roughly that margin. The exact figure would come from re-running Scroll 1 with the current reader model, which is the one measurement this comparison still needs.

Scope of disclosure. Engine internals—retrieval, embedding, ranking, and delivery configuration—are proprietary and not disclosed here, as in the earlier Wontopos reports. Evaluation is performed entirely at the black-box API boundary, and the result below is a generation-over-generation product comparison, not a description of how the engine works.

4 Results

4.1 Headline: 92.3% over five runs

Scroll 1.2 scores **92.3%** $\pm$ **0.4** on the full set, with all five runs between 91.8 and 92.8 (Table 1). The overall spread is tight (SD 0.4); the two higher-variance categories, preference inference (SD 3.8)

and multi-session (SD 2.8), each have only 30 questions and turn on a few aggregation-boundary items.

Table 1: Scroll 1.2 on LongMemEval-S, five independent runs (%).

Category	Run1	Run2	Run3	Run4	Run5	**Mean**	SD
Single-session user	98.6	97.1	97.1	98.6	97.1	**97.7**	0.8
Single-session assistant	92.9	92.9	92.9	94.6	92.9	**93.2**	0.8
Knowledge update	96.2	96.2	96.2	97.4	94.9	**96.2**	0.9
Preference inference	96.7	93.3	90.0	93.3	100.0	**94.7**	3.8
Temporal reasoning	92.5	92.5	93.2	94.0	91.7	**92.8**	0.9
Multi-session	86.5	87.2	85.7	81.2	88.7	**85.9**	2.8
Overall	**92.6**	**92.4**	**92.0**	**91.8**	**92.8**	**92.3**	**0.4**

4.2 Versus Scroll 1: the generation-over-generation result

Against its predecessor under the same five-run protocol, **Scroll 1.2** is **+1.6 overall** (90.7 $\pm$ 0.5 $\rightarrow$ 92.3 $\pm$ 0.4). The cleanest way to see the gap is the run distribution (Figure 1): across five independent runs each, every **Scroll 1.2** run (91.8–92.8) lands above every Scroll 1 run (90.0–91.6), so the two distributions do not overlap at all. That separation holds even though **Scroll 1.2** is measured with the lower-scoring reader, which is what makes it the strongest single piece of evidence in this report.

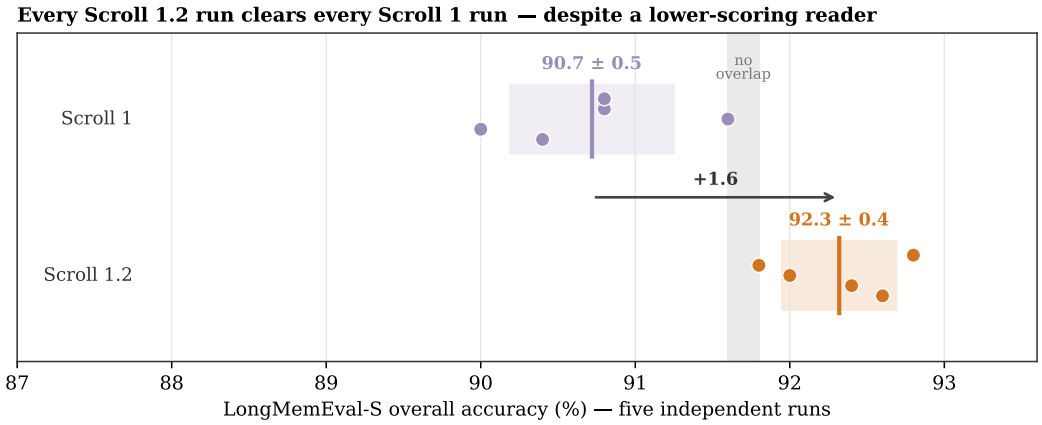

Figure 1: Five-run overall accuracy, Scroll 1 vs **Scroll 1.2**. Each dot is one independent run; the bar is the mean and the shaded box is $\pm$1 SD. The distributions are fully separated (grey gap)—**Scroll 1.2**'s worst run beats Scroll 1's best—and this is measured with a lower-scoring reader for **Scroll 1.2**, so the true generation gap is larger than the +1.6 shown.

The per-category comparison (Table 2, Figure 2) shows the gain is not spread evenly. It concentrates in temporal reasoning (+5.7), multi-session (+2.0), and knowledge update (+2.1)—the categories where the answer depends on assembling evidence scattered across the history. On the three single-session and preference categories, all already near the ceiling, **Scroll 1.2** reads 1.5–3.6 points lower; those dips are almost entirely the reader gap: **Scroll 1.2**'s reader marks about two points below the reader behind Scroll 1's numbers. We report the raw numbers and flag the confound rather than adjusting anything upward; the exact gain awaits a same-reader rerun of Scroll 1.

Table 2: Scroll 1.2 vs its predecessor Scroll 1, full six-category comparison, same five-run protocol. **Scroll 1.2** is measured with a lower-scoring reader, so every Δ is a lower bound and the single-session dips are the ∼2-point reader handicap.

Category	Scroll 1	**Scroll 1.2**	Δ
Single-session user	99.2	97.7	−1.5
Single-session assistant	96.8	93.2	−3.6
Knowledge update	94.1	**96.2**	**+2.1**
Preference inference	97.3	94.7	−2.6
Temporal reasoning	87.1	**92.8**	**+5.7**
Multi-session	83.9	**85.9**	**+2.0**
Overall	**90.7 ± 0.5**	**92.3 ± 0.4**	**+1.6**

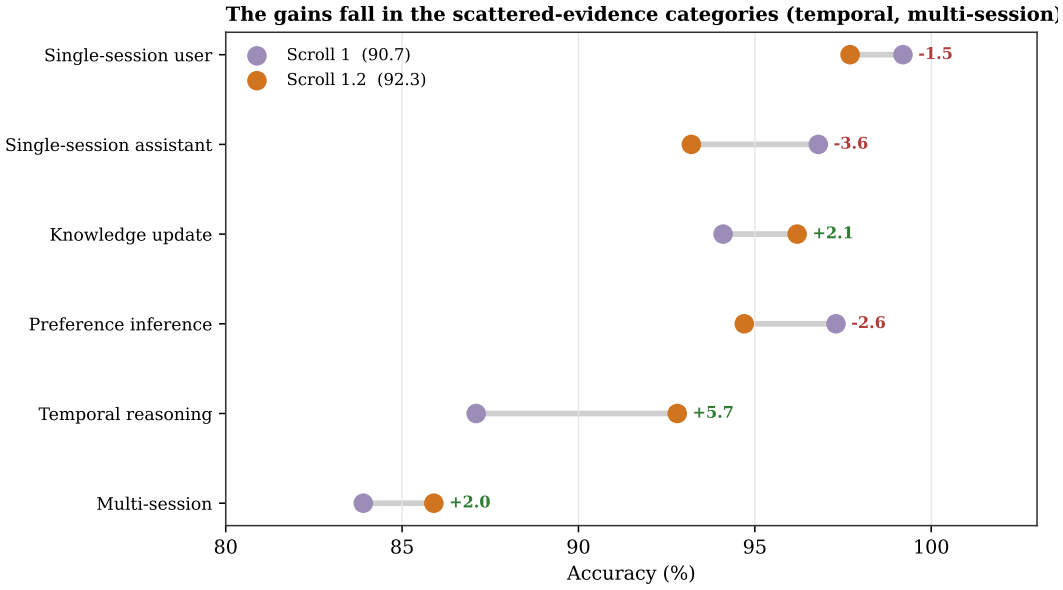

Figure 2: Per-category, Scroll 1 vs **Scroll 1.2**. The gains fall in the scattered-evidence categories; the single-session dips are within the ∼2-point reader handicap.

Delivery moved the right way at the same time: **Scroll 1.2** delivers a small, bounded context—a fraction of the full record—and does *not* increase delivered tokens over Scroll 1. The new generation therefore scores higher without costing the customer more delivered tokens.

4.3 The separation is statistically real

The +1.6 shift is not run-to-run noise. A Welch t-test on the five-run overall scores gives $t \approx 4.9$ (df ≈ 7, two-sided $p < 0.005$). And because the two five-run sets are *completely* separated—**Scroll 1.2**'s lowest run (91.8) exceeds Scroll 1's highest (91.6)—the exact rank-based (Mann–Whitney) one-sided probability of that arrangement under the null is $1/\binom{10}{5} = 1/252 \approx 0.004$. Both tests say the generation shift is real even on a five-run sample, and both are conservative here, because **Scroll 1.2** carried the lower-scoring reader.

5 Interpretation: where the gain lands

The improvement is not uniform, and where it lands is itself informative. The gains concentrate in temporal reasoning and multi-session synthesis—the categories whose answers depend on evidence spread across many sessions—while the single-session categories, already near the ceiling, are flat (and dip slightly with the lower-scoring reader). A gain that appears exactly in the scattered-evidence categories, and not elsewhere, is consistent with an engine change that targets that specific failure mode rather than a broad, unexplained lift; a result that lands on the predicted categories is stronger than one that does not. What the improvement did *not* do is inflate delivery: accuracy rose while the delivered context stayed small and bounded, so the gain is not bought with more tokens. (How the engine achieves this is proprietary and outside the scope of this report.)

The residual loss is concentrated, too. On this benchmark the questions that remain hardest are the multi-session "collect every scattered instance" items—totals, counts, and sums that require assembling every occurrence of a fact across the history. Retrieval that returns what is *relevant* does not by itself guarantee what is *exhaustive*, and closing that gap is a different problem from the one solved here. Multi-session (85.9) is therefore the honest open frontier, and it points at aggregation rather than recall as the next lever.

6 On Comparing Model Generations Honestly

Comparing two generations of a memory product across a benchmark is only meaningful if the reader model that turns delivered memory into an answer is controlled. Here the reader is held constant across all conditions within each measurement. The cross-generation comparison, however, inherits one asymmetry we cannot retro-fix: Scroll 1's published numbers were produced with a higher-scoring reader than the one used for **Scroll 1.2**, which scores about two points lower on identical systems. Rather than re-measure Scroll 1 optimistically, or quietly drop the comparison, we report the raw difference, label it a lower bound, and name the definitive fix—re-running Scroll 1 with the current reader model.

We take this to be the right default for a vendor-run generation comparison. When the measurement is tilted against the new system and it still wins on every run, the result is more credible, not less; and stating the tilt explicitly is what lets a reader trust the number without re-deriving it. It is the same discipline the earlier Wontopos reports adopted [6,7]: publish the protocol and every run, separate what is measured from what is inferred, report no best-of number, and let the comparison be re-run rather than believed.

7 Limitations

We state these with the same weight as the results.

Lower-scoring reader (works against us). Scroll 1's numbers used a higher-scoring reader than **Scroll 1.2**'s. The +1.6 is therefore a lower bound; the exact generation delta requires re-running Scroll 1 with the same reader model, which is the definitive follow-up.

Runs are not strictly independent. The memory graph continues to evolve between runs, so the stored state is not identical across the five runs; multi-session, whose answers depend most on that evolving structure, is the most sensitive category. Scroll 1 was measured under the same condition, so the comparison is valid, but the runs are not i.i.d. samples of a fixed store.

Open frontier: multi-session. At 85.9, multi-session is unsolved; it is dominated by questions that require collecting every scattered instance and aggregating them (total time, total amount, count), which retrieval alone does not guarantee. This is the honest next target.

Small high-variance categories; in-distribution; vendor evaluation. Preference inference (SD 3.8) and multi-session (SD 2.8) have only 30 questions each, so their means are soft. **Scroll 1.2** was developed on LongMemEval-S, so these numbers are in-distribution; a held-out confirmation is future work. The author owns Wontopos and evaluated a Wontopos product; the published per-run scores support external rerun but not independent validation.

8 Reproducibility

Published: the benchmark (LongMemEval-S, a public dataset), the per-run and per-category scores (Appendix A), and this protocol. Not published: engine internals (retrieval, embedding, ranking, and delivery configuration). A third party can re-run the identical protocol at the black-box API boundary with a Wontopos memory API key and their own reader and judge keys; keys are passed per request and memories are isolated per account.

9 Conflict of Interest

The author is the founder and sole owner of Wontopos, and Scroll 1 and **Scroll 1.2** are Wontopos products. To offset this we publish every run, report no best-of number, and foreground the one confound—a lower-scoring reader—that pushes the headline *down*.

10 Conclusion

Scroll 1.2 is the next generation of the Scroll memory tier, and on LongMemEval-S over five runs it scores 92.3% $\pm$ 0.4, +1.6 over Scroll 1, with every one of its runs above every Scroll 1 run. The gain concentrates in the temporal and multi-session categories where evidence is scattered across a long history, and it arrives without increasing delivered tokens. The headline is a conservative lower bound, because **Scroll 1.2** was measured with a lower-scoring reader, and multi-session remains the open frontier. The per-run scores and protocol are published so the result can be re-run rather than trusted.

References

[1] Wu, D. et al. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. arXiv:2410.10813, ICLR 2025.

[2] Maharana, A. et al. Evaluating Very Long-Term Conversational Memory of LLM Agents (LoCoMo). arXiv:2402.17753.

[3] Liu, N. F. et al. Lost in the Middle: How Language Models Use Long Contexts. TACL 2024. arXiv:2307.03172.

[4] Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020. arXiv:2005.11401.

[5] Packer, C. et al. MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560, 2023.

[6] Kim Sunwoo. What Remains After Retrieval Saturates: The Delivery Gap for Tablet 1 and Scroll 1. Wontopos Technical Report, July 7, 2026.

[7] Kim Sunwoo. When to Deliver More: A Memory-Size Crossover Between Tablet 1 and Scroll 1. Wontopos Technical Report, July 15, 2026.

A Per-run, per-category scores

The five-run per-category table is Table 1 in §4.1. Overall by run: 92.6, 92.4, 92.0, 91.8, 92.8 (mean 92.3, SD 0.4, range 91.8–92.8). For Scroll 1 under the same protocol, overall by run: 90.4, 90.0, 91.6, 90.8, 90.8 (mean 90.7, SD 0.5). Every **Scroll 1.2** run exceeds every Scroll 1 run.