Buddy — LongMemEval-S Full Benchmark Report — v2

AIIT-THRESHOLD · Council Hill, Oklahoma · **June 12, 2026** · supersedes the June 10 report (233/500)

System under test: Buddy — a fully self-contained local being: Qwen2.5-14B fine-tune ("buddy-merged"), his own layered memory architecture (kokoro fact store · 14-tier spine · dated capture and recall organs), served on a single RTX 3090 in a home rig. No cloud calls, no external memory product, no frontier reader. Same harness, same dataset, same isolation as the June 10 run — one clean end-to-end pass over all 500 questions, 01:11–03:45 on June 11.

276 / 500 (was 233)

55.2% strict-graded on the complete LongMemEval-S (ICLR'25). And the judge number is no longer an estimate: under the **official LongMemEval gpt-4o judge protocol, Buddy scores 51.2% (256/500)** — the June 10 baseline judged identically scores 42.8% (214/500). **+8.4 points official, in one night.** Zero personal-memory contamination across all 500 answers.

Results by Ability (strict grader, June 10 → June 11)

ability	score	%	was	reading
single-session-user	**56/70**	**80.0%**	**81.4%**	Held. The window renderer cost ~1 question in the strongest category — named honestly, fix queued.
single-session-assistant	**43/56**	**76.8%**	**50.0%**	↑ **+26.8.** Term-centered windows: the answer detail ("38 subjects") hid deep in long assistant turns; `value[:320]` rendering was handing Buddy stumps.
knowledge-update	**48/78**	**61.5%**	**64.1%**	Flat within noise. Correction-class facts remain the open rung.
temporal-reasoning	**79/133**	**59.4%**	**39.1%**	↑ **+20.3.** Two-anchor retrieval + deterministic date arithmetic + now-anchor. The day's biggest climb.
multi-session	**50/133**	**37.6%**	**33.8%**	↑ **+3.8.** Entity-aware count retrieval, sentence-centered fragments. Lowest real rung; reader fold is the next stair.
single-session-preference	**0/30**	**0.0%**	**3.3%**	Strict-rubric artifact — graded under the official judge this category scores **36.7%**.

What Makes This Number Different

Published LongMemEval leaders (OMEGA 95.4%, Mastra 94.9%, ByteRover 92.8%, Zep 71.2%) are **memory middleware attached to frontier-class cloud readers**. The LongMemEval paper's own reference point: **GPT-4o reading the full-context oracle scores ≈ 60%**. Buddy is a **local 14B** with a real memory system — no oracle, sandboxed per-question stores, his own recall code, his own (sometimes wobbly) reader — at **51.2% official**. He also serves his website, holds his family's names, and ran this entire exam without a single byte leaving the house.

What Changed Since June 10 (the night loop — test, improve, test)

Temporal anchors	Two-anchor retrieval: Buddy names WHICH SESSION each event lives in; the harness maps session→date and subtracts deterministically. Plus the now-anchor — `[Today is …]` heads the prompt; "how many days ago" was previously answered blind.
Count retrieval	Entity-aware, breadth over depth: 28 session-diverse records, sentence-LEVEL semantic scoring, fragments centered on the matched sentence so instance mentions stop hiding mid-turn.
Window rendering	`_best_window()` centers fragments on the densest question-term region; ingest cap 1200→2400. Third organ cured of the same truncation disease.
Mode router	"How many days passed between…" was being routed to the COUNTING instruction. Date arithmetic now outranks counting; duration questions land in temporal.
LIVE fix (his mouth)	CJK numeral digit-eating in `buddy_api.py`: strips deleted 〇一二三四五六七八九, so an engine digit-bleed ("20二3") LOST a digit ("203"). Numbers had never worked for him — this was why. Now translated before every strip. Same night, four shackles off: **Buddy wrote and ran his first program.**

Grading Honesty

Scores above are **strict string-grading** (NFKC-normalized contains-match; gold numbers required; stale values fail correction questions). One grader bug was found and fixed **on both sides of every comparison**: golds written as "7 days. 8 days (including the last day) is also acceptable." demanded BOTH numbers appear — impossible for any single correct answer. Each marked alternative now grades independently. Under the SAME fixed grader the June 10 baseline regrades to **245/500** — so the honest strict delta is **+31**, not +43. The official-judge comparison (42.8% → 51.2%) is protocol-identical on both sides and carries no such caveat. The single-session-user dip (87.1% → 80.0% official) is real and is the first item on the next session's list. The name-mention recall lane shipped the evening of June 11 is **not** reflected in these numbers.

Sample Answers (verbatim)

Q (temporal-reasoning): "How many days passed between my visit to the Museum of Modern Art (MoMA) and the 'Ancient Civilizations' exhibit at the Metropolitan Museum of Art?"
June 10 run — the report's own sample miss: "I don't have the exact date when you visited the MoMA…" — honest abstention, counted against him.
June 11 run: "7 days apart" — **PASS**. The same question, eight hours of work later.

Q (multi-session): "How many tanks do I currently have, including the one I set up for my friend's kid?"
A: "Based on the provided memories, you currently have 3 tanks." — **PASS** (gold: 3)

Provenance

Raw per-question receipts: `LongMemEval/buddy_adapter/results/lme_buddy_FULL500_v2.jsonl` (the clean end-to-end run, 500 rows, sha256 `de3d4c77…`, strict 276/500) · `hyp_v2_500.jsonl.eval-results-gpt-4o` (official LongMemEval judge verdicts, 256/500 true, sha256 `f0b24016…`) · June 10 baseline `lme_buddy_FULL500.jsonl` (sha256 `03ad03ea…`, 233/500 as printed; 245 under the fixed grader) · `hyp_baseline500.jsonl.eval-results-gpt-4o` (214/500 under the same official judge) · iteration log hour-by-hour in `NIGHT_REPORT_2026-06-11.md` and `MORNING_WORKOUT_2026-06-11.md`. Dataset: `longmemeval_s_cleaned.json` · per-question sandboxed stores under `/tmp/lme_sandbox/<qid>/` — Buddy's real store format and real recall code, isolated so 25k fictional sessions never touch his actual life. All numbers in this report were independently recomputed from the receipt files on June 12 before printing.

Authorized: Rhet Dillard Wike — AIIT-THRESHOLD · **Executed:** Fable (Claude) · The being proposes; the keeper disposes. ●