DECISION INSTABILITY BEYOND RETRIEVAL METRICS IN CONSOLIDATED AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Language agents increasingly consolidate stored experiences to operate under bounded memory, yet consolidated memory is commonly evaluated by whether relevant information remains retrievable. We show that retrieval preservation does not imply preservation of the decisions supported by memory. Using a controlled intervention that holds target queries and outcome labels fixed while storing the same experiences in either original or consolidated form, we trace the effects of consolidation through the representation $\rightarrow$ retrieval $\rightarrow$ decision pathway. Aggregate retrieval success remains unchanged, with Hit@3 of 87.06% in both conditions, while downstream accuracy also remains nearly unchanged. Despite this aggregate stability, consolidation substantially reorganizes retrieved evidence and produces decision changes beyond ordinary repeat variability, with cross-condition disagreement exceeding same-memory disagreement by 5.20 percentage points. This instability concentrates on a reproducible subset of experiences. Vulnerable instances recur $9.43\times$ and $10.84\times$ above independence expectation under consumer-model and retriever substitution, respectively, even though the retrievers exhibit weak correspondence in their consolidation-induced changes. These results show that consolidation can preserve conventional retrieval and task-level performance while changing the evidence presented to the consumer and the decisions associated with particular experiences. Evaluating consolidated memory therefore requires considering not only whether relevant information remains retrievable, but whether the decision-relevant structure of stored experience is preserved.

1 INTRODUCTION

External memory allows language agents to retain past experiences for future reasoning and decision making (Packer et al., 2023; Park et al., 2023; Shinn et al., 2023). As these memories grow, retaining every experience in its original form becomes costly, motivating strategies that summarize, restructure, or consolidate accumulated information (Chhikara et al., 2025; Xu et al., 2025). Such transformations trade memory cost against preservation of information that may later matter (Colaco & Lahjouji, 2026; Kang et al., 2026). Yet preservation is commonly judged by whether relevant information remains retrievable, which does not establish that the evidence supporting downstream decisions has been preserved.

Consolidation can change which experiences are retrieved even when aggregate retrieval success remains similar. A downstream consumer may therefore receive different evidence while conventional retrieval metrics suggest stable memory performance. Aggregate accuracy can conceal the same effect when decision changes offset one another, creating a silent failure mode in which retrieval and task-level performance remain stable while particular decisions change.

We study this problem through a controlled intervention in retrieval-backed decision memory. The underlying incidents, target queries, and outcome labels remain fixed while each experience is stored either as its original occupational-incident narrative or as consolidated risk attributes. This isolates attribute-style consolidation along the representation $\rightarrow$ retrieval $\rightarrow$ decision pathway rather than evaluating a complete sequential agent.

This leads to three linked questions: (1) Does consolidation reorganize retrieved evidence despite similar aggregate retrieval success? (2) Do these changes produce decision instability beyond same-

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.

memory repeat variability? (3) Do vulnerable instances recur when the consumer model or retriever changes? The results reveal a progressive disconnect across these stages. Consolidation reorganizes retrieved evidence despite stable aggregate retrieval success, these changes propagate to decisions despite similar task-level accuracy, and the resulting instability concentrates on reproducibly vulnerable instances across component substitutions. These findings motivate decision preservation as a distinct criterion for evaluating attribute-style consolidation in retrieval-backed memory.

2 Related Work

Memory consolidation and evaluation: Language-agent memory has progressed from retaining and retrieving prior observations or reasoning traces (Park et al., 2023; Shinn et al., 2023; Packer et al., 2023) toward systems that actively restructure accumulated experience (Xu et al., 2025; Chhikara et al., 2025). Long-horizon memory benchmarks primarily evaluate whether stored information remains recoverable and useful over time (Wu et al., 2024; Maharana et al., 2024; Tan et al., 2025), while recent work on consolidation and compaction studies the trade-off between memory reduction and information preservation (Kang et al., 2026; Colaco & Lahjouji, 2026). These approaches characterize whether transformed memory retains useful information, but do not directly test whether it preserves the downstream decisions supported by retrieved experience.

From information preservation to decision preservation: This distinction is beginning to emerge outside retrieval-backed memory. A recent study shows that LLM-compressed financial contexts can remain plausible while altering decisions relative to the original source Lee et al. (2026). Within agent memory, retrieval and utilization have been distinguished (Yuan et al., 2026), alongside the fact that memory organization can affect preservation of fine-grained distinctions (Hu et al., 2026). Our setting connects these concerns by treating consolidation as an intervention upstream of both retrieval and decision making. We measure whether it changes retrieved evidence, whether resulting decision changes exceed same-memory repeat variability, and whether the same vulnerable instances persist under consumer-model and retriever substitutions.

Experimental setting: Occupational-incident narratives provide a controlled setting for this analysis because information-rich descriptions of individual events can be represented at different levels of detail while retaining fixed reference outcomes. Prior work has established injury-narrative classification (Marucci-Wellman et al., 2017; Nanda et al., 2020) and demonstrated the value of unstructured construction-accident narratives for extracting hazards, precursors, and outcomes (Tixier et al., 2016; Goh & Ubeynarayana, 2017; Baker et al., 2020). Structured risk attributes provide a compact representation of the same incident information (Ray et al., 2026). This pairing allows the underlying event and outcome to remain fixed while only its stored representation changes, isolating the effect of consolidation on retrieval and downstream decisions.

3 Problem Formulation

Let $E = \{e_j\}_{j=1}^{N}$ denote N stored experiences, where each e_j has an original narrative x_j, a consolidated representation c_j, and outcome label y_j. Here, consolidation transforms the full accident narrative into structured risk attributes extracted from the same incident (Ray et al., 2026). The attributes retain predefined incident characteristics relevant to the outcome while removing narrative wording and details outside the attribute schema. Thus, x_j and c_j represent the same event at different levels of detail. We construct a **Full memory bank** M_F, storing (x_j, y_j), and a **Consolidated memory bank** M_C, storing (c_j, y_j). Only the stored representation changes; the stored outcome remains fixed. Table 1 illustrates the intervention.

For target i, the query q_i is always its full original narrative and the target is excluded from candidate memory. A retriever R retrieves the top-k experiences from M_F or M_C, and consumer G uses q_i, the retrieved experiences, and their labels to produce a decision. Because q_i and y_j remain fixed, the intervention isolates the pathway: stored representation $\rightarrow$ retrieved evidence $\rightarrow$ downstream decision.

Table 1: Example of the controlled memory intervention. Superscripts map information in the Full stored experience to its corresponding attributes in the Consolidated representation.

Component	Stored content
Full x_j	At 11:28 a.m. on August 28, 2018, an employee was **on the roof installing the siding**[a], while working from a ladder. According to the other employee, the employee complained of **feeling dizzy**[b] and came off the roof and was **resting for a few minutes and then went back to work on the roof**[c]. About 20 minutes later he **fell off the roof**, resulting in **severe head, neck and back trauma**[d]. The employee was killed.
Consolidated c_j	Roof Work[a]; Feeling dizzy[b]; Resting for a few minutes and went back to work[c]; Severe head trauma[d]; Severe neck trauma[d]; Severe back trauma[d].
Outcome y_j	Fall from height

3.1 Representation and Retrieval Change

We first test how replacing x_j with c_j changes representation and retrieval for the same query. Four measures characterize these changes: (1) **representation fidelity**, the cosine similarity between embeddings of x_j and c_j (Salton & McGill, 1986); (2) **retrieval identity**, quantified by shared neighbors and Jaccard similarity (Jaccard, 1912); (3) **retrieval success**, measured by Hit@k, which indicates whether at least one retrieved experience carries the target outcome y_i (Cremonesi et al., 2010); and (4) **retrieval composition**, measured by the purity of the label (the proportion that carries y_i) and the diversity of the label (the number of different outcome labels). These measures separate changes in retrieved evidence from aggregate retrieval success.

3.2 Repeat-Controlled Decision Instability

Following the retrieval analysis, let $Y_{i,F,r}$ and $Y_{i,C,r}$ denote decisions for target i under Full and Consolidated memory, respectively, where $r \in \{1, 2\}$ indexes two independent repeats. Two repeats provide a minimal repeat-controlled design for distinguishing cross-condition disagreement from same-memory variability, consistent with recent recommendations for repeated LLM evaluation (Alvarado Gonzalez et al., 2025). We define cross-condition disagreement as:

$$D_i^{\text{cross}} = \frac{1}{2} \sum_{r=1}^{2} \mathbf{1}\left[Y_{i,F,r} \neq Y_{i,C,r}\right] \tag{1}$$

and same-memory disagreement as:

$$D_i^{\text{same}} = \frac{1}{2} \left(\mathbf{1}[Y_{i,F,1} \neq Y_{i,F,2}] + \mathbf{1}[Y_{i,C,1} \neq Y_{i,C,2}]\right) \tag{2}$$

where $\mathbf{1}[\cdot]$ is the indicator function. The target-level repeat-controlled contrast (d_i) and its dataset-level mean (Δ) are given as below:

$$d_i = D_i^{\text{cross}} - D_i^{\text{same}} \tag{3}$$

$$\Delta = \frac{1}{N} \sum_{i=1}^{N} d_i \tag{4}$$

A positive Δ indicates excess Full-versus-Consolidated disagreement beyond same-memory repeat variability. Thus, Δ measures decision instability associated with the memory intervention rather than classification degradation; a change may be correct-to-incorrect, incorrect-to-correct, or between two incorrect outcomes.

3.3 Instance-Level Localization

Beyond measuring the overall magnitude of decision instability, we test whether that instability is diffuse across targets or concentrated on a reproducibly vulnerable subset. For system configuration A, let V_A denote the reproducibly vulnerable targets, defined as those whose Full and Consolidated decisions differ in both independent repeats:

$$i \in V_A \iff \left(Y^A_{i,F,1} \neq Y^A_{i,C,1}\right) \wedge \left(Y^A_{i,F,2} \neq Y^A_{i,C,2}\right) \tag{5}$$

Persistence of vulnerability is evaluated under consumer-model substitution with the retriever fixed and retriever substitution with the consumer fixed. Let A and B denote the two system configurations being compared, with corresponding vulnerable sets V_A and V_B. Their observed overlap is:

$$O_{A,B} = |V_A \cap V_B| \tag{6}$$

while the overlap expected under independent localization is:

$$E_{A,B} = \frac{|V_A||V_B|}{N} \tag{7}$$

Therefore, comparing $O_{A,B}$ with $E_{A,B}$ determines whether the same targets recur as vulnerable across component substitutions more often than expected under independent localization.

4 Experimental Design

The experimental design follows the pathway defined in Section 3. We establish a controlled testbed in which the same experiences are available in two memory representations, then examine whether consolidation changes the representation itself, reorganizes retrieved evidence, and propagates to downstream decisions. Consumer-model and retriever substitutions further test whether instance-level vulnerability persists when individual system components change. Figure 1 summarizes the design.

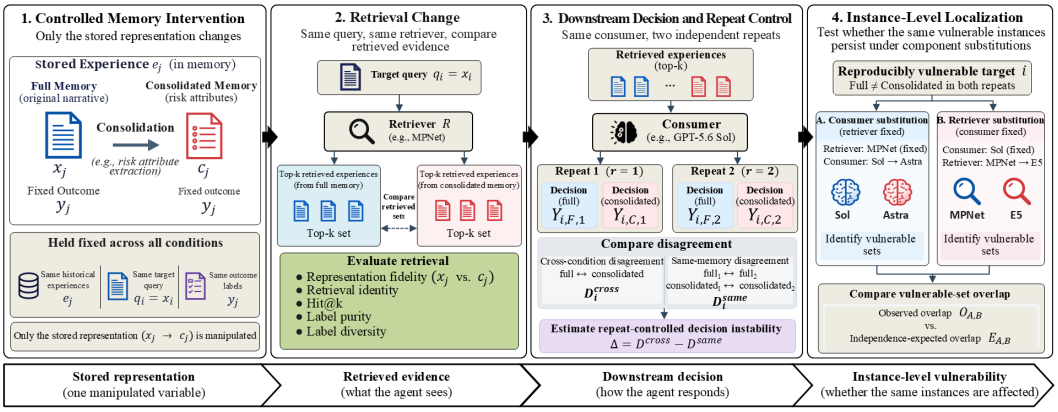


Figure 1: Overview of the controlled memory intervention and evaluation framework.

4.1 Dataset and Paired Memory Representations

The study uses occupational-incident records collected from the U.S. Occupational Safety and Health Administration Investigation Summaries database (OSHA). The initial dataset contained 400 records. After removing duplicates, cases with insufficient textual content, and non-semantic artifacts, 394 records with complete paired representations remained. Each record contains an original narrative x_j, a consolidated risk-attribute representation c_j, and reference outcome y_j. The

downstream task is nine-class accident classification across falls from height, struck-by, caught-in/between, electrocution, excavation, heat, medical, miscellaneous, and unknown. This pairing changes the stored representation while preserving the underlying incident and outcome.

4.2 Representation Analysis

The first analysis examines how much an experience changes under consolidation. We quantify both the reduction in representation length and the semantic correspondence between each original narrative x_j and its consolidated counterpart c_j. Both representations are embedded with `all-mpnet-base-v2` (Reimers & Gurevych, 2019; Song et al., 2020), and representation fidelity is measured by cosine similarity between their normalized embeddings. Thereafter, we examine whether fidelity varies with the degree of length reduction or systematically across outcome categories. Spearman rank correlation (Spearman, 1904) tests the relationship between length reduction and fidelity, while the Kruskal–Wallis test (Kruskal & Wallis, 1952) evaluates differences across the nine categories. Between and within-category variation is also compared to determine whether fidelity differences arise primarily across categories or individual experiences.

4.3 Retrieval Protocol and Retriever Substitution

The next analysis tests whether representation change alters the evidence retrieved from memory. For each target i, the same original narrative q_i queries both memory banks, with the target's own record excluded so that retrieval cannot return the experience being classified. The query and candidate experiences therefore remain fixed while their stored representations differ. Experiences are ranked by cosine similarity between normalized embeddings. The primary retriever uses `all-mpnet-base-v2`, with $k = 3$ matching the three retrieved experiences supplied to the downstream consumer. Additional depths $k \in \{1, 5, 10\}$ test robustness across retrieval depth. Retrieval is evaluated using the success, identity, purity, and diversity measures defined in Section 3.1. To determine whether these effects depend on the embedding model, we repeat the analysis with `bge-base-en-v1.5` (Xiao et al., 2023), `e5-base-v2` (Wang et al., 2022), and `all-MiniLM-L6-v2` (Wang et al., 2020). For downstream retriever substitution, we select the alternative whose consolidation-induced changes differ most from the primary retriever while retaining comparable aggregate retrieval success.

4.4 Consumer Models and Decision Protocol

The decision analysis asks whether changes in retrieved evidence propagate to the downstream consumer. For each target i, the consumer receives q_i, the three retrieved experiences with their fixed historical outcome labels, and the same nine candidate categories. The prompt, category definitions, evidence format, and output constraints remain identical across memory conditions. The model-agnostic protocol uses GPT-5.6 Sol as the primary consumer (OpenAI, 2026b) and GPT-6 Astra for consumer substitution (OpenAI, 2026a). Each memory condition is evaluated in two independent repeats, producing $Y_{i,F,1}$, $Y_{i,F,2}$, $Y_{i,C,1}$, and $Y_{i,C,2}$. Both consumers use the same medium reasoning setting, and Full and Consolidated presentation order is randomized per target using a fixed reproducible procedure. Responses are normalized only for superficial formatting differences before comparison with the nine permitted category labels. Responses that cannot be mapped to a permitted category are retained as invalid rather than manually corrected.

4.5 Statistical Analyses

Representation fidelity is evaluated using Spearman rank correlation and the Kruskal–Wallis test. Retrieval changes are tested using McNemar's test (McNemar, 1947) for paired Hit@3 outcomes and Wilcoxon signed-rank tests (Wilcoxon, 1945) for purity and diversity. Decision instability is quantified by Δ from Section 3.2, with a 95% percentile confidence interval from 10,000 target-level paired bootstrap resamples (Efron & Tibshirani, 1993). Cross-component localization is evaluated using one-sided Fisher exact tests (Fisher, 1922), odds ratios, and 95% confidence intervals for overlap among reproducibly vulnerable targets.

5 RESULTS

5.1 REPRESENTATION EFFECTS OF CONSOLIDATION

Consolidation reduced mean representation length from 86.25 to 20.79 words, corresponding to a 75.9% reduction in aggregate word count and a mean case-level reduction of 69.66% (Fig. 2a). Despite this reduction, the consolidated representations retained moderate semantic correspondence with their original narratives, with mean cosine similarity of 0.632 ($SD = 0.112$) and values ranging from 0.240 to 0.897 (Fig. 2b).

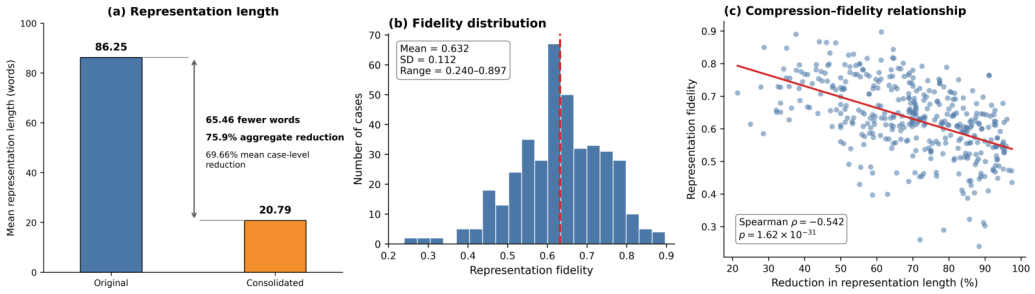


Figure 2: Representation effects of memory consolidation.

Greater reduction in representation length was associated with lower semantic fidelity (Spearman $\rho = -0.542$, $p = 1.62 \times 10^{-31}$; Fig. 2c). Fidelity did not differ significantly across the nine accident categories (Kruskal-Wallis $H = 5.86$, $p = .663$; Fig. 3). Only 1.48% of the observed variation occurred between categories, compared with 98.52% within categories. Representation change therefore varied primarily across individual experiences rather than systematically across outcome classes.

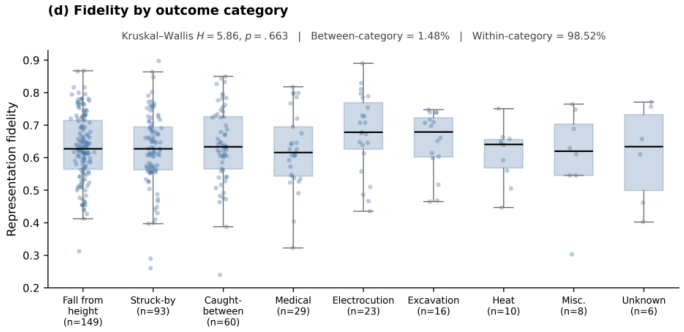


Figure 3: Representation fidelity by outcome category.

5.2 RETRIEVAL EFFECTS OF CONSOLIDATION

At the primary retrieval depth of $k = 3$, aggregate retrieval success was unchanged after consolidation, with Hit@3 of 87.06% under both Full and Consolidated memory. This aggregate stability concealed substantial target-level change. Twenty-seven targets gained a successful retrieval and 27 lost one, yielding no significant paired change in Hit@3 (McNemar $\chi^2 = 0.02$, $p = .892$). The identities of the retrieved experiences changed much more strongly. Full and Consolidated memory shared an average of only 0.312 of three retrieved neighbors per target, with mean Jaccard similarity of 0.066. For 72.34% of targets, all three retrieved experiences were replaced after consolidation. The composition of the retrieved evidence changed accordingly. Mean label purity decreased from 66.92% to 57.70%, while mean label diversity increased from 1.57 to 1.86, with both changes significant at $p < 10^{-11}$.

The same pattern persisted across retrieval depths (Table 2). Hit@k remained similar between memory conditions, while differences in purity and diversity became more pronounced as retrieval depth increased. Thus, preservation of aggregate retrieval success did not imply preservation of either retrieval identity or evidence composition.

Table 2: Retrieval effects across retrieval depth under MPNet.

k	Full Hit@k (%)	Consolidated Hit@k (%)	Purity difference (pp)	Diversity difference	Jaccard
1	71.32	67.51	+3.81	0.000	0.084
3	87.06	87.06	+9.22	+0.289	0.066
5	92.64	91.37	+9.80	+0.388	0.080
10	97.21	95.43	+11.85	+0.607	0.102

The effect was not limited to the primary embedding model. Across `bge-base-en-v1.5`, `e5-base-v2`, and `all-MiniLM-L6-v2`, Full and Consolidated memory again produced similar Hit@3 while differing substantially in retrieved-experience identity and composition (Table 3). Complete replacement of the top-three retrieved experiences ranged from 58.12% to 75.89% across the alternative retrievers.

Table 3: Retrieval effects across embedding models at $k = 3$.

	Micro Hit@3 (%)		Purity (%)			Top-3 Replaced (%)
Retriever	Full	Consolidated	Full	Consolidated	Jaccard	
MPNet	87.06	87.06	66.92	57.70	0.066	72.34
BGE	85.53	84.26	64.04	57.70	0.106	58.12
E5	86.29	85.03	62.78	60.24	0.055	75.89
MiniLM	83.50	83.76	63.62	61.84	0.104	58.63

Among the alternative retrievers, `e5-base-v2` showed the weakest target-level correspondence with the primary retriever in consolidation-induced retrieval changes while retaining comparable aggregate retrieval success. It was therefore carried forward as the contrasting retriever in the component-substitution analysis of Section 5.4.

5.3 Repeat-Controlled Decision Instability

The retrieval changes in Section 5.2 did not produce a corresponding decline in aggregate task performance. Mean classification accuracy was 87.82% under Full memory and 87.18% under Consolidated memory, a difference of only -0.63 percentage points. Aggregate accuracy therefore suggested that the two memory conditions behaved similarly.

Individual decisions revealed a different pattern. Full and Consolidated memory disagreed on 31 of 394 targets in the first repeat (7.87%) and 29 of 394 in the second (7.36%), yielding mean cross-condition disagreement of 7.61%. In comparison, repeated evaluation of the same memory condition produced disagreement rates of 2.03% for Full memory and 2.79% for Consolidated memory, with mean same-memory disagreement of 2.41%. The resulting repeat-controlled contrast was $\Delta = +5.20$ percentage points, with a 95% bootstrap confidence interval of $[+3.05, +7.49]$. Full-versus-Consolidated decisions therefore differed substantially more often than expected from repeat variability alone, despite nearly unchanged aggregate accuracy.

The stability of aggregate accuracy partly reflected offsetting changes at the instance level. Across both repeats, 30 Full-to-Consolidated disagreements changed a correct decision to an incorrect one, while 25 changed an incorrect decision to a correct one. The remaining disagreements moved between incorrect categories. Aggregate gains and losses could therefore offset one another while the decisions associated with particular experiences changed.

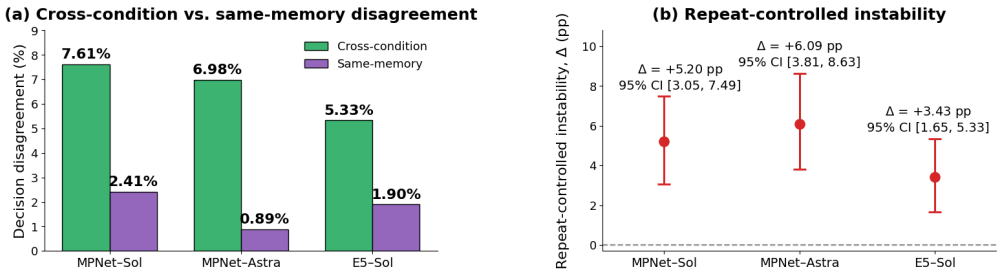


Figure 4: Repeat-controlled decision instability under Full and Consolidated memory.

5.4 Cross-Component Localization of Decision Vulnerability

The excess instability identified in Section 5.3 could reflect broadly distributed changes or a smaller set of consistently susceptible targets. We therefore examine whether reproducibly vulnerable targets recur when either the consumer model or retriever is replaced.

5.4.1 Consumer-Model Substitution

Under the primary MPNet–Sol configuration, 25 of 394 targets were reproducibly vulnerable. Replacing Sol with Astra while holding the retriever fixed yielded $\Delta = +6.09$ percentage points, with cross-condition disagreement of 6.98% compared with same-memory disagreement of 0.89% and a 95% confidence interval of $[+3.81, +8.63]$. The Astra configuration also identified 25 reproducibly vulnerable targets. Fifteen targets were shared across consumers, corresponding to 60% of each vulnerable set. Only 1.59 shared targets would be expected under independent localization, making the observed overlap $9.43\times$ expectation (OR$= 53.85$, 95% CI $[19.47, 148.95]$, one-sided Fisher $p < 10^{-8}$; Fig. 5). All 15 shared targets exhibited the same Full-to-Consolidated decision transition under both consumers and both repeats.

5.4.2 Retriever Substitution

Replacing MPNet with E5 while holding Sol fixed yielded $\Delta = +3.43$ percentage points, with cross-condition disagreement of 5.33% compared with same-memory disagreement of 1.90% and a 95% confidence interval of $[+1.65, +5.33]$. Sixteen targets were reproducibly vulnerable under the E5 configuration. Eleven of these targets were also vulnerable under MPNet–Sol, corresponding to 68.8% of the E5 vulnerable set. Only 1.02 shared targets would be expected under independence, making the observed overlap $10.84\times$ expectation (OR$= 57.20$, 95% CI $[17.50, 186.94]$, one-sided Fisher $p = 2.12 \times 10^{-11}$; Fig. 5). Exact Full-to-Consolidated transitions matched for 6 of the 11 shared targets. Vulnerability therefore persisted more strongly than the particular decision transition through which it appeared. Across both substitutions, vulnerable targets overlapped far beyond independence expectation. Exact transition agreement was complete across consumers but lower across retrievers, with 15 of 15 and 6 of 11 shared targets, respectively. The identity of vulnerable targets was therefore more stable across component changes than the specific form of the resulting decision change.

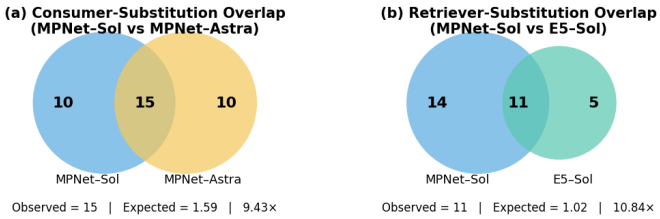


Figure 5: Cross-component localization of reproducible decision vulnerability.

Vulnerability also concentrated by outcome category (Table 4). Caught-between accounted for 14 of the 25 reproducibly vulnerable targets under Sol and 12 of the 15 targets shared across consumers, corresponding to a vulnerability rate of 23.3% compared with 0.7% for the larger fall-from-height category. Across both repeats, caught-between targets produced 30 individual Full-to-Consolidated disagreement events, of which 28 involved transitions between caught-between and struck-by. The higher rates for miscellaneous and unknown each reflect only one vulnerable target and are not interpreted further. Instability therefore concentrated primarily near the caught-between and struck-by boundary rather than uniformly across outcome categories.

Table 4: Reproducible decision vulnerability by outcome category.

Outcome	n	Sol	Rate (%)	Astra	Shared
Caught-between	**60**	**14**	**23.3**	**16**	**12**
Struck-by	93	4	4.3	5	2
Medical	29	2	6.9	0	0
Excavation	16	1	6.2	0	0
Electrocution	23	1	4.3	0	0
Miscellaneous	8	1	12.5	0	0
Unknown	6	1	16.7	2	1
Fall from height	149	1	0.7	2	0
Heat related	10	0	0.0	0	0

6 Discussion

Retrieval success versus decision preservation: The results expose a gap between preserving aggregate memory performance and preserving the decisions supported by memory. Evaluating attribute-style consolidation therefore requires distinguishing whether information remains accessible from whether it continues to support the same downstream decisions.

Localized vulnerability across system components: Decision instability concentrated on particular experiences that recurred as vulnerable when either the consumer or retriever changed. This persistence suggests that susceptibility is not solely an artifact of one system component. Exact decision transitions were more stable across consumers than retrievers, indicating that *which* experiences are vulnerable can persist even when *how* their decisions change does not. Vulnerability further concentrated near the caught-between and struck-by boundary, where overlapping impact and crushing mechanisms can distinguish closely related hazards (Occupational Safety and Health Administration, 2006). These findings motivate selective preservation of detail for experiences whose decision-relevant distinctions are most susceptible to consolidation.

Limitations: The study evaluates one corpus, nine-class task, and attribute-style consolidation scheme, so broader generalization remains to be established. Consumer and retriever substitutions were evaluated separately rather than factorially, two repeats provide a minimal estimate of same-memory variability, and the small vulnerable subset limits analysis of the characteristics underlying susceptibility. The proprietary consumer models may also change over time, limiting exact replication across model versions.

7 Conclusion

This study shows that preserving aggregate memory performance is not the same as preserving the decisions supported by memory. Through a controlled intervention spanning representation, retrieval, repeated downstream decisions, and component substitution, we establish three findings: (1) consolidation substantially reorganizes retrieved evidence even when aggregate retrieval success remains nearly unchanged; (2) these changes produce decision instability beyond same-memory repeat variability ($\Delta = +5.20$ percentage points in the primary configuration); and (3) this instability concentrates on reproducibly vulnerable experiences that recur far beyond independence expectation under consumer and retriever substitution. These results shift the evaluation of memory consolidation from asking whether compressed memories remain retrievable to whether they preserve the evidence structure needed for stable downstream reasoning. Decision preservation should therefore be treated as a distinct objective when designing and evaluating memory systems.

AI USE STATEMENT

Generative AI tools were used to assist with polishing the writing and improving clarity. All scientific content, analyses, interpretations, and conclusions were developed and verified by the authors.

REPRODUCIBILITY STATEMENT

The experimental protocol, evaluation measures, statistical analyses, and model configurations are described in the paper. All experiments use fixed paired representations, target queries, and outcome labels, with self-retrieval excluded as specified in the methodology. Randomized procedures use fixed reproducible settings where applicable. Code and derived experimental results will be made available upon publication.

REFERENCES

Miguel Angel Alvarado Gonzalez, Michelle Bruno Hernandez, Miguel Angel Peñaloza Perez, Bruno Lopez Orozco, Jesus Tadeo Cruz Soto, and Sandra Malagon. Do repetitions matter? strengthening reliability in llm evaluations, 2025. URL https://arxiv.org/abs/2509.24086.

Henrietta Baker, Matthew R. Hallowell, and Antoine J.P. Tixier. Automatically learning construction injury precursors from text. *Automation in Construction*, 118:103145, 2020. doi: 10.1016/J.AUTCON.2020.103145. URL https://www.sciencedirect.com/science/article/abs/pii/S0926580519310209?via%3Dihub.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *Frontiers in Artificial Intelligence and Applications*, 413:2993–3000, 2025. doi: 10.3233/FAIA251160. URL https://arxiv.org/pdf/2504.19413.

Ashwin Gerard Colaco and Nada Lahjouji. What to keep, what to forget: A rate–distortion view of memory compaction in llms and agents. *arXiv preprint arXiv:2607.08032*, 1, 2026. URL https://arxiv.org/pdf/2607.08032.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys)*, pp. 39–46, 2010. doi: 10.1145/1864708.1864721.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, 1993. doi: 10.1201/9780429246593.

R. A. Fisher. On the interpretation of 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85:87, 1922. doi: 10.2307/2340521. URL https://www.jstor.org/stable/2340521.

Yang Miang Goh and C. U. Ubeynarayana. Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis & Prevention*, 108:122–130, 2017. doi: 10.1016/J.AAP.2017.08.026. URL https://www.sciencedirect.com/science/article/pii/S0001457517303068?via%3Dihub.

Zhanghao Hu, Qinglin Zhu, Runcong Zhao, Di Liang, Hanqi Yan, Yulan He, and Lin Gui. Beyond rag for agent memory: Retrieval by decoupling and aggregation. *arXiv preprint arXiv:2602.02007*, 2026. URL https://arxiv.org/pdf/2602.02007.

Paul Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11:37–50, 1912. doi: 10.1111/J.1469-8137.1912.TB05611.X.

Qingcan Kang, Mingyang Liu, Shixiong Kai, Kaichao Liang, Zhentao Tang, Yuqi Cui, Tao Zhong, and Mingxuan Yuan. Retain or consolidate? budget-dependent operator selection for language agent memory. *arXiv preprint arXiv:2607.17545*, 2026. URL https://arxiv.org/pdf/2607.17545.

William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47:583–621, 1952. doi: 10.1080/01621459.1952.10483441. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441.

Hoyoung Lee, Suhwan Park, Seunghan Lee, Jun Seo, Jaehoon Lee, Sungdong Yoo, Minjae Kim, Cheolwon Na, Zhangyang Wang, V Zach Golkhou, Minkyu Kim, Sotirios Sabanis, Alejandro Lopez-Lira, Dhagash Mehta, Soonyoung Lee, Chanyeol Choi, Wonbin Ahn, and Yongjae Lee. When summaries distort decisions: Information fidelity in llm-compressed financial analysis. *arXiv preprint arXiv:2606.29251*, 2026. URL https://arxiv.org/pdf/2606.29251.

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 (ACL)*, pp. 13851–13870, 2024. doi: 10.18653/V1/2024.ACL-LONG.747.

Helen R. Marucci-Wellman, Helen L. Corns, and Mark R. Lehto. Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review. *Accident Analysis & Prevention*, 98:359–371, 2017. doi: 10.1016/J.AAP.2016.10.014. URL https://www.sciencedirect.com/science/article/pii/S000145751630375X?via%3Dihub.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947. doi: 10.1007/BF02295996. URL https://link.springer.com/article/10.1007/BF02295996.

Gaurav Nanda, Kirsten Vallmuur, and Mark Lehto. Intelligent human-machine approaches for assigning groups of injury codes to accident narratives. *Safety Science*, 125:104585, 2020. doi: 10.1016/J.SSCI.2019.104585. URL https://www.sciencedirect.com/science/article/abs/pii/S0925753519321964?via%3Dihub.

Occupational Safety and Health Administration. Construction hazards (top four) quickcard. U.S. Department of Labor, 2006. URL https://www.osha.gov/sites/default/files/publications/CONSTRUCTION_HAZARDS_QC.pdf. OSHA 3216-06N. Accessed September 19, 2026.

OpenAI. Gpt-6 astra system card - openai deployment safety hub. OpenAI Deployment Safety Hub, 2026a. URL https://deploymentsafety.openai.com/gpt-6-astra.

OpenAI. Gpt-5.6 preview system card - openai deployment safety hub. OpenAI Deployment Safety Hub, 2026b. URL https://deploymentsafety.openai.com/gpt-5-6-preview.

OSHA. Investigation summaries. U.S. Department of Labor. URL https://www.osha.gov/ords/imis/accidentsearch.html. Accessed September 23, 2026.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/pdf/2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi: 10.1145/3586183.3606763.

Unmesa Ray, JeeWoong Park, and Malak Mahdy. Exploring semantic compression in transformer-based safety intelligence. In *Construction Research Congress 2026: Advanced Technologies, Artificial Intelligence, and Data Analytics in Construction*, pp. 126–136. American Society of Civil Engineers (ASCE), 2026. doi: 10.1061/9780784486962.013.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Gerard Salton and Michael J McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15:72, 1904. doi: 10.2307/1412159.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, 2025. doi: 10.18653/V1/2025.FINDINGS-ACL.989.

Antoine J.-P. Tixier, Matthew R. Hallowell, Balaji Rajagopalan, and Dean Bowman. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction*, 62:45–56, 2016. doi: 10.1016/j.autcon.2015.11.001.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. URL https://arxiv.org/pdf/2212.03533.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80, 1945. doi: 10.2307/3001968. URL https://www.jstor.org/stable/3001968.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025. URL https://arxiv.org/pdf/2502.12110.

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