FROM BOTTLENECK LABELS TO REPAIR SURFACES: FACTORIAL DIAGNOSIS OF AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

An oracle repair may recover many tasks from a native memory-agent pipeline, yet tell us little about where the original failure occurred if another repair recovers the same tasks. We examine this ambiguity by crossing repairs to upstream state, retrieval, and read context. In the primary setting, upstream-state and read-context repairs each add about 91 pp from the native cell, but no more than $+0.45$ pp once the other is active; they correct 199 of the same 206 native errors. Across all seven complete surfaces, the two repairs have negative accuracy interactions. Retrieval repair has smaller native gains across eleven settings, and paired intervals do not resolve a change in its gain after upstream repair. The same trajectories yield nominal $p < .05$ retrieval-background contrasts with individual judges, but none with adjudicated labels. Stage-level interpretation therefore depends on both the repair background and the scoring rule.

1 INTRODUCTION

A memory-augmented task can fail because an earlier answer enters memory incorrectly, because retrieval misses a needed record, or because the solver does not use information in its read context. Stage-level, counterfactual, and trace-based diagnostics examine these possibilities (He et al., 2026; Zhu et al., 2026; Liu et al., 2026b; Deng et al., 2026; Shi et al., 2026), often by replacing one interface with a gold counterpart and treating the largest recovered-accuracy gain as a *bottleneck label*.

That label is difficult to interpret when two repairs deliver the same missing information through different routes: correcting earlier answers before storage and supplying them directly in the solver's read context can each recover a task, although the second repair adds little once the first is active. In that case, the gain against the native pipeline does not locate a unique faulty component.

We test this dependence by crossing repairs to upstream state ($S^\star$), retrieval ($R^\star$), and read context ($U^\star$). The eight accuracies in a complete setting form a *factorial repair surface*, so a repair's gain can be compared across the native pipeline and backgrounds where another interface has already been repaired. The contrast between those two gains is the point of the design: native recovery measures what can be corrected from the observed pipeline, while the conditional gain asks what the same repair adds after another route has supplied information. Across the complete settings, their conditional gains shrink sharply; task-level overlap is measured directly in the primary setting.

Because the surface is computed from correctness labels, it can also shift when judging errors vary by intervention cell, even with the trajectories fixed. We therefore use one adjudication protocol for the primary results and rescore the same trajectories with individual reference-guided judges. This sensitivity analysis, together with the paired repair contrasts, exposes what is lost when one native-cell gain is treated as a stage label.

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.

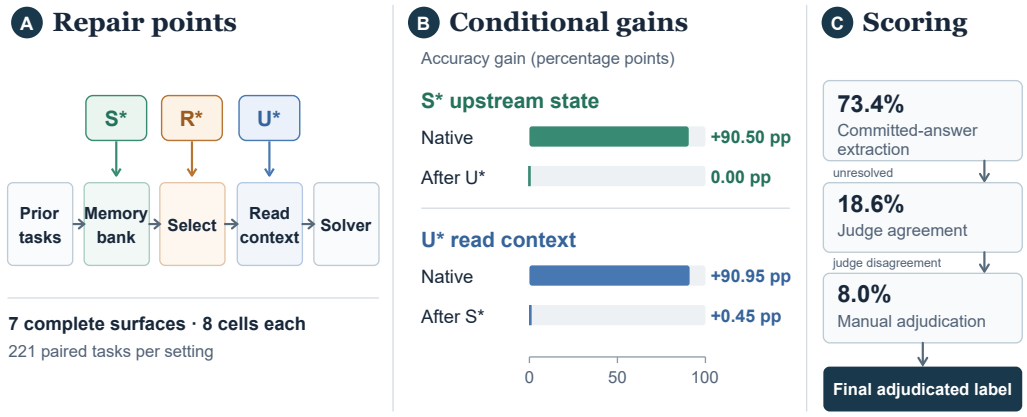

Figure 1: **Repair backgrounds and scoring in the factorial design.** **(A)** Upstream-state ($S^\star$), retrieval ($R^\star$), and read-context ($U^\star$) repairs give eight cells in each of seven complete settings. Four variants observe only $\{\varnothing, S^\star, R^\star, S^\star R^\star\}$. All cells within a setting share 221 task IDs. **(B)** In the primary setting, native-cell gains are $+90.50$ pp for $S^\star$ and $+90.95$ pp for $U^\star$; the repairs correct 199 of the same 206 native errors, and either adds at most 0.45 pp after the other. **(C)** Correctness labels come from deterministic committed-answer extraction (73.4%), dual-judge agreement (18.6%), and manual adjudication (8.0%).

2 Related Work

MemoryArena connects memory formation to later agentic tasks in the progressive-search setting used here (He et al., 2026). Paired counterfactual probes and stage-separated evaluation examine write, retrieval, and answer-time failures (Zhu et al., 2026; Shi et al., 2026); trace-based work extends the view to multimodal interaction and evolving memory graphs (Liu et al., 2026b; Deng et al., 2026), while other stress tests involve stale, changing, or latent user state (Chao et al., 2026; Xie et al., 2026; Ma et al., 2026). The repair question here concerns overlap across those interfaces: when two interventions each help, do they correct the same task errors?

Component comparisons pose a different question. On LoCoMo, Yuan et al. (2026) cross three write strategies with three retrieval methods, reporting about 20 pp of average-accuracy variation across the latter versus 3–8 pp across the former; their probes also examine memory relevance and use at the retrieval-to-generation boundary. MemoryLake (Zhan et al., 2026) compares complete MemoryArena backends that bundle writing, retrieval, consolidation, budgeting, and prompt assembly. Changing a retrieval method compares choices within a system; $R^\star$ instead re-selects chunks from the bank already produced by the agent, and the factorial cells measure how that repair's gain changes after another interface is repaired.

We describe these conditional comparisons with the Möbius/Harsanyi decomposition, which assigns main effects and interactions to a set function (Harsanyi, 1963); Shapley values average marginal contributions over intervention orders (Shapley, 1953). On the repair accuracy surface, nonzero interactions identify background-dependent gains (Proposition 1).

Conditional gains also depend on correctness labels. LLM-as-judge evaluation is widely used (Liu et al., 2023; Zheng et al., 2023), although studies have reported position, verbosity, and self-preference biases (Zheng et al., 2023; Wang et al., 2024; Panickssery et al., 2024). Here the relevant quantity is disagreement by intervention cell: if a reference-guided judge marks noncommitting answers correct more often in one cell than another, the estimated contrast can change without any change to the trajectories.

Repair	Intervention	Interpretation
$S^\star$	Replace each prior subtask's predicted answer with its gold answer immediately before the result enters memory.	Value of correct upstream state entering memory with the write mechanism fixed.
$R^\star$	Re-select chunks from the agent's own memory bank by strict match to the gold answer under the original top-k budget; use recent chunks when none match.	Value of selecting among the agent's existing chunks; missing content remains unavailable.
$U^\star$	Place prior subtasks' gold answers in the solver's read context, never the current task's answer.	Value of gold prior answers at read time with the solver unchanged.

Table 1: Repair operators. $S^\star$ and $U^\star$ supply gold prior answers at different pipeline locations; crossing them measures the overlap in their conditional gains.

3 Factorial Repair Diagnosis

3.1 Tasks and settings

We use the progressive-search environment of MemoryArena (He et al., 2026), whose tasks are derived from BrowseComp-Plus (Chen et al., 2026): each task asks a hard factoid question assembled from a fixed local corpus, and later subtasks depend on answers produced earlier in the interaction. Every setting and condition evaluates the same 221 task identifiers. The primary agent is Qwen3-8B (Yang et al., 2025) with temperature-0 local decoding and BM25, a sparse lexical retriever, for memory retrieval (Robertson & Zaragoza, 2009). Cross-setting evaluation also includes Qwen3.5-9B (Qwen Team, 2026), Meta-Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Olmo-3-7B-Instruct (Olmo et al., 2025), Nemotron-Labs-Diffusion-8B in autoregressive mode (hereafter Nemotron-Diffusion-8B) (Fu et al., 2026), Ministral-3-8B-Instruct-2512 (Liu et al., 2026a), and Qwen3-4B. The dense-memory setting replaces BM25 with a normalized dense vector index implemented in FAISS (Johnson et al., 2021) and built with Qwen3-Embedding-0.6B (Zhang et al., 2025); the Mem0 setting uses the open-source Mem0 backend (Chhikara et al., 2025). Each task–condition run has isolated task-scoped memory, and all reported comparisons are paired by task ID.

3.2 Repair operators

Each complete setting observes all eight subsets of $\{S^\star, R^\star, U^\star\}$; four additional variants observe only $(\varnothing, S^\star, R^\star, S^\star R^\star)$, so neither $U^\star$ nor its interaction with $S^\star$ is measured there. BM25 consolidation is an identity operation for this backend and is excluded from the factorial design (Appendix G), whereas Mem0's backend-specific consolidation is evaluated only within its own setting. Every comparison requires its planned cells, valid structured evaluator output, consistent trajectories, and verified intervention events; all eleven settings pass these checks (Appendix H).

3.3 Conditional marginals and paired transitions

Let $N = \{s, r, u\}$ index the repairs and $v(A)$ be mean binary correctness under repair set $A \subseteq N$. The eight values of v form the *factorial repair surface*; each repair combination is a *cell*, with $v(\varnothing)$ denoting the native cell. For repair j and background $A \subseteq N \setminus \{j\}$, the conditional marginal is

$$B_j(A) = v(A \cup \{j\}) - v(A). \tag{1}$$

Proposition 1 (Scalar summaries require vanishing interactions)**.** *For repair j, $B_j(A)$ is identical across all backgrounds $A \subseteq N \setminus \{j\}$ if and only if every Möbius/Harsanyi coefficient $m(T)$ with $j \in T$ and $|T| \geq 2$ vanishes.*

In the Möbius/Harsanyi decomposition, $v(A) = \sum_{T \subseteq A} m(T)$. The proof appears in Appendix B, with primary coefficients and Shapley averages in Appendix D. A nonzero interaction containing j changes its marginal on at least one background; a reported single-cell gain is tied to that background.

Task-level transitions decompose the marginal. Let $a(A)$ be background accuracy, $r_j(A)$ the fraction of background errors rescued by repair j, and $c_j(A)$ the fraction of background-correct tasks that it breaks. Then

$$B_j(A) = (1 - a(A))\, r_j(A) - a(A)\, c_j(A). \tag{2}$$

The first term weights corrections by the remaining error rate, whereas the second subtracts regressions among tasks already correct on the background; thus a small net gain need not imply that the repair seldom corrects errors, since few errors may remain or corrections may be offset by regressions.

3.4 Adjudicated correctness

We apply one three-stage correctness protocol to all 15,912 records. First, tolerant answer-marker parsing and a commitment scan extract the answer presented as final, excluding entities merely mentioned in the trajectory; this resolves 9,763 clean matches, 1,571 explicit abstentions (scored incorrect), and 352 scan cases, or 73.4% of records combined. Unresolved records pass to Mistral-7B-Instruct-v0.3 and Meta-Llama-3.1-8B-Instruct, which independently label them under the same strict JSON contract; their agreements are accepted, and the remaining 1,266 records (8.0%) are manually adjudicated. These final labels underlie the primary results, while both judges also score Stage 1 records so that the sensitivity analysis can compare their labels with adjudication in every cell (Section 4.3).

For each contrast, 10,000 paired bootstrap resamples draw task IDs while carrying each task's matched outcomes across cells. We report percentile 95% intervals and two-sided bootstrap tail probabilities $p = 2\min\{\Pr(\hat{\delta} \le 0), \Pr(\hat{\delta} \ge 0)\}$ capped at 1. Accuracy differences are reported in percentage points (pp).

4 Results

4.1 Upstream and read-context repairs recover the same native errors

From 15 of 221 tasks (6.79%) solved natively, accuracy rises to 97.29% under $S^\star$ and 97.74% under $U^\star$ (Figure 2A). Applying both reaches 97.74%, compared with 98.19% under all three repairs: the native gains of $+90.50$ and $+90.95$ pp shrink to $B_S(U^\star) = 0.00$ and $B_U(S^\star) = +0.45$ pp once the other repair is active. Their interaction is -90.50 pp (95% CI $[-94.12, -86.43]$, $p < .001$), quantifying their non-additivity on the accuracy scale.

The task-level transitions explain why these large native gains are nearly substitutable: $S^\star$ corrects 201 of 206 native errors, $U^\star$ corrects 202, and 199 tasks appear in both corrected-error sets (Figure 2C). Retrieval repair behaves differently. On the native background, $R^\star$ corrects 15 errors but breaks 7 of the 15 correct tasks, moving accuracy from 6.79% to 10.41% ($+3.62$ pp; 95% CI $[-0.45, +7.69]$, $p = .10$); after $S^\star$, it corrects four of six remaining errors and breaks three correct tasks, leaving a $+0.45$ pp net gain. Both the smaller error pool and the regressions enter Equation (2).

All seven complete surfaces show negative $S^\star \times U^\star$ accuracy interactions, ranging from -91.0 to -62.9 pp (all $p < .001$), while gains after the other repair are bounded by $B_U(S^\star) \le 9.50$ pp and $B_S(U^\star) \le 6.33$ pp. The largest observed conditional gain rises to $+14.03$ pp when $R^\star$ is also active, so the conditional gains remain nonzero on some backgrounds. Each nonzero interaction violates the background-invariance condition in Proposition 1.

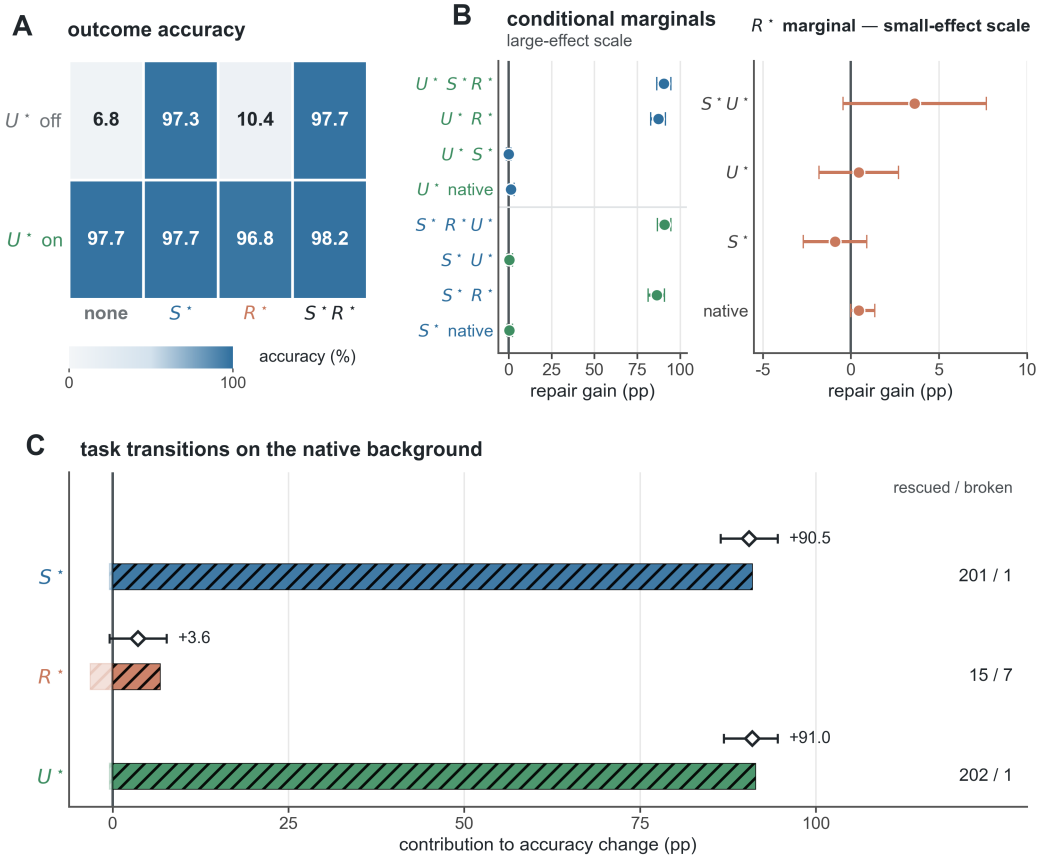

Figure 2: **Primary adjudicated repair surface (221 paired tasks). (A)** Accuracy in all eight cells, including 6.79% natively, 97.29% under $S^\star$, and 97.74% under $U^\star$. **(B)** Conditional marginals: the native gains of $+90.50$ and $+90.95$ pp fall to 0.00 and $+0.45$ pp, respectively, when the other repair is active. Error bars mark paired 95% bootstrap intervals. **(C)** Native-background transitions. Dark bars show corrected errors, light bars newly incorrect tasks, and diamonds net changes with paired 95% bootstrap intervals; the right column gives the corresponding counts. $S^\star$ and $U^\star$ correct 201 and 202 of 206 native errors; $R^\star$ corrects 15 and breaks 7 of the 15 native-correct tasks.

4.2 All seven complete settings show $S^\star \times U^\star$ non-additivity

Using the final adjudicated labels, native accuracy ranges from 1.36% to 9.05% across the eleven settings (Table 2). Upstream-state repair adds 69.2–92.3 pp in every setting, and read-context repair adds 72.4–95.0 pp on the seven complete surfaces. Among the four cells shared with the primary experiment, the BM25 replication differs by at most 2.27 pp; the corresponding maximum for the dense replication is 0.90 pp.

Retrieval repair contributes a much smaller native-cell change, from -1.81 to $+4.08$ pp across the eleven settings. Its gain changes by -2.27 to $+7.24$ pp after upstream-state repair, measured as $B_R(\varnothing) - B_R(S^\star)$, but every paired interval includes zero (smallest $p = .0998$; Figure 3A). The intervals therefore do not resolve the direction of this contrast.

Setting	Native	$S^\star$	$U^\star$	Full	$B_R(\varnothing)$	$m(\{s,u\})$
		accuracy (%)			pp	
Qwen3-8B / BM25	6.79	97.29	97.74	98.19	+3.62	−90.50
Qwen3-8B / dense	7.69	98.19	96.38	98.19	+0.91	−88.69
Qwen3.5-9B / BM25	4.52	95.02	97.74	97.74	+1.36	−90.50
Llama-3.1-8B / BM25	9.05	89.14	85.97	90.50	−1.36	−76.47
Olmo-3-7B / BM25	4.07	92.31	92.76	94.57	−0.90	−88.24
Nemotron-Diffusion-8B (AR) / BM25	1.36	92.76	96.38	97.74	+0.90	−90.95
Ministral-3-8B / BM25	7.24	76.47	79.64	85.52	−1.81	−62.90
Qwen3-8B / Mem0†	3.17	73.76	–	–	+2.71	–
Qwen3-8B / BM25 replication†	4.52	96.83	–	–	+4.08	–
Qwen3-4B / BM25†	4.07	96.38	–	–	+4.07	–
Qwen3-8B / dense replication†	8.14	98.19	–	–	+1.36	–

Table 2: Accuracy and repair effects on 221 paired tasks per setting. The first seven rows have complete 2^3 surfaces; † marks four-cell variants. $B_R(\varnothing)$ is the native retrieval gain, and $m(\{s,u\})$ the $S^\star \times U^\star$ interaction. All observed cell accuracies appear in Appendix A (Table 4).

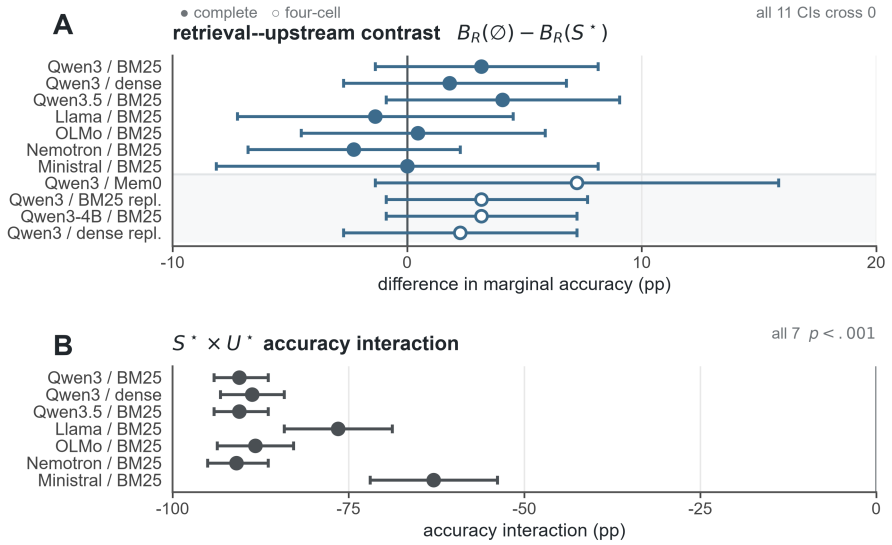

Figure 3: **Repair contrasts across settings.** **(A)** The change in retrieval gain after $S^\star$, $B_R(\varnothing) - B_R(S^\star)$, has eight positive, two negative, and one zero point estimates. Every paired 95% interval includes zero (smallest $p = .0998$). **(B)** The $S^\star \times U^\star$ interaction is -91.0 to -62.9 pp in all seven complete settings (all $p < .001$).

4.3 Single-judge labels change retrieval contrasts

Single-judge disagreement with adjudicated labels is 1.6–7.6× higher in the native/$R^\star$ pair than in the other repaired cells (Table 3); Mistral-7B alone produces 1,187 false positives and 16 false negatives in this pair. Two cases in Appendix F show reference-guided judges marking a noncommitting response correct, although the excerpts omit the full input visible to each judge.

Pooled across settings, Mistral false positives relative to adjudicated labels make up 21.8% of native-cell records and 27.1% of $R^\star$-cell records; the $R^\star$ fraction is higher in 8 of 11 settings (Figure 4B).

The retrieval-background contrasts reach nominal $p < .05$ in 6/11 settings under Mistral labels, compared with 5/10 for self-judging and 4/11 for Llama labels; their respective maxima are $+17.2$, $+13.1$, and $+13.1$ pp. Under adjudication, all eleven intervals include zero.

Scoring layer	Disagreement with adjudicated label: Native/$R^\star$ cells	Disagreement with adjudicated label: Other repaired cells	Retrieval contrast: Nominal $p < .05$	Retrieval contrast: Max. (pp)
Self-judge (10 settings)	33.4% (1392/84)	4.4% (239/189)	5 / 10	+13.1
Mistral-7B	24.7% (1187/16)	4.4% (261/225)	6 / 11	+17.2
Llama-3.1-8B	18.1% (839/43)	11.5% (189/1078)	4 / 11	+13.1
Adjudicated	–	–	**0 / 11**	+7.2

Table 3: Judge disagreement relative to final adjudicated labels. Parentheses give false-positive/false-negative counts. The native/$R^\star$ pair defines $B_R(\varnothing)$; all other repaired cells form the comparison group. The last two columns summarize $B_R(\varnothing) - B_R(S^\star)$ using nominal, unadjusted $p < .05$ counts and the largest observed contrast.

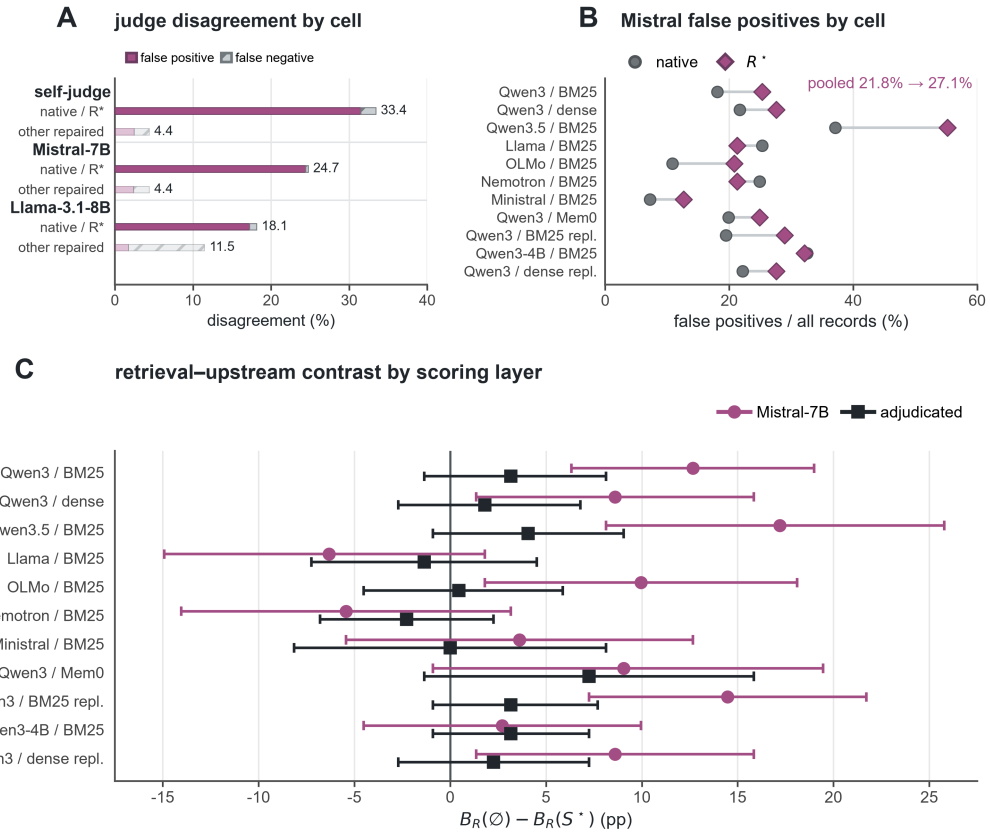

Figure 4: **Cell-dependent judge disagreement and retrieval contrasts.** **(A)** Disagreement with adjudicated labels, split into false positives (purple) and false negatives (hatched). Dark bars pool native/$R^\star$ cells; light bars pool other repaired cells. **(B)** Mistral false positives relative to adjudicated labels, divided by all records in each native or $R^\star$ cell. The pooled fractions are 21.8% and 27.1%, respectively. **(C)** Retrieval-background contrasts from Mistral labels (purple circles) and adjudicated labels (black squares). All adjudicated intervals include zero.

Scoring also changes the estimated size of the primary $S^\star \times U^\star$ interaction: it ranges from -58 to -74 pp under individual judges and is -90.5 pp after adjudication. Unlike the retrieval-background contrast, its sign remains negative under every scoring layer examined.

5 Implications for Diagnostic Design

The overlap between $S^\star$ and $U^\star$ changes what a large oracle gain can mean. One repair changes prior answers before storage; the other supplies them at read time, yet either route can restore information needed for the task. Recovery through both routes does not identify the native failure location. The native $R^\star$ gain concerns only selection from the agent's existing memory bank under the retriever used here. A stage-level report should give the native gain alongside conditional gains for the observed backgrounds, naming the repair and background in each case. Task-level rescued and broken counts then show how remaining error headroom and regressions contribute to a small net gain.

Because a repair contrast subtracts accuracies from different cells, unequal label errors need not cancel. The observed imbalance in single-judge disagreement would be hidden by one aggregate agreement rate. A cellwise error audit and an explicit committed-answer field would make that measurement more transparent.

6 Limitations

$S^\star$ changes upstream answers while leaving the write mechanism intact, whereas $U^\star$ supplies prior answers at read time without changing the solver; their overlap therefore identifies substitutable information routes, not write fidelity or solver utilization in isolation. The study uses one benchmark family, one strict committed-answer endpoint, and 221 paired tasks per setting, with the same task IDs reused across settings rather than independent task samples. Most retrieval-background intervals span about ± 5 pp, and the widest spans about ± 8.6 pp, so 8/11 positive point estimates do not establish a pooled direction. Large negative accuracy interactions can also reflect limited headroom after either strong repair; task-level overlap is measured directly only for the primary setting. Dual-judge agreements contribute to 18.6% of final labels, so disagreement with adjudication is not an independent human-reference error estimate. One manual adjudicator resolves 8.0%, and we did not obtain independent human labels for every record or measure annotation time and cost. Finally, each additional repair doubles the cells required for a complete factorial design.

7 Conclusion

In the primary setting, upstream-state and read-context repairs recover many of the same native errors, though they supply prior answers at different points in the pipeline. Across seven complete settings, their accuracy gains are strongly non-additive. Their conditional effects show what an isolated native-cell gain cannot: how much of the recovery survives another repair. Retrieval contrasts also change when the same trajectories receive different correctness labels. These results motivate treating the repair surface, rather than a bottleneck label from the unrepaired pipeline alone, as the diagnostic object.

Reproducibility Statement

The task pairing, repair semantics, factorial cells, scoring protocol, and bootstrap procedure are specified in Sections 3–4.3. The appendices give the full surfaces, uncertainty tables, transition counts, adjudication examples, and intervention checks. Reported statistics use fixed task-paired experimental records. Model checkpoints and the full trajectory store are not embedded in the manuscript package, so an end-to-end rerun requires the external experiment archive. The Llama-3.1-8B self-judge sensitivity layer is omitted because its corresponding extraction output failed validity checks; the independent Mistral and final adjudicated layers remain included.

Ethics Statement

This work uses benchmark data and local model inference and collects no new personal data. Manual adjudication is limited to benchmark responses under a fixed correctness protocol.

AI USE STATEMENT

Generative AI tools were used for code and analysis review, experiment orchestration, figure preparation, and drafting and language editing of portions of the manuscript. All AI-assisted outputs were reviewed by the authors, and numerical claims were checked against the experimental records. The authors take responsibility for the final content of the paper.

REFERENCES

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. Stale: Can llm agents know when their memories are no longer valid?, 2026. URL https://arxiv.org/abs/2605.06527.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-plus: A fair and disentangled evaluation benchmark for deep search agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22349–22370, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1023. URL https://aclanthology.org/2026.acl-long.1023/.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

Xinle Deng, Ruobin Zhong, Hujin Peng, Xiaoben Lu, Yanzhe Wu, Guang Li, Buqiang Xu, Yunzhi Yao, Jizhan Fang, Haoliang Cao, Junjie Guo, Yuan Yuan, Ziqing Ma, Yuanqiang Yu, Rui Hu, Baohua Dong, Hangcheng Zhu, and Ningyu Zhang. Memtrace: Tracing and attributing errors in large language model memory systems, 2026. URL https://arxiv.org/abs/2605.28732.

Yonggan Fu, Lexington Whalen, Abhinav Garg, Chengyue Wu, Maksim Khadkevich, Nicolai Oswald, Enze Xie, Daniel Egert, Sharath Turuvekere Sreenivas, Shizhe Diao, Chenhan Yu, Ye Yu, Weijia Chen, Sajad Norouzi, Jingyu Liu, Shiyi Lan, Ligeng Zhu, Jin Wang, Jindong Jiang, Morteza Mardani, Mehran Maghoumi, Song Han, Ante Juki, Nima Tajbakhsh, Jan Kautz, and Pavlo Molchanov. Nemotron-labs-diffusion: A tri-mode language model unifying autoregressive, diffusion, and self-speculation decoding, 2026. URL https://arxiv.org/abs/2607.05722.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie

Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. doi: 10.2307/2525487.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Arthur Wu, Siru Ouyang, Zihan Wang, Jiaxin Pei, Julian McAuley, Yejin Choi, and Alex Pentland. Memoryarena: Benchmarking agent memory in interdependent multi-session agentic tasks, 2026. URL https://arxiv.org/abs/2602.16313.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle

Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3, 2026a. URL https://arxiv.org/abs/2601.08584.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, Charese Smiley, Xiaomo Liu, James Zou, Sheng Liu, Yuheng Bu, Songyou Peng, and Xin Eric Wang. Worldmemarena: Evaluating multimodal agent memory through action-world interaction, 2026b. URL https://arxiv.org/abs/2605.29341.

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 Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Enze Ma, Yufan Zhou, Wei-Chieh Huang, Jie Yang, Huanhuan Ma, Zixuan Wang, Chengze Li, Chunyu Miao, Philip S. Yu, and Zhen Wang. Memprobe: Probing long-term agent memory via hidden user-state recovery, 2026. URL https://arxiv.org/abs/2606.24595.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2197. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f1f0218e45f5414c79c0679633e47bc-Abstract-Conference.html.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

L. S. Shapley. A value for n-person games. In H. W. Kuhn and A. W. Tucker (eds.), *Contributions to the Theory of Games (AM-28), Volume II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–318. Princeton University Press, 1953. URL https://www.jstor.org/stable/j.ctt1b9x1zv.24.

Zitong Shi, Yixuan Tang, and Anthony Kum Hoe Tung. A-tma: Decoupling state-aware memory failures in long-term agent memory, 2026. URL https://arxiv.org/abs/2607.01935.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Wenya Xie, Shengming Zhou, Zelin Li, Pouya Parsa, Shuang Zhou, Xinheng Ding, Chinmay Arvind, Guanchu Wang, Vladimir Braverman, Ali Payani, Yantao Zheng, and Zirui Liu. Dynamicmem: A long-horizon memory benchmark in real-world settings, 2026. URL https://arxiv.org/abs/2606.22877.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng

Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Boqin Yuan, Yue Su, and Kun Yao. Diagnosing retrieval vs. utilization bottlenecks in llm agent memory, 2026. URL https://arxiv.org/abs/2603.02473.

Chaoqun Zhan, Qiang Zhou, Guannan Li, Zhenqiang Huang, and Qianjin Wang. Memorylake on memoryarena: A matched study of agent memory backends, 2026. URL https://arxiv.org/abs/2608.13883.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

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, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Jianing Zhu, Yeonju Ro, John Robertson, Kevin Wang, Junbo Li, Haris Vikalo, Aditya Akella, and Zhangyang Wang. Your agents are aging too: Agent lifespan engineering for deployed systems, 2026. URL https://arxiv.org/abs/2605.26302.

A Full Cross-Setting Repair Surfaces

Setting	Native	$S^\star$	$R^\star$	$U^\star$	$S^\star R^\star$	$S^\star U^\star$	$R^\star U^\star$	$S^\star R^\star U^\star$
Qwen3-8B / BM25	6.79	97.29	10.41	97.74	97.74	97.74	96.83	98.19
Qwen3-8B / dense	7.69	98.19	8.60	96.38	97.29	98.19	96.83	98.19
Qwen3.5-9B / BM25	4.52	95.02	5.88	97.74	92.31	97.74	98.19	97.74
Llama-3.1-8B / BM25	9.05	89.14	7.69	85.97	89.14	89.59	76.47	90.50
Olmo-3-7B / BM25	4.07	92.31	3.17	92.76	90.95	92.76	90.95	94.57
Nemotron-Diffusion-8B (AR) / BM25	1.36	92.76	2.26	96.38	95.93	96.83	97.29	97.74
Ministral-3-8B / BM25	7.24	76.47	5.43	79.64	74.66	85.97	80.09	85.52
Qwen3-8B / Mem0†	3.17	73.76	5.88	–	69.23	–	–	–
Qwen3-8B / BM25 replication†	4.52	96.83	8.60	–	97.74	–	–	–
Qwen3-4B / BM25†	4.07	96.38	8.14	–	97.29	–	–	–
Qwen3-8B / dense replication†	8.14	98.19	9.50	–	97.29	–	–	–

Table 4: Final adjudicated accuracy (%) for every observed repair cell. † denotes four-cell settings. Each row uses the same 221 paired task IDs across its observed conditions.

B Factorial Algebra and Scalar Stability

For each operator j, define the observed marginal range

$$I_j \;=\; \max_{A\subseteq N\setminus\{j\}} B_j(A) \;-\; \min_{A\subseteq N\setminus\{j\}} B_j(A). \tag{3}$$

$I_j = 0$ exactly when the measured marginal is constant across all backgrounds of the complete design. In the Möbius/Harsanyi decomposition, $v(A) = \sum_{T\subseteq A} m(T)$, where each repair set T has a unique coefficient and $|T| \geq 2$ denotes an interaction. Its contribution to the marginal is $B_j(A) = \sum_{T\subseteq A} m(T\cup\{j\})$.

Proposition 2 (Restatement of Proposition 1)**.** *For operator j, $B_j(A)$ is identical for every background $A \subseteq N \setminus \{j\}$ if and only if $m(T) = 0$ for every T containing j and at least one other operator.*

Proof. If all interactions containing j vanish, then for every A, $B_j(A) = \sum_{T\subseteq A} m(T\cup\{j\}) = m(\{j\})$. Conversely, the set function $A \mapsto B_j(A) - B_j(\varnothing)$ over $N \setminus \{j\}$ has Möbius coefficients $m(T\cup\{j\})$ for nonempty $T \subseteq N \setminus \{j\}$. A constant marginal makes this function zero everywhere, and uniqueness of the transform sets all those coefficients to zero. A fractional design can check constancy only on its observed backgrounds. □

For example, $m(\{s,r\}) = v(\{s,r\}) - v(\{s\}) - v(\{r\}) + v(\varnothing)$; the other pairwise coefficients follow the same rule. Exact Shapley values average each repair's marginal over six orders. Those averages need not reveal how the marginal changes by background (Appendix D).

C Primary Conditional Marginals with Uncertainty

Operator	Background A	$B_j(A)$ (pp)	Paired 95% CI (pp)
$S^\star$	native	+90.50	$[+86.43, +94.57]$
	$R^\star$	+87.33	$[+82.81, +91.40]$
	$U^\star$	$+0.00$	$[-1.36, +1.36]$
	$R^\star U^\star$	+1.36	$[+0.00, +3.17]$
$R^\star$	native	+3.62	$[-0.45, +7.69]$
	$S^\star$	+0.45	$[-1.81, +2.71]$
	$U^\star$	-0.90	$[-2.71, +0.90]$
	$S^\star U^\star$	+0.45	$[+0.00, +1.36]$
$U^\star$	native	+90.95	$[+86.88, +94.57]$
	$S^\star$	+0.45	$[-0.90, +2.26]$
	$R^\star$	+86.43	$[+81.45, +90.95]$
	$S^\star R^\star$	+0.45	$[-0.90, +2.26]$

Table 5: Primary-setting conditional marginals with paired 95% intervals from 10,000 task bootstrap resamples. The $S^\star$ gain falls from $+90.50$ pp natively to 0.00 pp after $U^\star$; the $U^\star$ gain falls from $+90.95$ to $+0.45$ pp after $S^\star$.

D Primary Möbius Decomposition and Shapley Averages

Term	Value (pp)	Paired 95% CI (pp)
$m(\{s\})$	+90.50	–
$m(\{r\})$	+3.62	$[-0.45, +7.69]$
$m(\{u\})$	+90.95	–
$m(\{s,r\})$	-3.17	$[-8.14, +1.36]$
$m(\{s,u\})$	-90.50	$[-94.12, -86.43]$
$m(\{r,u\})$	-4.52	$[-9.05, -0.00]$
$m(\{s,r,u\})$	$+4.52$	$[-0.45, +9.50]$

Table 6: Möbius/Harsanyi coefficients of the primary adjudicated surface ($p < .001$ for $m(\{s,u\})$; $p = .19, .046, .067$ for the remaining interaction terms in the order listed).

Across the six repair orders, the exact Shapley averages are $+45.17$ pp for $S^\star$, $+1.28$ pp for $R^\star$, and $+44.95$ pp for $U^\star$. The shared $S^\star/U^\star$ gain is therefore allocated roughly evenly, although the average hides its variation by background and can be shared by different repair surfaces.

E TRANSITION DETAILS

Repair	Background	Acc. (%)	Errors	Rescued	Broken	$B_j(A)$ (pp)
$S^\star$	native	6.79	206	201 (97.6%)	1/15	+90.50
$U^\star$	native	6.79	206	202 (98.1%)	1/15	+90.95
$R^\star$	native	6.79	206	15 (7.3%)	7/15	+3.62
$R^\star$	after $S^\star$	97.29	6	4 (66.7%)	3/215	+0.45
$R^\star$	after $U^\star$	97.74	5	1 (20.0%)	3/216	−0.90

Table 7: Task transitions in the primary setting (221 paired tasks). "Rescued" counts background errors corrected by the repair, with its percentage in parentheses; "Broken" counts background-correct tasks turned incorrect. $S^\star$ and $U^\star$ correct 199 of the same 206 native errors.

F ADJUDICATION PROTOCOL AND JUDGE FALSE POSITIVES

Adjudication counts. Deterministic extraction resolves 11,686/15,912 records (73.4%): 9,763 committed-answer matches, 1,571 explicit abstentions, and 352 commitment-scan resolutions. Dual-judge agreement resolves 2,960 more (18.6%); 1,266 (8.0%) require manual arbitration. Both passes return parseable structured JSON for every record, using 24,000-token head+tail compaction for long inputs. Each scoring layer uses seed 0 and 10,000 paired bootstrap resamples with percentile intervals.

Noncommitting answers scored correct. Table 8 gives two primary-setting responses that do not commit to an answer: at least one reference-guided judge marks each correct, although both final labels are incorrect. The commitment scan assigns the first response an incorrect label despite judge agreement; the second case reaches manual arbitration. Because the excerpts omit full judge input, neither case establishes the cause of disagreement.

Response excerpt	Gold answer	Mistral	Llama	Final
"Unable to determine the name of the school." ($R^\star$ cell)	University of Ottawa	✓	✓	incorrect
"No person that fits all the given details was identified." (native cell)	Lambros Vorloou	✓	×	incorrect

Table 8: Two noncommitting responses scored correct by reference-guided judges in the primary setting. Checkmarks mark single-judge correct labels; in the first example, deterministic commitment parsing assigns the final incorrect label despite agreement between the judges.

G CONSOLIDATION CONTROL

BM25 consolidation returns each input chunk unchanged, so a fourth factorial axis would duplicate existing cells. Matched runs bound drift in final answers and retrieval lists; the toggle enters no reported surface or statistic. Mem0 consolidation is backend-specific and evaluated only within its own setting.

H COVERAGE AND INTERVENTION VERIFICATION

All conditions cover the same 221 task IDs, with one valid record per task–condition. Trajectory and intervention-event audits confirm that each repair fired at its intended events in isolated task-scoped memory. The replication (Section 4.2) shares no execution state with the primary run; companion per-task labels reproduce the reported accuracies, marginals, interactions, and transitions.