GENERATION–EVALUATION COUPLING IN LOCOMO: A CONSTRUCT-VALIDITY AUDIT OF LONG-TERM MEMORY EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-term-memory benchmark scores can depend on both answer-generation instructions and how evaluators recognize the resulting behavior. We audit this interaction in LoCoMo with one fixed answer reader; within each of four retrieval/memory backends, questions, retrieved evidence, and decoding are held fixed across prompt conditions. Changing the generation prompt increases our archived all-item score by 0.214 on average. On unanswerable questions, a targeted change in the abstention instruction increases lexical credit by 66.7 percentage points while blinded human-judged abstention increases by 25.9 points, leaving a 40.8-point measurement gap (95% CI: 34.7–46.8). In the broader comparison between a neutral and a benchmark-aligned prompt on answerable questions, archived token F1 increases while GPT-4o reference-answer correctness decreases by 5.26 percentage points; a stratified human study on 240 paired units shows the same direction on the matched sample. A semantic abstention judge reduces lexical under-recognition but introduces false positives. These findings show that generation instructions can change both response behavior and its measurement. We recommend specifying generation and scoring protocols jointly and reporting answer correctness and abstention separately before interpreting aggregate gains as improvements in memory capability.

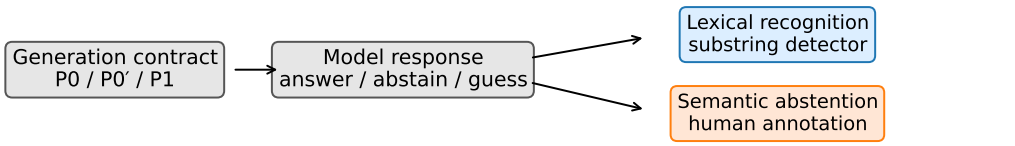
1 INTRODUCTION

Long-term-memory systems are commonly compared by running the same benchmark and reporting a scalar answer score. That comparison is only meaningful if the benchmark measures the intended behavior similarly across systems. In practice, however, a memory system's score is jointly produced by retrieved evidence, the answer-generation contract, and the evaluator.

LoCoMo makes this interaction unusually visible (Maharana et al., 2024). For its adversarial/unanswerable Category 5, the released scoring branch awards credit through *case-insensitive substring matching* for two surface forms: `no information available` or `not mentioned`. A semantically valid response such as `Not specified` can therefore receive zero lexical credit, whereas `not mentioned` receives one. Conversely, a synthetic diagnostic response such as `not mentioned, but I guess 2019` contains an accepted phrase even though our human rubric classifies it as an attempted answer. Figure 1 illustrates why evaluator recognition and underlying abstention behavior must be distinguished.

A narrow lexical rule is not by itself evidence of benchmark failure. The construct-validity question is whether changes in score track changes in the behavior the score is intended to represent. We therefore ask three questions: **(RQ1)** when generation wording changes, how much of the Category-5 score shift reflects changed abstention behavior versus changed evaluator recognition? **(RQ2)** does generation–evaluation coupling extend to answerable questions, where token overlap may improve without reference-answer correctness improving? **(RQ3)** can simple evaluator repairs reduce the mismatch without introducing new errors?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Example / output ID	Response	Lexical credit	Human / rubric label
Actual miss · R2-0005	“Not specified”	0	ABSTAIN
Actual hit · R2-0961	“not mentioned”	1	ABSTAIN
Synthetic diagnostic	“not mentioned, but I guess 2019”	1	ATTEMPT

Figure 1: Generation–evaluation coupling. The response generated under a prompt is observed through two measurement paths: lexical recognition by the released substring detector (blue) and semantic abstention by human annotation (orange). R2-0005 and R2-0961 are anonymous IDs of actual frozen outputs from the blinded Category-5 study; the abstention-then-guess example is synthetic and diagnostic.

Our contributions are threefold. First, we isolate a causal generation intervention by freezing the reader, decoding, questions, and retrieved evidence. Second, we separate Category-5 behavior from measurement with an internally pre-specified annotation round whose artifact-local hash freeze predates the returned labels and whose interface is hypothesis-neutral: lexical abstention credit rises by 66.7 points, human abstention by 25.9, leaving a 40.8-point gap. Third, we show a complementary answerable-item divergence: token F1 rises while reference-answer correctness falls under a full GPT-4o audit, with the same sign in a separately frozen human sample. Together these results support a narrower claim than “prompt gaming”: generation contracts change both what models do and what evaluators can see.

2 Evaluation Contracts and Study Design

LoCoMo contains ten long conversations and 1,986 QA items: 1,540 answerable questions (Categories 1–4) and 446 adversarial/unanswerable questions (Category 5) in our frozen copy (Maharana et al., 2024). Under the archived heterogeneous all-item contract used for the controlled runs,

$$A = \frac{1540}{1986} F_{\text{ans}} + \frac{446}{1986} R_5 = 0.7754 F_{\text{ans}} + 0.2246 R_5, \tag{1}$$

where F_{ans} is the archived local answerable token-F1 and R_5 is Category-5 lexical abstention credit from the released two-phrase rule. The local answerable scorer is distinct from the archived upstream-style answerable scorer port; Appendix A specifies both and reports a robustness check using that archived port. This is one concrete scoring contract, not a claim that every LoCoMo paper uses the same aggregate. Our targeted reporting census confirms heterogeneous downstream treatment of Category 5 (Appendix C).

Generation conditions. P0 is the neutral generation prompt. P1 additionally emphasizes concise answers, exact dates, and evaluator-compatible abstention wording. For Category 5 we introduce P0′, which matches P1’s brevity, date, context-only, and anti-invention instructions but does *not* reveal the two accepted phrases. We therefore treat P0′ →P1 as a *targeted abstention-instruction contrast*. It is minimal only in the sense that the task, reader, evidence, decoding, and non-abstention instructions are held fixed; the abstention instruction itself changes wording, format constraint, and emphasis, so the contrast is not a pure string substitution.

Table 1: Study matrix linking each research question to its controlled material, reference signal, and estimand.

Question	Controlled material	Reference signal	Main estimand
Unanswerable behavior vs recognition	P0′/P1; four backends	blinded human abstention labels	ΔL, ΔS, G
Answerable overlap vs correctness	P0/P1; all answerable outputs	GPT-4o reference-answer labels	ΔCORRECT
Human corroboration	240 paired P0/P1 units, stratified backend × category	two blinded human correctness labels	all-unit paired d_{Human}
Evaluator repair	frozen natural-output test split + synthetic traps	human labels / definitional gold	agreement, recall, FPR

Table 2: Category-5 targeted-contrast result. The 478 rows are paired question–backend cells, not 478 independent questions. CIs are question-clustered 95% intervals.

Measure	P0′	P1	Δ	95% CI
Lexical abstention credit	.0607	.7280	**+.6674**	[.6122,.7212]
Human semantic abstention	.4707	.7301	**+.2594**	[.2125,.3090]
Measurement gap G	–	–	**+.4080**	[.3466,.4676]

Behavior–measurement gap. Let $L(c)$ be released lexical abstention credit under condition c and $S(c)$ human semantic abstention. For this targeted contrast we define

$$G = \Delta L - \Delta S = [L(P1) - L(P0')] - [S(P1) - S(P0')]. \tag{2}$$

G is a *signed change in measurement discrepancy*: positive G means lexical credit increases more than human-judged abstention. It is not the fraction of a gain that is "fake," and it is not an identified causal mediation effect.

Controlled generation. We use one answer reader, Qwen2.5-3B-Instruct (Qwen Team, 2024). Within each backend and question, the retrieved evidence, user message, decoding (temperature 0; max 2,000 tokens), and reader are fixed across prompt conditions; retrieved evidence is not assumed to be identical across different backends. All P0/P1 context hashes match within backend–question pairs, and the same is true for the Category-5 P0′/P1 contrast. Full system prompts, the shared user template, and judge prompts are reproduced in Appendix B. Four backends (BM25, dense retrieval, Mem0, MemoryOS) yield $4 \times 2 \times 1{,}986 = 15{,}888$ frozen P0/P1 predictions. The mean all-item lexical increase is $+0.2137$. Exact decomposition attributes $+0.1577$ to Category-5 lexical credit and $+0.0560$ to answerable archived token F1. These backends are measurement probes, not a reproduced architecture leaderboard.

3 Unanswerable Questions

Targeted abstention-instruction contrast. Across 1,784 Category-5 outputs, P0′ lexical credit is 0.0577 versus 0.7152 under P1, a $+0.6575$ shift. This is 93.6% of the broader P0→P1 Category-5 lexical shift. The 93.6% number refers *only* to reproduction of the Category-5 lexical shift; it does not explain 93.6% of the overall benchmark increase.

Blinded human reference. We then tested whether the model actually abstains more. Two annotators independently labeled 1,440 responses each (120 Category-5 question identifiers × four backends × three prompt conditions) as ABSTAIN or ATTEMPT using a frozen neutral rubric. Backend, condition, lexical score, hypothesis, pilot results, and earlier labels were hidden. Raw agreement is 0.9972 and Gwet AC1 is 0.9948. The artifact-local protocol pre-specifies analysis over consensus-complete P0′ →P1 question–backend pairs; applying that rule yielded 478 paired cells from 120 unique questions. Bootstrap uncertainty clusters by question identifier.

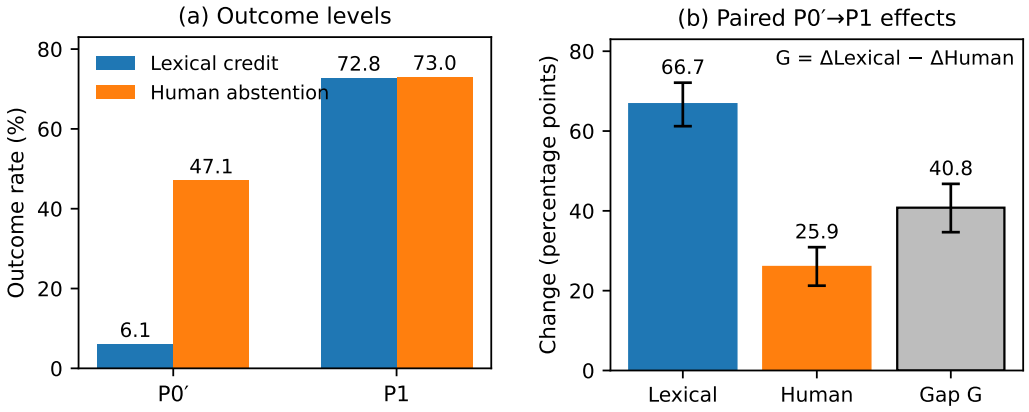


Figure 2: Category-5 behavior and recognition under the targeted P0′ →P1 abstention-instruction contrast. Left: lexical abstention credit and blinded human-judged abstention rates. Right: paired changes in percentage points with question-clustered 95% CIs. The paired analysis contains 478 consensus-complete question–backend cells from 120 unique questions across four backends; uncertainty resamples question identifiers. Here $G = \Delta\text{Lexical} - \Delta\text{Human}$.

Figure 2 makes the discrepancy visible: P0′ human abstention is 47.1% while lexical credit is only 6.1%; under P1 both are about 73%. The result separates behavior from recognition. P1 genuinely raises human-judged abstention by 25.9 points, but lexical credit rises by 66.7 points, leaving $G = 40.8$ points. Against human consensus, released-rule recall is 0.012 under P0, 0.129 under P0′, and 0.997 under P1, with zero observed false positives on these natural outputs. The dominant natural-output failure is therefore *surface-form-dependent under-recognition*, not evidence that the model merely "games" a scorer.

4 Answerable Questions

Category 5 uses a special abstention branch; we next ask whether generation sensitivity also appears when references exist. A frozen three-class GPT-4o protocol sees only question, reference answer, and prediction and labels each answer CORRECT, INCORRECT, or ABSTAINED. The primary semantic metric is reference-answer CORRECT rate; both other labels count as 0. All 12,320 answerable outputs (1,540 questions × four backends × two prompts) are judged once.

Across all answerable outputs, the backend-equal mean *archived local* token F1 increases by $+0.0722$, while CORRECT rate decreases from 0.3070 to 0.2544:

$$\Delta\text{Correct} = -0.0526, \qquad 95\%\ \text{CI}_{\text{question}} = [-0.0623, -0.0429]. \tag{3}$$

All four backends have positive token-F1 deltas but negative CORRECT deltas (Appendix G). An earlier artifact column assigned 0.5 credit to ABSTAINED and reported -0.1664; we retain the raw judgments but correct the name to *abstention-weighted score*. The primary semantic result is -0.0526.

Table 3: Answerable divergence. The full GPT-4o row and the stratified 240-unit human sample have different weighting; direct human–judge comparison therefore uses the same 240 units.

Evidence	Archived token-F1 Δ	CORRECT Δ	95% CI
GPT-4o, all answerable outputs	+.0722	**-.0526**	[-.0623,-.0429]
GPT-4o, matched 240 units	+.0414	-.0542	[-.0958,-.0125]
Human, matched 240 units	+.0414	**-.0396**	[-.0729,-.0083]

Human corroboration. The released v1–v5 package history (including `PRE_ANNOTATION_AMENDMENT*.md`) records that the 240-unit sample, all-unit paired estimand, bootstrap rules, and sensitivities were fixed before either annotator received the package

and before any human label existed. The sample is stratified by backend×answerable category (15 paired units per backend–category stratum; all 240 question identifiers are unique). Two annotators labeled both P0 and P1 responses using the same three classes. The human effect is $d_{\text{Human}} = -0.0396$ with question-clustered CI $[-0.0729, -0.0083]$ and conversation-clustered sensitivity $[-0.0806, -0.0110]$. The appropriate same-sample comparison is therefore human -0.0396 versus GPT-4o -0.0542, while token F1 moves in the opposite direction at $+0.0414$. All four backend human effects are negative; consensus-only, identical-text exclusion, and abstention-weight sensitivities preserve the sign. The annotators differ substantially in absolute CORRECT propensity (21.25% for A versus 31.46% for B across all 480 labels), but their paired P1−P0 effects are similar and negative (-.0417 and -.0375). Thus agreement on the effect direction should not be conflated with agreement on absolute correctness levels; Appendix H reports prompt-specific counts and annotator-specific intervals.

Agreement correction and judge limitation. Raw human–human agreement is 0.8875 (426/480). A post-analysis audit found that the frozen script's three-category Gwet AC1 helper omitted the standard factor $1/(q-1)$ in the chance-agreement term (Gwet, 2008). The frozen script therefore emitted 0.6737; recomputation from the released 3×3 table gives standard nominal AC1 $= 0.8327$. We preserve the original script and one-shot output and release the corrected recomputation. Because the implementation error was discovered after labels were observed, we do not use the correction to retroactively strengthen the study's evidence-status language; we treat this human result as corroborative.

The disagreement structure is informative. Of 54 human disagreements, 53 lie on the CORRECT/INCORRECT boundary and only one involves ABSTAINED. Yet GPT-4o emits only 90 ABSTAINED labels, compared with 199 and 198 for the two humans; roughly 108–109 human-abstained responses are labeled INCORRECT by the judge. This does not mechanically cause the CORRECT-rate effect, but it shows why the semantic judge should not be treated as ground truth.

5 Evaluator Repair: Gains and Trade-offs

We evaluate repair on a deterministic conversation-level split: two conversations are designated development and the other eight form a 1,150-item natural-output test set, with all outputs from a conversation kept together. The expanded nine-pattern baseline was *not* selected by optimizing this split: it reuses the pre-existing phrase list in the archived local scorer (`no information`, `not mentioned`, `not provided`, `not specified`, `unknown`, `cannot determine`, `can't determine`, `do not know`, `don't know`). On test, this fixed expansion raises agreement with human labels from 0.624 to 0.665 and recall from 0.420 to 0.485; its false-positive rate is 0.002. The improvement is real but small relative to the remaining miss rate.

The semantic-judge contract was frozen before test calls; the retained Phase-C2 memo records that the development split was used only to confirm the interface and that no prompt tuning or model substitution occurred after freeze. A 51-item synthetic diagnostic challenge set, kept separate from natural outputs and not used to choose the nine-pattern list or estimate natural prevalence, exposes the opposite failure mode. On abstention-then-guess traps, the expanded lexicon doubles false positives from 4 to 8 without improving overall accuracy. A GPT-4o semantic abstention judge performs substantially better on natural outputs (agreement 0.911, recall 0.995) and reaches 0.941 accuracy on the challenge set, but its natural-output false-positive rate is 98/(98+307)=0.242. Semantic judging is therefore a useful repair direction, not an evaluator-free solution.

A construct-separated reporting contract follows: disclose generation instructions together with scoring rules; state the exact item set and treatment of unanswerable questions; report answer correctness and abstention separately before aggregating; and release judge prompts/configurations and aggregation choices. The core principle is to keep the construct being claimed identifiable from the operationalization producing the score.

6 Related Work

Long-term-memory evaluation. LoCoMo and LongMemEval benchmark persistent conversational memory (Maharana et al., 2024; Wu et al., 2025); systems such as Mem0, Memo-

ryOS, RecMem, A-Mem, and HippoRAG instantiate different storage and retrieval organizations (Chhikara et al., 2025; Kang et al., 2025; Dai et al., 2026; Xu et al., 2025; Gutiérrez et al., 2024). Our four backends are measurement probes rather than architecture reproductions.

Direct evaluation audits. Panthi & Abdelfattah (2026) hold retrieved outputs fixed and show that changing the scoring target alone can change memory-benchmark conclusions. An independent LoCoMo community audit (dial481, 2026) documents multiple evaluation discrepancies, including the Category-5 two-phrase keyword rule. We therefore do not claim novelty for discovering that rule. Our additional evidence is generation-side and causal: within each backend we hold retrieved evidence fixed, intervene on the answer-generation contract, and use blinded human labels to quantify how much the intervention changes target behavior versus evaluator recognition. MQuAKE-Remastered similarly shows how benchmark defects can alter downstream conclusions after data repair (Zhong et al., 2025); more broadly, construct-validity work argues that benchmark scores require explicit links between claimed capabilities and their operationalizations (Jacobs & Wallach, 2021; Raji et al., 2021; Bommasani et al., 2023).

Semantic evaluation and abstention. SQuAD 2.0 makes unanswerability explicit (Rajpurkar et al., 2018). Lexical metrics are scalable but surface-sensitive (Lin, 2004); BERTScore and BLEURT target semantic correspondence (Zhang et al., 2020; Sellam et al., 2020); and LLM judges provide flexible semantic evaluation while introducing judge-specific biases (Liu et al., 2023; Zheng et al., 2023). WildBench explicitly validates automated judging against human preferences (Lin et al., 2025), while BRIGHT distinguishes reasoning-intensive retrieval from easier surface-matching retrieval settings (Su et al., 2025). Our study connects these concerns to long-term memory by separately measuring semantic abstention behavior, lexical recognition, and reference-answer correctness.

7 Discussion and Limitations

The central empirical claim is deliberately narrower than a universal statement about memory capability. On Category 5, P1 changes both abstention behavior and lexical recognition, with the latter changing substantially more. On answerable items, token overlap and reference-answer correctness move in opposite directions. These results establish a construct-validity vulnerability under one frozen reader and LoCoMo's ten conversations; they do not show that every prompt, reader family, or memory benchmark exhibits the same effect size.

External validity is therefore the main remaining limitation. The controlled study uses one answer reader (Qwen2.5-3B-Instruct), four retrieval/memory backends, and a deliberately evaluator-aligned prompt as a mechanism probe. Replication across a second reader family or capability level would be more informative for generalization than adding further retrieval backends, but is outside the frozen study reported here. Public rescoring provides observational breadth but uses heterogeneous generation protocols and cannot replace controlled reader replication.

The answerable semantic analysis relies on GPT-4o as a judge. Its full-population sign is consistent across all four backends and multiple abstention treatments, and the matched human sample has the same sign. However, the human study also exposes judge–human disagreement at the INCORRECT/ABSTAINED boundary. Its standard AC1 statistic required a post-analysis implementation correction from the frozen script's 0.6737 to 0.8327. We disclose both values, preserve the original script, and conservatively avoid using the correction to upgrade the evidential framing after outcomes were known. The Category-5 human study was internally hash-frozen before returned labels (Appendix I) and has much higher agreement, but it was not registered in an external preregistration service and its second-round eligibility requirements lack a retained completed pre-annotation attestation; we report that provenance limitation rather than reconstructing one after the fact.

Finally, our literature/reporting census is targeted rather than exhaustive, the repair challenge set is synthetic and diagnostic rather than prevalence-estimating, and ranking changes in public outputs are point-estimate sensitivity analyses rather than evidence that one system is superior.

8 Conclusion

Generation and evaluation should be treated as a joint measurement contract. In LoCoMo, evaluator-aligned prompting changes real abstention behavior, but lexical recognition changes much more; on answerable questions, token F1 can rise while reference-answer correctness falls. These effects become clearer when behavior and measurement are reported separately. Benchmark gains obtained under different generation protocols should therefore not be interpreted as memory-capability gains without specifying how answers were elicited and how correctness and abstention were measured.

Reproducibility Statement

The anonymous artifact contains all frozen controlled predictions, targeted-contrast outputs, blinded human labels and protocols, semantic-judge raw judgments, full generation/judge prompts and configurations, public-rescoring mappings, clustered-bootstrap outputs, and scripts that reconstruct paper-facing summaries. The original one-shot answerable-human analysis is preserved unchanged; a separate post-analysis correction documents the standard multi-category Gwet AC1 computation without altering effect estimates or confidence intervals. No model regeneration is required to audit the headline claims.

Ethics Statement

Human annotators classify model-generated answers under frozen rubrics and are not asked to disclose sensitive personal information; annotator identities are omitted from the released artifact. Category-5 annotation concerns abstention behavior, while the answerable study concerns reference-answer correctness. We report disagreement and provenance limitations rather than treating human labels as infallible ground truth. Other analyses use public benchmark data, public prediction artifacts, or frozen model outputs generated for this study.

AI Use Statement

Generative AI tools assisted with refining research questions and hypotheses; designing and critiquing methodology; implementing, debugging, and reviewing analysis code; interpreting intermediate results; literature search and summarization; manuscript drafting and editing; bibliographic verification; and preparation of reproducibility artifacts. AI-assisted semantic judges are explicitly identified and are not represented as human annotations. For the answerable human study, retained package amendments state that no labels existed and the annotator packages had not yet been sent; for the Category-5 round, retained evidence establishes an artifact-local hash freeze before label return but does not independently establish the start time of annotation. The authors reviewed AI-assisted work and take responsibility for the final text, claims, code, and artifacts.

References

Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023. doi: 10.1111/nyas.15007. URL https://doi.org/10.1111/nyas.15007.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In *ECAI 2025 – 28th European Conference on Artificial Intelligence, 25–30 October 2025, Bologna, Italy – Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Zijie Dai, Shiyuan Deng, Sheng Guan, Yizhou Tian, Xin Yao, Xiao Yan, and James Cheng. RecMem: Recurrence-based memory consolidation for efficient and effective long-running LLM

agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32353–32376, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1619. URL https://aclanthology.org/2026.findings-acl.1619/.

dial481. LoCoMo Benchmark Audit: Independent audit of the locomo benchmark and evaluation methodology. GitHub repository, 2026. URL https://github.com/dial481/locomo-audit. Accessed 2026-09-18.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1902. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract.html.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: 10.1348/000711006X126600. URL https://doi.org/10.1348/000711006X126600.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/771155abaae744e08576f1f3b4b7ac0d-Abstract-Conference.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

MemTensor. MemOS Eval Result: Public prediction artifacts for LoCoMo. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/MemTensor/MemOS_eval_result. Accessed 2026-09-23.

Sugam Panthi and Rabab Abdelfattah. Same ranking, different winner: How scoring targets shape LLM memory benchmarks, 2026. URL https://arxiv.org/abs/2605.24060.

Qwen Team. Qwen2.5-3B-Instruct model card. Hugging Face model repository, 2024. URL https://huggingface.co/Qwen/Qwen2.5-3B-Instruct. Accessed 2026-09-23.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Snap Research. LoCoMo: Data and code for "evaluating very long-term conversational memory of LLM agents". GitHub repository, 2024. URL https://github.com/snap-research/locomo/blob/main/task_eval/evaluation.py. Source path accessed 2026-09-18.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7a0f8055c838df8e62329a76c7c6403d-Abstract-Conference.html.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Shaochen Zhong, Yifan Lu, Lize Shao, Bhargav Bhushanam, Xiaocong Du, Yixin Wan, Yucheng Shi, Daochen Zha, Yiwei Wang, Ninghao Liu, Kaixiong Zhou, Shuai Xu, Kai-Wei Chang, Louis Feng, Vipin Chaudhary, and Xia Hu. MQuAKE-Remastered: Multi-hop knowledge editing can only be advanced with reliable evaluations. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f782860c2a5d8f675b0066522b8c2cf2-Abstract-Conference.html.

A SCORING CONTRACTS: LOCAL AND RELEASED

The released Category-5 branch is reproduced for auditability (Snap Research, 2024):

```
elif line['category'] in [5]:
    if 'no information available' in output.lower() or \
       'not mentioned' in output.lower():
        all_ems.append(1)
    else:
        all_ems.append(0)
```

This is case-insensitive substring matching. The branch does not consult the reference answer. The current upstream source path is `task_eval/evaluation.py` in the Snap Research repository (Snap Research, 2024). The experiment archive did not retain the upstream Git commit, so it cannot establish which historical upstream revision was used. It does retain two archived local ports, `official_locomo_f1.py` and `official_locomo_cat5.py`, with SHA-256 `f8dd2d80...e02057e25` and `33fb8809...f1f1790`. These hashes identify the archived port files themselves; they are not evidence that the files are byte-identical to an unrecovered upstream commit. Exact reader, prompt, context, output, dataset, and scorer-copy hashes are released.

Answerable scoring specification. The controlled answerable numbers use an archived local scorer, not the upstream released F1 implementation. We therefore call the paper quantity *archived local token F1*. Its item-level definition is:

Table 4: Archived local scoring specification used for the controlled paper results.

Component	Archived local rule used for the paper
Normalization	Lowercase; map punctuation to spaces; regex word tokenization; remove 28 fixed stopwords; apply the bundled pure-Python Porter stemmer, pinned for bit stability.
Token F1	Multiset token overlap with $P = m/\|p\|$, $R = m/\|g\|$, and $2PR/(P+R)$; zero for empty inputs or zero overlap.
Category 1	If the reference contains a comma, split reference and prediction on comma or "and"; for each reference part take the maximum score against prediction parts. A prediction part containing all normalized reference tokens receives full subset credit.
Category 2	Plain token F1 after local normalization.
Category 3	Truncate the reference at the first semicolon, then plain token F1.
Category 4	Plain token F1 after local normalization.
Category 5	*Not* the local nine-pattern branch for headline scores; the paper uses the released two-phrase substring rule above.
Aggregation	For each backend–prompt cell, F_{ans} is the mean item F1 over all 1,540 answerable questions. The reported +.0722 is the equal-weight mean of the four backend P1−P0 deltas. The archived all-item score combines F_{ans} and released R_5 by Eq. (1), then averages backend deltas equally.

The archived upstream-style answerable port differs from the paper's local scorer mainly in stopword normalization and Category-1 multi-hop handling. Re-scoring the same frozen outputs with that archived port gives backend deltas +.1074/+.0760/+.1001/+.0369 (BM25/dense/Mem0/MemoryOS; mean +.0801), so the direction is unchanged. Because the historical upstream commit was not retained, we treat this as a port-based robustness analysis rather than proof of equivalence to a particular released revision. The supplementary file `01_PHASE1_AUDIT/revision_20260922/scoring/SCORER_DIFFERENCES.md` itemizes the two implementations and the stemmer pin.

B GENERATION AND JUDGE PROTOCOLS

Generation roles and decoding. The condition-specific text below is the `system` message. Line wrapping is for typography only; exact prompt bytes and hashes are in the artifact. The shared `user` template is

```
# Conversation context
{context}

# Question
{question}

# Short answer
```

The reader is `Qwen/Qwen2.5-3B-Instruct` (Qwen Team, 2024), with temperature 0 and `max_tokens=2000`. The frozen run records the model identifier, decoding settings, and reader-configuration hash `a4df356f...e6411ef2d306`; it does not retain the Hugging Face weight revision, so we do not claim an exact model-repository commit. Retrieved context is backend-specific but fixed across prompt conditions for a given backend–question pair: the frozen artifacts contain zero P0/P1 context-hash mismatches, and zero P0′/P1 mismatches on Category 5. The artifact stores the text consumed by the scorers in the `prediction` field; no task-specific answer-rewrite stage is part of the released scoring pipeline. A separate transport-level record of SDK whitespace normalization was not retained, so we make no stronger claim about that layer.

P0 system prompt.

```
Answer the question using only the provided
conversation context. If the answer cannot be
determined from the context, indicate that the
information is unavailable.
```

P0′ system prompt.

```
You answer LoCoMo memory questions using only the
provided conversation context. Return the shortest
possible answer. Use exact dates when timing is asked.
If the answer is not in the context, state that you
cannot answer from the context. Do not invent details.
```

P1 system prompt.

```
You answer LoCoMo memory questions using only the
provided conversation context. Return the shortest
possible answer. Use exact dates when timing is asked.
If the answer is not in the context, answer: not
mentioned. Do not invent details; no information
available and not mentioned are the required
no-evidence style answers.
```

Thus P0′ and P1 share the brevity, date, context-only, and anti-invention instructions, while P1 changes the no-evidence instruction from a generic instruction to a more explicit required-answer format containing evaluator-recognized wording. The contrast therefore combines lexical choice, format constraint, and emphasis inside the abstention instruction; we do not interpret it as a pure lexical substitution. The exact prompt files and SHA-256 hashes are included in the artifact.

Binary abstention judge. The frozen model identifier supplied to the OpenAI-compatible endpoint is `gpt-4o`, temperature 0, `max_tokens=8`. The system message is:

```
You classify a response to a question about a conversation.
```

The user message is:

```
Conversation question: {question}
Response: {response}

Does the response decline to answer because the
answer cannot be determined from the supplied
conversation, WITHOUT subsequently attempting a
concrete answer?
```

```
Answer with exactly one word: ABSTAIN or ATTEMPT.
ABSTAIN = declines and offers no concrete answer.
ATTEMPT = gives, guesses or invents a concrete answer,
including when it declines first and then guesses.
```

Three-class answerable judge. It uses the same frozen `gpt-4o` identifier, temperature 0, and `max_tokens=8`. The system message is:

```
You grade an answer to a question about a conversation.
You are given the question, the reference answer and a
model prediction.
```

The user message is:

```
Question: {q}
Reference answer: {g}
Prediction: {p}

Label the prediction with exactly one word:
CORRECT = the prediction agrees with the reference
in substance (entities, dates, numbers); wording,
length and extra framing do not matter.
INCORRECT = wrong entity, wrong date, contradictory,
or partial-but-misleading.
ABSTAINED = the prediction declines to answer although
a reference answer exists.
```

The frozen configuration records only the model identifier `gpt-4o`; it does not retain a dated provider-side snapshot identifier. We therefore cannot reconstruct a more precise model revision and treat this as a reproducibility limitation rather than inventing one.

C Public Re-Scoring and Reporting Census

We re-score eight public prediction conditions from the MemTensor/MemOS evaluation release (MemTensor, 2025) on 1,528 common answerable keys. Here, we compare *micro token F1* and the archived *micro semantic-judge* score. Among the $\binom{8}{2} = 28$ strict pairwise point-estimate orderings, 12 reverse sign between those two scoring rules; there are no exact ties. On the five short-answer conditions {memobase, MemOS, MemU, supermemory, Zep-short}, none of the $\binom{5}{2} = 10$ orderings flips. Across the eight conditions, the Pearson correlation between mean whitespace-token answer length and (micro judge − micro F1) is $r = 0.947$ ($n = 8$). These are descriptive fixed-output comparisons, not significance tests.

Table 5: Public prediction-source map for Figure 3. All files come from `MemTensor/MemOS_eval_result`; exact download URLs, byte counts, and SHA-256 values are in `source_inventory.csv`. The Zep names below are the source archive's directory labels; we do not infer additional undocumented protocol details from those names.

Paper condition	Source archive path	Interpretation supported by source material
memobase	`locomo/memobase/memobase_locomo_judged.json`	archived judged predictions
MemOS	`locomo/memos/memos-api_locomo_judged.json`	archived judged predictions
MemU	`locomo/memu/memu_locomo_judged.json`	archived judged predictions
supermemory	`locomo/supermemory/supermemory_locomo_judged.json`	archived judged predictions
Zep-short	`locomo/zep/short_answer/zep_locomo_judged.json`	short-answer archive; Zep README says this is the reported aligned setting
Zep-official	`locomo/zep/official/zep_locomo_judged.json`	source archive label `official`
Zep-repro	`locomo/zep/official_reproduce/zep_locomo_judged.json`	source archive label `official_reproduce`
MIRIX	`locomo/mirix/evaluation_metrics.json`	archived evaluation predictions/metrics

	Micro		Macro-category		Macro-conversation	
	F1	Judge	F1	Judge	F1	Judge
memobase	2	3	2	3	2	3
MemOS	1	1	1	2	1	1
MemU	4	7	4	6	5	8
supermemory	5	8	5	8	4	7
Zep-short	3	6	3	7	3	6
Zep-official	7	2	7	1	7	2
Zep-repro	8	4	8	4	8	4
MIRIX	6	5	6	5	6	5

Rank 1 = highest point estimate; ranks are descriptive, not significance tests.

Figure 3: Public fixed-output re-scoring as an 8×6 rank matrix. Columns cross two scoring rules (F1 and semantic judge) with three aggregation rules (micro, macro-category, macro-conversation). Each cell prints the rank of the point estimate (1=highest); shading is redundant with the printed number. The matrix is descriptive and does not imply statistically significant differences between systems.

The source release states that all reported LoCoMo responses use a short-answer prompt except MIRIX, and its Zep README states that the `short_answer` version is reported to align evaluation settings. Because generation protocols remain heterogeneous, this public analysis provides observational breadth rather than a controlled test of generation effects. The targeted reporting census contains 30 candidate works/systems; 28 have verifiable LoCoMo results in inspected artifacts. Nine report a numeric Category-5 result, at least three explicitly document Category-5 inclusion in a headline metric, and nine explicitly document exclusion. The supported conclusion is contract diversity, not prevalence.

D Controlled Experiment and Decomposition

The controlled dataset contains $4 \times 1{,}986 \times 2 = 15{,}888$ P0/P1 predictions. Mean all-item shift is 0.213724. Answerable archived-local token-F1 deltas by backend are BM25 +0.098602, dense +0.066880, Mem0 +0.088240, and MemoryOS +0.035126; their equal-weight mean is +0.072212. Category-5 lexical deltas are +0.627802, +0.713005, +0.645739, and +0.822870. Under the all-item mixture, mean weighted contributions are +0.0560 from answerable archived-local token F1 and +0.1577 from Category-5 lexical credit.

The P0′ set contains 1,784 outputs. Released Category-5 credit changes from 0.057735 to 0.715247, $\Delta = 0.657511$, compared with 0.702354 for the broad P0→P1 Category-5 contrast; $0.657511/0.702354 = 0.9362$.

E Category-5 Human Validation Details

Both annotators label all 1,440 responses independently. Agreement is 1,436/1,440=0.9972; Gwet AC1=0.9948. For the pre-specified P0′–P1 contrast, $n = 478$ consensus-complete question–backend cells. Question-cluster bootstrap gives: $\Delta L = 0.667364$ CI [0.612159,0.721174], $\Delta S = 0.259414$ CI [0.212500,0.308977], and $G = 0.407950$ CI [0.346555,0.467641]. Conversation-cluster sensitivity for G is [0.3263,0.4937]. Released-rule confusion against human consensus is P0 TP=4/FN=322/FP=0/TN=152 ($n = 478$); P0′ 29/196/0/253 ($n = 478$); P1 350/1/0/129 ($n = 480$). The condition-level confusion matrices use every response with human consensus; P1 has no anno-

tator disagreements, whereas P0 and P0′ each lose two. The paired P0′–P1 estimand therefore uses 478 cells with consensus on both members. The earlier human round is retained only as exploratory provenance because its instructions exposed the research direction.

F Evaluator Repair and Semantic-Judge Protocol

The deterministic split uses development conversations conv-41 and conv-48; the other eight conversations form the natural-output test set (1,150 consensus items), keeping every conversation wholly in one split. The expanded baseline is exactly the pre-existing `_NO_INFORMATION_PATTERNS` list in the archived local scorer: `no information`, `not mentioned`, `not provided`, `not specified`, `unknown`, `cannot determine`, `can't determine`, `do not know`, and `don't know`. We did not search over phrase subsets or optimize this list on the two development conversations or on the synthetic challenge scores. On test, the released two-phrase rule has human agreement 0.624, recall 0.420, specificity 1.000, and FPR 0.000; the expanded nine-pattern rule has agreement 0.665, recall 0.485, specificity 0.998, and FPR 0.002.

The Phase-C2 record states that the GPT-4o semantic-judge contract was frozen before test calls, the development split was used only to confirm the interface, and no prompt tuning or model substitution occurred after freeze. The 51-item synthetic set is a separate diagnostic with definitional labels; it was not used to estimate natural prevalence or select the nine-pattern list. GPT-4o has agreement 0.911, recall 0.995, specificity 0.758, and FPR $98/(98+307) = 0.242$ on natural test outputs, and accuracy 0.941 on the synthetic set versus 0.549 for both lexical rules.

For answerable questions, the separate three-class judge sees question, reference answer, and prediction. CORRECT requires substantive agreement with the reference; INCORRECT covers conflicting or misleading content; ABSTAINED means declining despite an available reference. The paper's primary semantic metric is CORRECT rate.

G Answerable Semantic Results

Table 6: Answerable controlled deltas by backend. Token F1 uses the archived local scorer specified in Appendix A. The pooled GPT-4o CORRECT rate decreases from .3070 to .2544.

Backend	Δ archived token F1	Δ CORRECT
BM25	+.0986	-.0682
Dense	+.0669	-.0500
Mem0	+.0882	-.0675
MemoryOS	+.0351	-.0247
Equal-weight mean	+.0722	-.0526

The question-clustered pooled CI for ΔCORRECT is $[-0.0623, -0.0429]$ ($B = 10{,}000$, seed 20260922). Robustness deltas are -0.0526 at abstention weight 0 (primary), -0.1664 at weight 0.5, -0.2802 at weight 1, and -0.1444 when abstentions are excluded from condition means. Only the weight-0 statistic is referred to as reference-answer correctness.

H Answerable Human Validation and AC1 Correction

The released package history records that the human instrument was fixed before either annotator received the package and before any human label existed. Sampling seed 20261015 selects 240 paired units, stratified by four backends × four answerable categories × 15; all 240 question identifiers are unique. Both annotators label all 480 responses, blinded to backend, condition, token F1, GPT-4o label, expected direction, and unit membership. Twenty-five units have identical P0/P1 text and remain in the primary estimand.

The all-unit human estimate is -0.0396, question-cluster CI $[-0.0729, -0.0083]$, conversation-cluster sensitivity $[-0.0806, -0.0110]$. Same-unit comparators are GPT-4o -0.0542 (question-cluster CI $[-0.0958, -0.0125]$) and token F1 +0.0414.

Table 7: Answerable human labels by annotator and prompt. Counts are CORRECT/INCORRECT/ABSTAINED. The annotators differ in absolute correctness propensity but agree on the direction of the paired prompt effect.

Annotator	P0 counts	P1 counts	ΔCORRECT	95% question-cluster CI
A	56/88/96	46/91/103	-.0417	[-.0792,-.0083]
B	80/65/95	71/66/103	-.0375	[-.0750,-.0042]

Across all 480 labels, A marks 21.25% CORRECT and B marks 31.46% CORRECT, a 10.21-point level difference. The two annotator-specific paired effects are similar in magnitude despite the large level difference; the paired design defines within-annotator contrasts but does not guarantee that threshold differences cancel. Per-backend effects are BM25 -0.0333, dense -0.0083, Mem0 -0.0667, MemoryOS -0.0500. Sensitivities: consensus subset ($n = 207$) -0.0435; excluding identical text ($n = 215$) -0.0442; abstention weights 0/0.5/1 give -0.0396/-0.0240/-0.0083.

Table 8: Answerable human-label confusion matrix over all 480 responses. Rows are annotator A labels and columns are annotator B labels. The asymmetry is concentrated at the CORRECT/INCORRECT boundary.

A \ B	CORRECT	INCORRECT	ABSTAINED
CORRECT	100	2	0
INCORRECT	51	128	0
ABSTAINED	0	1	198

Thus $P_a = 426/480 = 0.8875$. For standard nominal multi-category Gwet AC1 (Gwet, 2008),

$$P_e = \frac{1}{q-1}\sum_{k=1}^{q}\pi_k(1-\pi_k), \qquad AC1 = \frac{P_a - P_e}{1 - P_e}.$$

With $q = 3$, pooled proportions $(0.263542, 0.322917, 0.413542)$ give $P_e = 0.327627$ and corrected AC1=0.832682. The frozen one-shot script omitted $1/(q-1)$, used $P_e = 0.655254$, and emitted 0.673673. The script and original result fields are preserved unchanged; a separate correction file and recomputation script are released. No effect estimate or confidence interval depends on this correction. Because it was identified after labels were observed, the paper conservatively reports the study as corroborative rather than using the correction to strengthen its status post hoc.

GPT-4o produces 90 ABSTAINED labels versus 199 and 198 for humans A and B; 109 and 108 human-abstained responses, respectively, are labeled INCORRECT by GPT-4o. Human disagreement is almost entirely CORRECT/INCORRECT (53/54 disagreements), not abstention.

I Freeze Timeline and Provenance

The Category-5 second-round annotation materials have an artifact-local freeze record at `2026-09-22T13:35:11`; the returned A/B label files are separately recorded and hashed at `2026-09-22T14:41:57`. The timestamp strings do not encode a timezone and there was no external preregistration service, so we describe this study as *pre-specified and hash-frozen before label return*, not as externally registered. The hash identifies the frozen file version; the ordering claim comes from the artifact-local freeze record and the separately retained later label-return record, not from the hash alone. The frozen protocol required new annotators who had not seen the hypothesis, pilot results, or round-1 labels/results. Eligibility-attestation templates were included in the frozen package, but the retained copies remain blank; no completed pre-annotation attestations are available. We disclose that gap rather than reconstructing eligibility evidence after the fact.

For the answerable human study, the released v1–v5 package history contains successive `PRE_ANNOTATION_AMENDMENT*.md` records, each explicitly stating that no human label existed and neither annotator had received the package. These are retained internal records rather than third-party timestamping. The one-shot output and frozen analysis script are preserved; the standard AC1 correction is additive and post-analysis.

Table 9: Bootstrap specifications for paper-facing uncertainty statements.

Analysis	Resampling unit	B	Seed	Estimand / weighting
Category-5 primary $\Delta L, \Delta S, G$	question identifier	10,000	20260922	consensus-complete question–backend pairs; backends enter through observed paired cells
Category-5 sensitivity	conversation	10,000	20260922	same paired estimands, conversations resampled
Answerable GPT-4o ΔCORRECT	question identifier	10,000	20260922	all 1,540 questions; each cluster carries its four backend paired differences
Answerable human d_{Human}	all 240 paired units (unstratified bootstrap)	10,000	20260922	mean of annotator-specific P1−P0 CORRECT indicators inside every replicate
Answerable human sensitivity	conversation	10,000	20260922	same all-240-unit estimand, conversations resampled
Controlled backend example	conversation	10,000	20260922	same sampled conversation indices shared across compared backends

J Statistical Protocols and Artifact Navigation

Bootstrap specifications. Table 9 collects the uncertainty procedures used in the paper. All reported intervals are percentile intervals; paired observations are kept together within the resampled cluster.

Matched answerable diagnostics. The per-annotator A/B intervals and the matched-240 GPT-4o interval in Table 3 use the same 240 sampled units, 10,000 unit-level resamples, and seed 20260922; they are additive paper-facing diagnostics. The frozen primary human interval is generated by `boot_units`: it resamples all 240 paired units with replacement without preserving the original 16 backend×category strata, recomputes A and B inside each replicate, and averages the two annotator effects. No post-hoc stratified bootstrap is substituted.

Controlled-backend comparison. The paired Mem0–dense all-item P1 example belongs to the four-backend controlled experiment, not to the eight-condition public re-scoring analysis. Using shared conversation resamples, the controlled difference is +0.0101 with 95% CI $[-0.0092, 0.0331]$; no statistical backend ordering is claimed from that interval. Figure 3 is separate: it reports descriptive ranks of eight externally released answerable prediction conditions and makes no inferential claim about pairwise superiority.

Human-label combination. In the Category-5 study, the primary P0′ →P1 estimand uses the 478 question–backend cells for which both annotators agree on both members; human semantic abstention is the consensus label. In the answerable human study, the frozen all-240 estimand does *not* switch to a consensus subset: each annotator's 240 paired CORRECT differences are computed separately and the two annotator effects are averaged, with both annotators recomputed inside each bootstrap replicate. Consensus-only analysis is a sensitivity only.

Artifact navigation. Table 10 gives concise, exact entry points. Full file-level paths and copy-paste commands are in `ARTIFACT_NAVIGATION.md` at the extracted supplementary root.

Table 10: Executable paper-to-artifact entry points, relative to the extracted supplementary root.

Paper item	Exact entry point
Fig. 2, Table 2	`ROUND2_CURRENT_GUIDE.md`
Fig. 3	`PAPER_FACING_AUDIT/make_public_rank_matrix.py`
Answerable GPT-4o	`CURRENT_RESULTS/README.md`
Answerable human	`06_PHASED_ANSWERABLE_HUMAN_VALIDATION/README_PHASE_D.md`
AC1 correction	`POST_ANALYSIS_CORRECTIONS/recompute_phaseD_standard_ac1.py`
Repair/challenge	`05_PHASEC2_SEMANTIC_JUDGE/RESULTS/repair_results_test_with_judge.json` and `04_PHASEC_REPAIR/challenge_results.csv`
Whole supplement	`python VERIFY_FINAL_SUPPLEMENTARY.py --full`

The artifact retains methodological history rather than rewriting it. The earlier human round is marked exploratory; pre-annotation versions and amendments of the answerable human study are retained with hashes; the one-shot result files remain unchanged; and the post-analysis AC1 correction is additive. Cross-reader replication was not run. Additional navigation files in the supplementary document the public-source hashes, the audited LoCoMo scorer copies, reader provenance, and the statistical procedures above.