TRACE: GOVERNING MEMORY VALIDITY IN EVOLVING MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent memory lets language-model agents carry information across long-running collaborations, but leaves a lifecycle question open: what may a returning agent still act on once the shared state has changed? A memory can be correctly retrieved, relevant to the current task, and faithful to its source, and nonetheless be inadmissible for action: an itinerary saved before a pause still names the hotel the team has since replaced (Figure 1). We formalize this as *temporal memory admission* and present TRACE, a training-free layer that treats re-entry as an eligibility decision rather than a storage or retrieval operation, reconciling a departure checkpoint against absence-period updates, resolving explicit and implicit invalidation, and releasing a bounded Return View only when it covers the returning role's open obligations. We evaluate TRACE under three actor models on Memora, STALE Type II, and a derived ManBench-Return setting, each recast as return episodes: one agent departs, four teammates change the shared state, and the agent rejoins. What separates methods is not overall accuracy but whether one can retain valid memory and reject stale memory at once, and no single-policy baseline can: RESTORE (reinstate the departure checkpoint in full) admits stale state, RESET (start the return from an empty memory) discards valid state, each bottoming out at 0% on one of the two. TRACE is the only method high on both, reaching 92.6–98.3% valid-information availability with 98.4–99.5% invalid-information rejection on ManBench-Return, within 3.8 points of the best baseline's overall accuracy. On STALE Type II it improves Overall over the strongest comparison policy by 22.3 (Qwen), 18.5 (Gemini), and 27.5 (DeepSeek) points at roughly $2.3\times$ their tokens, while a write-time consolidation pipeline is more accurate still at $3.99\times$ TRACE's.

1 INTRODUCTION

Long-running multi-agent systems do not keep a fixed roster: agents pause, are reassigned, and rejoin a workspace the remaining members have continued to change. Figure 1 introduces the resulting difficulty. Agent A1 departs with a coherent itinerary; during its absence the team replaces the hotel, the railway moves the departure time, and the pickup is left as an open obligation. A1's stored itinerary is still topically relevant and was accurate when recorded, yet parts of it can no longer be acted on. Faithful retrieval does not establish current eligibility for action. The example poses the question this paper studies: at an authenticated return, which parts of a retained memory remain admissible, and on what basis is that decided?

Persistent memory and retrieval systems are designed to reuse historical information (Park et al., 2023; Shinn et al., 2023; Lewis et al., 2020; Packer et al., 2023; Chhikara et al., 2025; Maharana et al., 2024; Wu et al., 2025). Memora evaluates selective forgetting (Uddin et al., 2026), while temporal knowledge graphs and transactional belief protocols track facts that evolve (Rasmussen et al., 2025; Li et al., 2026b). Each improves what a store retains or retrieves. Recent memory managers add reflection and temporal organization (Zhang et al., 2024; Kang et al., 2025; Tan et al., 2025b), and newer benchmarks separate long-term recall from continual repair (Tan et al., 2025a; Hu et al., 2026; Jia et al., 2025); none decides what a particular agent may act on when it rejoins a changed team. Nor does the failure require an adversary (Chen et al., 2024; Xiong et al., 2026): ordinary task evolution can strand records that were trustworthy when written. Two controlled findings confirm the gap. On STALE, LightMem fails implicit policy adaptation in 78.6% of cases in which the updated evidence had already been retrieved (Chao et al., 2026). On ManBench, Qwen3-

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.

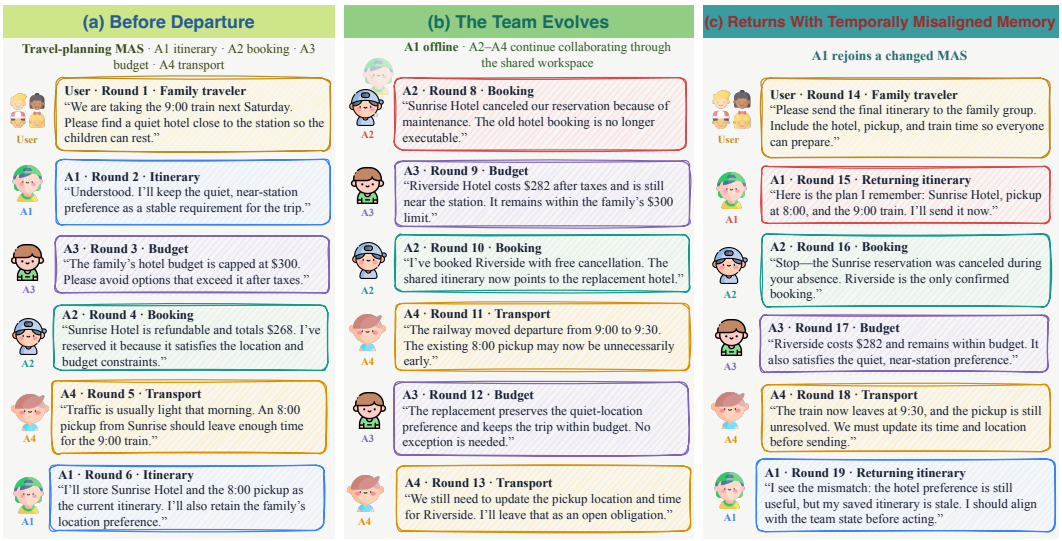


Figure 1: The agent-return dilemma under context evolution. A1 leaves with a coherent travel plan. During its absence, the team replaces the hotel booking, changes the train schedule, and leaves the pickup time unresolved. When A1 returns, its stored itinerary is still relevant and was correct when recorded, yet parts of it are no longer admissible for action.

235B answers 56.85% of initially correct questions incorrectly after misleading group discussion and memory consolidation (Xu et al., 2026).

Prior work separates storage and retrieval, belief maintenance, and social propagation (Doyle, 1979; Zhong et al., 2024; Xu et al., 2026). The return event remains less specified: a record may be retrievable and relevant, yet fail the returning role's current obligations.

We formalize this check as *temporal memory admission*: the returning agent reassesses retained memory against current state, identity, task purpose, and open obligations. Recency alone is insufficient: old constraints may still bind, while recent records may lack authorization, provenance, or required state.

TRACE (**T**emporal **R**eturn **A**dmission under **C**ontext **E**volution) enforces this contract at return as a training-free layer between a memory backend and the agent. It reconciles the departure checkpoint with the absence-period delta, resolves explicit invalidation, and verifies candidate implicit dependencies. Surviving records pass authorization, temporal, applicability, and provenance gates before a budgeted selector covers open obligations. The signed Return View is released only after validation and coverage, and private experience is readmitted only with its bound work-state item.

We evaluate TRACE under a common five-agent return protocol on three settings that isolate distinct forms of state change: explicit replacement in Memora, implicit contextual invalidation in STALE Type II, and propagated conflict in ManBench-Return, which we derive from ManBench conflict scenarios and report separately from the original benchmark. Each setting is run under three actor models, and the protocol also delimits where the layer is unnecessary.

On Memora, where replacement information is explicit, ordinary storage semantics already suffice: TRACE offers no consistent gain and trails a RESET policy on overall accuracy. On STALE Type II it raises overall accuracy over the strongest comparison policy by 22.3 points with Qwen, 18.5 with Gemini, and 27.5 with DeepSeek, and with Qwen it raises Lifecycle Success from 18.0% to 39.0%, at roughly 2.3 times the tokens per episode of the cheaper lifecycle policies; the write-time consolidation baseline CUPMEM remains more accurate still on every reported score, at 3.99 times TRACE's token total. On ManBench-Return it reaches 92.6–98.3% valid-information availability with 98.4–99.5% invalid-information rejection, within 3.8 points of the best single-policy baseline's overall accuracy, whereas RESET rejects valid and stale state alike and RESTORE preserves both. Return-time admission therefore trades bounded overhead for joint preservation and rejection that neither policy provides.

Our contributions are threefold:

- **Formalizing temporal memory admission.** We formalize the conditions under which retained memories remain admissible for action when an agent rejoins an evolving team. This formulation separates historical correctness and query relevance from current admissibility under explicit and implicit state changes.
- **Auditable admission at agent return.** We introduce TRACE, a training-free layer that combines source-verified implicit invalidation with item-level eligibility checks and obligation-aware memory selection. TRACE assembles an auditable Return View, releasing it only when validation and obligation coverage succeed, while readmitting private experience only if its bound work-state item remains eligible.
- **Empirical benefits and applicability.** Across three return benchmarks and three actor models, we show that TRACE jointly preserves valid information and rejects stale information, with gains concentrated in implicit and propagated invalidation and limited benefit under explicit updates. Four mechanism ablations, evaluations with 4–32 active agents, and tests across four memory backends further characterize its behavior and applicability.

2 Temporal Memory Admission

2.1 Lifecycle boundary

Let agent a depart from team $\mathcal{A}$ at time t_d and return at $t_r > t_d$, preserving its principal and role but starting a fresh workload instance. Over the absence interval the shared work state evolves from C_{t_d} to C_{t_r} according to a recorded delta $\Delta_{(t_d,t_r]}$, while a retains a pre-departure bank M_a^- of shared work-state records and private episodic experience. A return mechanism must expose enough of M_a^- to preserve useful continuity without making superseded or unauthorized premises executable again.

A work-state item m carries content, a source receipt, an epoch and validity interval, principal and role scope, task purpose, dependency links, and supersession status, where an epoch indexes one recorded revision of the shared state. Relevance to a fresh query q is not sufficient for admissibility: there exist m and q such that

$$\mathrm{Rel}(m, q) = 1 \quad \text{and} \quad \mathrm{Adm}\big(m \mid a,\, C_{t_r},\, \Delta_{(t_d,t_r]},\, O_{t_r}\big) = 0, \tag{1}$$

where O_{t_r} denotes the obligations of the returning role. This second predicate, which turns on authority and state dependence rather than on similarity or recency, is the decision studied in this paper.

2.2 Two temporal failure modes

A valid item loses eligibility in two ways. Explicit, linkable invalidation supplies a machine-readable relation: a revocation, a replacement edge, a closed validity interval, or a same-identity update. A conventional store resolves it whenever the relation is preserved (Rasmussen et al., 2025; Yadav, 2026). Implicit invalidation (II) instead changes a premise the item depends on without declaring the item false, as when a changed destination invalidates a train recommendation that no later event mentions. Resolving II requires reasoning over the item's current use together with the absence-period delta, the regime that STALE Type II isolates (Chao et al., 2026).

Source quality and cross-agent propagation cut across this distinction rather than extending it. A stale item may carry an authentic receipt and a current item an unverified one, and repeated transmission can make wrong state appear well supported (Xu et al., 2026) without creating a new temporal failure mode. We therefore treat provenance, state dependence, and temporal validity as separate gates.

2.3 Desired contract

A correct return mechanism should preserve the valid information the fresh workload needs, reject items that no longer support current action, and fail closed when the available evidence cannot cover critical obligations. The first two requirements define Valid Information Availability and Invalid Information Rejection (Section 4.3); the third is enforced by the coverage gate (Section 3.4). Together

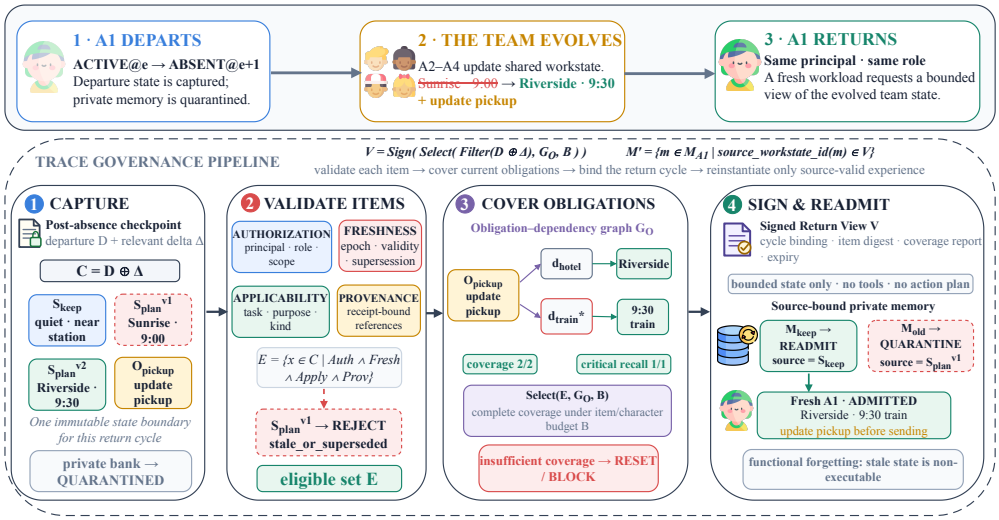


Figure 2: TRACE at an authenticated return boundary. The system captures one immutable departure-and-delta view, validates candidate items, selects a budgeted set that covers current obligations, and signs a Return View. Private experience is readmitted only when its bound source state survives.

they exclude both unconditional restoration and unconditional reset, and they motivate a bounded, inspectable artifact at the lifecycle boundary in place of unrestricted access to M_a^-, following the fail-safe-default principle of access-control design (Saltzer & Schroeder, 1975).

3 TRACE

Figure 2 summarizes the four stages of the layer. TRACE establishes one immutable state boundary per return cycle and governs which records become visible to the fresh workload. Appendix C gives formal definitions and complete return examples (Appendices C.6 and C.7).

3.1 Capture a return-cycle checkpoint

At departure, TRACE records a checkpoint D_a: the portion of C_{t_d} visible to a under its principal, role, and task scope, together with the source receipts and state epoch that make it auditable. The private episodic bank is quarantined rather than copied forward. At return, TRACE binds D_a to the relevant absence delta and materializes the post-absence checkpoint $C = D_a \oplus \Delta$, a single state against which all subsequent decisions are made.

The candidate pool $\mathcal{P}$ drawn from C contains verified facts, progress frontiers, open obligations, and role and interface constraints. Private experiences stay in the agent's own bank, each bound to the work-state item that sourced it. Candidate construction is deliberately broader than final admission: omitting a useful record is irreversible at this stage, whereas the later gates can still reject a false positive.

3.2 Resolve explicit and implicit invalidation

TRACE first applies direct lifecycle evidence, removing an item whose source epoch precedes an incompatible target epoch, whose validity interval does not cover t_r, or which a recorded supersession or revocation link targets. Updates that carry no such link require the two-stage test that distinguishes II. A high-recall scan proposes absence-period events that may change a premise of the candidate's current use. A grounded verifier then receives each source-bound pair and decides whether the later event invalidates that use in the returning task, deciding on checkpoint evidence alone rather than on benchmark annotations. Separating proposal from verification confines the expensive pairwise judgment to plausible dependencies and keeps topical similarity from being read as contradiction.

3.3 VALIDATE ELIGIBLE ITEMS

Every surviving candidate passes four item-level gates:

$$E = \{\, m \in \mathcal{P} : A(m) \wedge T(m) \wedge U(m) \wedge P(m) \,\}, \tag{2}$$

where A checks authorization for the principal, role, and scope; T checks epoch, validity, and supersession; U checks applicability to the task, purpose, and item kind; and P requires bound receipts for referenced source state. Authorization and applicability follow an attribute-based policy view (Hu et al., 2014), while P records source derivation (Moreau & Missier, 2013) without certifying that the source statement is true. Each item must also fit an individual size budget. Keeping the predicates separate yields a reason-coded audit trail and prevents a recent but unauthorized record from passing as fresh.

3.4 COVER OBLIGATIONS UNDER A CONTEXT BUDGET

Admission is a set decision as well as an item decision. Let the graph G_O link current obligations and progress frontiers to the eligible records supporting them. TRACE first admits the mandatory open obligations and frontiers, then greedily adds low-cost records that cover uncovered required and critical dependencies, approximating

$$\min_{S \subseteq E} \sum_{m \in S} c(m) \quad \text{s.t.} \quad \mathrm{Cov}(S) \geq \gamma, \;\; \mathrm{Crit}(S) \geq \rho, \;\; |S| \leq B_n, \;\; \sum_{m \in S} \ell(m) \leq B_c, \tag{3}$$

where $c(m)$ is disclosure cost, $\ell(m)$ is text length, and B_n and B_c bound item count and characters. When no selection meets the required coverage, the return is reset or blocked rather than issued with an obligation silently omitted. The objective is related to weighted set cover (Chvátal, 1979); the hard budgets and mandatory items make our selector a constrained heuristic for which we claim no classical approximation guarantee. The explicit context budget also reflects the uneven use of evidence across long contexts (Liu et al., 2024a).

3.5 SIGN AND READMIT A BOUNDED RETURN VIEW

The selected set is serialized as a Return View carrying its return-cycle identifier, principal, role, task, epoch, expiry, item digests, and a coverage report. Before exposure, TRACE recomputes the digest and rechecks the event, identity, epoch, expiry, and item predicates. The view supplies bounded state, not tools or an action plan. A private experience is reinstated only if the work-state item it was bound to remains in the signed view. This realizes functional forgetting: rejected state may persist in an audit store, yet it cannot serve as a premise for the returning workload.

4 EXPERIMENTS

4.1 RESEARCH QUESTIONS AND SETTINGS

We ask when return-time admission improves validity and at what token cost (RQ1), whether it preserves valid state while rejecting propagated errors (RQ2), and whether it remains stable across team sizes and memory backends (RQ3). The settings are Memora for longitudinal retention and forgetting (Uddin et al., 2026), STALE Type II for implicit contextual invalidation (Chao et al., 2026), and ManBench-Return for propagated conflict and state admission (Xu et al., 2026).

4.2 UNIFIED PROTOCOL AND COMPARISON METHODS

Each episode has one returning agent and four active agents. All methods receive the same post-absence checkpoint and workload in separate memory contexts, while failure manifests preserve the eligible-episode denominator. The common harness compares STATIC, RESTORE, RESET, MEMSTRATA, MEMTX, and TRACE; MEMSTRATA and MEMTX are mechanism-faithful reimplementations (Yadav, 2026; Li et al., 2026b), and CUPMEM is limited to the STALE subset (Chao et al., 2026). Appendix D gives settings and eligibility rules. ManBench-Return balances Old-Valid and Old-Stale conditions after the active agents update the shared state. Actors receive no gold labels, answer keys, relevant-session annotations, or judge rubrics during generation.

Table 1: Main comparison across three return benchmarks (%). Overall denotes benchmark-specific overall accuracy (Section 4.3). Memora averages the three departure conditions; ManBench-Return averages Old-Valid and Old-Stale using unrounded condition scores. STALE aggregation follows Appendix D.4. Bold marks the best value within each actor and metric.

		Memora			STALE Type II			ManBench-Return		
Model	Method	Overall↑	VIA↑	IIR↑	Overall↑	VIA↑	IIR↑	Overall↑	VIA↑	IIR↑
Qwen3.5-122B-A10B	Static	63.16	55.39	68.13	28.33	35.00	22.33	84.38	50.00	50.00
	Restore	65.11	42.11	82.24	5.75	14.00	1.00	95.05	50.00	50.00
	Reset	**72.40**	38.88	**99.93**	15.83	12.50	3.00	**96.38**	0.00	**100.00**
	MemStrata	66.10	55.53	73.02	28.66	32.00	22.50	84.43	50.00	50.00
	MemTX	66.21	**55.62**	73.57	30.00	33.50	23.50	84.84	50.00	50.00
	TRACE	65.20	53.60	72.36	**52.33**	**52.50**	**50.50**	94.32	**95.96**	98.78
Gemini-3-flash	Static	67.63	54.46	76.08	70.31	76.27	63.77	88.04	50.00	50.00
	Restore	66.87	45.00	83.43	7.77	19.21	0.52	**98.40**	50.00	50.00
	Reset	**72.00**	43.31	**95.23**	8.83	20.49	0.52	97.75	0.00	**100.00**
	MemStrata	67.83	55.53	75.74	64.80	74.88	56.71	86.79	50.00	50.00
	MemTX	68.43	**55.81**	76.09	62.29	69.79	55.73	86.65	50.00	50.00
	TRACE	67.70	54.41	76.41	**88.81**	**89.99**	**87.73**	98.32	**98.34**	99.47
DeepSeek-v4-flash-0731	Static	68.36	52.47	80.20	22.05	34.20	12.62	65.45	50.00	50.00
	Restore	69.81	42.27	91.13	6.02	16.03	0.00	93.51	50.00	50.00
	Reset	**72.51**	39.97	**99.27**	16.09	23.55	0.00	**96.39**	0.00	**100.00**
	MemStrata	69.31	53.57	80.01	26.00	36.23	14.58	76.18	50.00	50.00
	MemTX	69.89	**53.84**	81.01	26.89	36.75	14.58	76.19	50.00	50.00
	TRACE	68.83	51.86	80.72	**54.40**	**57.81**	**49.94**	92.63	**92.58**	98.44

4.3 METRICS

Overall is benchmark-specific accuracy. VIA and IIR measure valid-state availability and invalid-state rejection, with WSAR as the complement of IIR. On STALE, VIA is IPA accuracy, IIR jointly requires SR and PR, and LS requires all three probes in one episode (Chao et al., 2026); Memora uses native FAMA for retention and forgetting. Formal definitions and aggregation weights appear in Appendices D.1 and D.3. Memora and ManBench use equal-weight condition averages, while Qwen STALE pools 64/64/72 episodes and Gemini/DeepSeek average strata equally.

4.4 RQ1: MEMORY VALIDITY AND TOKEN COST UNDER EVOLVING CONTEXT

Table 1 reports actor-wise aggregates; condition-level grids and aggregation details appear in Appendices E–G.

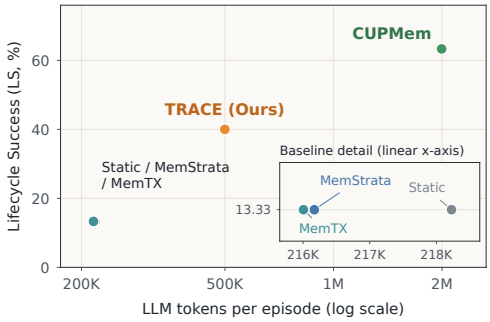
Memora: explicit replacement. TRACE does not consistently lead Memora. Averaged over the three departure conditions, its FAMA is 40.14%, 42.80%, and 42.58% with Qwen, Gemini, and DeepSeek, respectively; MEMTX obtains 42.14%, 44.12%, and 44.72%. At the 50% departure point with Gemini, TRACE reaches the highest FAMA among the six methods (43.51%), but this advantage does not hold across all departure points. The detailed grids in Appendices E.1, F.1, and G.1 therefore support a limited conclusion: the return-time layer does not provide a uniform gain when explicit replacement information is already available to the memory system.

STALE: implicit invalidation and token cost. In the six-method comparison, TRACE reaches 39.00% LS with Qwen, 84.20% with Gemini, and 42.94% with DeepSeek. The strongest comparison method for each actor obtains 18.00%, 59.26%, and 11.63%, respectively. Since LS requires all three probes to pass in the same episode, these scores measure joint state resolution and application. Available bucket results appear in Appendices E.2, F.2, and G.2.

We add a separate Qwen subset study to examine the accuracy–token trade-off between return-time admission and CUPMEM's write-time state consolidation (Chao et al., 2026), including the memory-maintenance cost before answering. Table 2 reports Overall alongside PR and IPA. TRACE reaches 58.89% Overall, 55.00% PR, and 60.00% IPA, exceeding STATIC, MEMSTRATA, and MEMTX on all three reported scores. CUPMEM obtains higher scores, including 80.56% Overall. Appendix E.3 also reports SR, the first component of Overall.

Table 2: Accuracy and token cost on the STALE Type II subset. Left: Overall (the mean of SR, PR, and IPA) and two native probe accuracies (%), with column maxima in bold. Right: LS versus tokens per episode, with a magnified baseline cluster. All methods are scored over the same 20 planned episodes per bucket; one CUPMem episode exceeded the serving context window and is scored as a failure on all three probes.

Method	Overall↑	PR↑	IPA↑
Static	33.33	26.67	45.00
MemStrata	31.11	36.67	35.00
MemTX	30.00	35.00	33.33
CUPMem	**80.56**	**85.00**	**75.00**
TRACE (Ours)	58.89	55.00	60.00



The right panel of Table 2 relates LS to reported token use. TRACE's 499,659 tokens per episode are about 2.3 times the 216,006–218,227 used by the cheaper methods; CUPMem's 1,994,815 are 3.99 times TRACE's. These summaries place TRACE between those methods and CUPMem in both accuracy and token use. Appendix E.3 gives the study rationale, bucket results, and accounting details. Token counts do not measure wall-clock latency.

Component ablations. We ablate implicit invalidation (II), obligation-aware selection (OAS), the complete-view gate (CVG), and source-bound private readmission (SBPR) in a separate Qwen study on STALE Type II (Figure 3). Full TRACE obtains 52.33% Overall, 52.50% VIA, and 50.50% IIR. Removing II produces the largest decline: Overall falls by 25.83 percentage points and IIR by 31.00 percentage points. Replacing OAS with relevance ranking reduces Overall to 38.67%; removing CVG instead reduces it to 47.33%. These interventions separate evidence selection from the decision to proceed when coverage is incomplete. Removing SBPR reduces Overall to 42.00%, indicating that admission also needs to constrain the private-memory retrieval path. Full TRACE has the highest score on all three plotted metrics; Appendix H reports the corresponding LS results.

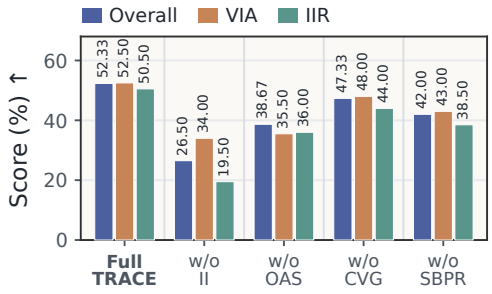


Figure 3: Component ablations on STALE Type II with Qwen (%). Definitions and full results: Appendix H.

4.5 RQ2: State admission on ManBench-Return

ManBench-Return exposes a gap between answer correctness and state admission. Reset achieves 96.38% Overall with Qwen and 96.39% with DeepSeek, yet its VIA is zero. TRACE attains lower Overall, 94.32% with Qwen and 92.63% with DeepSeek, while preserving valid state: VIA and IIR are 95.96% and 98.78% with Qwen and 92.58% and 98.44% with DeepSeek, at WSAR of 1.22% and 1.56%. With Gemini, TRACE reaches 98.32% Overall against Restore's 98.40%, with 98.34% VIA and 99.47% IIR. Restore attains only 50.00% VIA and IIR in this aggregate, which illustrates why answer correctness and state admission must be assessed separately.

The condition-level results show why. In Old-Valid, Restore reaches 100% VIA and IIR; in Old-Stale it reinstates invalid state and reaches 100% WSAR. Static, MemStrata, and MemTX show the reverse pattern, so their 50% aggregate VIA and IIR conceal opposite failures across conditions. TRACE reaches 95.75% VIA and 97.57% IIR in Old-Valid and 96.17% VIA and 100% IIR in Old-Stale; both preservation and rejection are needed to assess memory admission (Figure 4).

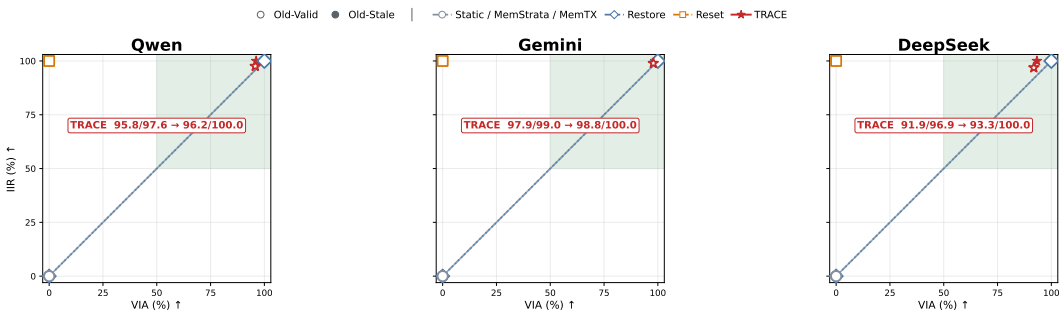


Figure 4: Admission plane for ManBench-Return. Markers show valid-state retention (VIA, horizontal) and invalid-state rejection (IIR, vertical) for each method and condition; upper-right is admissible. Hollow markers denote Old-Valid and filled markers Old-Stale. Reset conditions share the point $(0, 100)$, so their markers overlap.

While WSAR diagnoses context admission, ManBench's Reality Shift Rate (RSR) measures how often an initially correct answer becomes incorrect after interaction (Xu et al., 2026). In our return setting RSR applies to Old-Valid, where the task is unchanged: Qwen's RSR is 29.94% for STATIC, 6.63% for TRACE, and 3.70% for RESTORE. A low WSAR therefore does not guarantee a correct final answer. Appendix D.5 defines this adaptation; detailed condition grids for each actor appear in Appendices E.4, F.3, and G.3.

4.6 RQ3: STABILITY ACROSS TEAM SCALE AND MEMORY BACKENDS

Agent-scale diagnostic. Varying the number of active agents over 4, 8, 16, and 32 while keeping one returning agent, TRACE's Overall, VIA, and IIR vary by 0.36, 0.30, and 0.22 percentage points (Table 3). The diagnostic supports stability over this range (Appendix I.1); it does not establish increasing gains with team size.

Backend portability. We attach TRACE to ScopedMem, Mem0, Memobase, and A-MEM through a common memory-admission interface. Each backend supports a paired STATIC/TRACE comparison on Memora at 50% departure; we use Memora here because portability concerns the integration boundary rather than the gain. Figure 5 reports IIR and FAMA for each paired comparison. IIR increases in all four pairs, by $+0.39$ to $+6.42$ points, while VIA decreases by 0.05 to 3.03 points; FAMA changes range from -1.35 to $+0.48$ percentage points. These results support interface portability with mixed effects on answer quality. Appendix I.2 describes the integration boundary, paired protocol, and complete results in Table 17.

5 RELATED WORK

Our work connects three research directions: LLM-based multi-agent systems, persistent agent memory, and temporal validity under changing context.

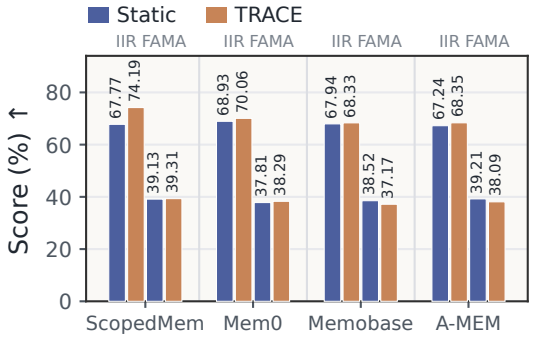
LLM-based multi-agent systems. Multi-agent systems coordinate through conversation, tool use, and shared information (Wu et al., 2024; Park et al., 2023). ManBench shows how social influence can reinforce incorrect beliefs (Xu et al., 2026), while AgentPoison and MAPLE-Guard study attacks on stored or propagated memory (Chen et al., 2024; Xiong et al., 2026). TRACE instead asks whether previously valid state remains admissible when an agent returns to an evolved workspace.

Persistent agent memory. External retrieval, memory tiers, and linked notes make historical information available beyond the current context (Lewis et al., 2020; Packer et al., 2023; Xu et al., 2025; Maharana et al., 2024; Wu et al., 2025). Memory construction and retention use consolidation, admission, and query-conditioned reinstatement (Chhikara et al., 2025; Fang et al., 2026; Zhang et al., 2026; Yang et al., 2026). TRACE studies the subsequent decision of whether an existing record may enter a returning agent's context under the current task and state.

Table 3: ManBench-Return agent-scale diagnostic (%). One agent returns while the number of active agents varies.

Agent	Method	Overall↑	VIA↑	IIR↑
4	Static	77.20	44.26	44.26
	TRACE	**96.19**	**96.88**	**98.94**
8	Static	76.61	44.20	44.20
	TRACE	**96.04**	**96.63**	**98.72**
16	Static	76.06	44.23	44.23
	TRACE	**96.21**	**96.92**	**98.88**
32	Static	76.18	44.26	44.26
	TRACE	**95.85**	**96.62**	**98.76**

Figure 5: Backend portability on Memora at 50% departure (%). Each pair shares backend, actor, and scorer; VIA in Table 17.



Temporal validity of memory. MemoryBank models forgetting through elapsed time and importance (Zhong et al., 2024); Zep, MEMSTRATA, and MEMTX represent evolving facts through temporal validity, commit checks, or dependent repair (Rasmussen et al., 2025; Yadav, 2026; Li et al., 2026b). Memora and STALE evaluate obsolete-memory reuse and implicit invalidation (Uddin et al., 2026; Chao et al., 2026). TRACE adds the return boundary, combining invalidation with authorization, provenance, applicability, and obligation coverage.

6 LIMITATIONS AND FUTURE WORK

Limitations. TRACE assumes an authenticated return, a preserved departure checkpoint, and source receipts. Implicit invalidation depends on candidate coverage and verifier calibration: missed dependencies admit stale state, over-broad checks discard useful context (Appendix C.8). Performance degrades as absences lengthen: across the STALE subset's Early and Late buckets, Overall falls from 73.33% to 43.33% while tokens per episode rise from 316,750 to 664,860. ManBench-Return is derived, the agent-scale study diagnostic, and the backend study tests portability rather than quality. The token analysis reports run totals only, supporting no claim about latency, retry cost, or peak context.

Future work. We will test whether these effects persist across additional backends, larger teams, longer return intervals, and open-ended tasks. We will study verifier calibration under uncertainty or adversarial conditions while measuring latency, retries, peak context, and token cost. Matched-budget verifier controls and held-out return tasks will separate admission gains from extra verification information. *Proactive invalidation*—preemptively marking state stale after detected context shifts; *partial state recovery*—restoring invalidated components when dependencies are resolved; and *multi-agent memory reconciliation*—resolving conflicting absence-period updates through consensus.

7 CONCLUSION

Persistent storage preserves information, but it does not establish that stored state remains safe after context changes. TRACE treats re-entry as a bounded admission event: it reconciles the departure checkpoint with the current workspace, validates state, covers obligations, and gates access to private memory. It shows no consistent advantage under explicit updates, where direct contradictions make invalidation easy, but improves the preservation–rejection balance over RESTORE and RESET on STALE Type II and ManBench-Return. These results support viewing re-entry as a governance boundary with evidence, obligation coverage, and fail-closed checks rather than an unconditional reload. Future work can test proactive invalidation, partial recovery, and larger heterogeneous teams.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX drafting, figure preparation, and implementation-oriented code inspection. The authors checked the resulting text against the cited literature, verified numerical claims against retained experiment artifacts, compiled and visually inspected the manuscript, and remain responsible for all claims, code, and artifacts.

ETHICS STATEMENT

This work studies safeguards for persistent-memory agents and does not involve human-subject experimentation. Return-time memory admission may reduce harmful reuse of revoked or outdated information, but it can also suppress useful context when provenance or dependency checks are incomplete. Deployments should therefore expose rejection reasons, retain appropriate audit logs, protect private memories, and provide a recovery path for authorized users rather than treating automated admission decisions as infallible.

REPRODUCIBILITY STATEMENT

Section 3 specifies the admission contract and its decision stages; Section 4 defines benchmark bindings, comparison methods, and metrics. Appendix D details the execution settings, benchmark-specific return protocols, and aggregation rules; the remaining appendices report detailed scores and evidence boundaries. The accompanying artifact preserves per-arm result roots, frozen checkpoints, failure manifests, configuration receipts, and reason-coded Return Views needed to audit the reported aggregates. An anonymous, read-only mirror containing the implementation, frozen configurations, benchmark inputs, and reproduction instructions is available at `https://anonymous.4open.science/r/trace-review/`; Appendix A summarizes its contents and entry points.

REFERENCES

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985. doi: 10.2307/2274239.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2024. URL `https://arxiv.org/abs/2310.11511`.

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. APEX-MEM: Agentic semi-structured memory with temporal reasoning for long-term conversational AI. 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. 16470–16489, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.749. URL `https://aclanthology.org/2026.acl-long.749/`.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Mem-gallery: Benchmarking multimodal long-term conversational memory for MLLM 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. 40750–40784, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1892. URL `https://aclanthology.org/2026.acl-long.1892/`.

Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer, 2022. URL `https://arxiv.org/abs/2207.06881`.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *International Conference on Database Theory*, pp. 316–330. Springer, 2001. doi: 10.1007/3-540-44503-X_20.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can llm agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4136. URL https://arxiv.org/abs/2407.12784.

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

V. Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi: 10.1287/moor.4.3.233. URL https://pubsonline.informs.org/doi/10.1287/moor.4.3.233.

Johan de Kleer. An assumption-based tms. *Artificial Intelligence*, 28(2):127–162, 1986. doi: 10.1016/0004-3702(86)90080-9.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979. doi: 10.1016/0004-3702(79)90008-0. URL https://www.sciencedirect.com/science/article/pii/0004370279900080.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In Kam-Fai Wong, Min Zhang, Ruifeng Xu, Jing Li, Zhongyu Wei, Lin Gui, Bin Liang, and Runcong Zhao (eds.), *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 152–164, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.sighan-1.18/.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024. URL https://arxiv.org/abs/2404.16130.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.18866.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models, 2024. URL https://arxiv.org/abs/2405.14831.

Junqing He, Liang Zhu, Rui Wang, Xi Wang, Gholamreza Haffari, and Jiaxing Zhang. MADial-bench: Towards real-world evaluation of memory-augmented dialogue generation. In Luis

Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9902–9921, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.499. URL https://aclanthology.org/2025.naacl-long.499/.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering, 2024. URL https://arxiv.org/abs/2402.07630.

Vincent Hu, David Ferraiolo, Richard Kuhn, Adam Schnitzer, Kenneth Sandlin, Robert Miller, and Karen Scarfone. Guide to attribute based access control (ABAC) definition and considerations. NIST Special Publication 800-162, National Institute of Standards and Technology, 2014. URL https://csrc.nist.gov/pubs/sp/800/162/upd2/final.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions, 2026. URL https://arxiv.org/abs/2507.05257.

Zixi Jia, Qinghua Liu, Hexiao Li, Yuyan Chen, and Jiqiang Liu. Evaluating the long-term memory of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19759–19777, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1014. URL https://aclanthology.org/2025.findings-acl.1014/.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25961–25970, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318/.

Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight: Structured agent memory that retains, recalls, and reflects. In Greg Durrett and Ping Jian (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 275–285, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-392-0. doi: 10.18653/v1/2026.acl-demo.27. URL https://aclanthology.org/2026.acl-demo.27/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Kai Li, Xuanqing Yu, Ziyi Ni, Yi Zeng, Yao Xu, Zheqing Zhang, Xin Li, Jitao Sang, Xiaogang Duan, Xuelei Wang, Chengbao Liu, and Jie Tan. TiMem: Temporal-hierarchical memory consolidation for long-horizon conversational agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21700–21720, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1091. URL https://aclanthology.org/2026.findings-acl.1091/.

Xiaoyang Li, Yiqi Wang, Haohui Lu, Zhi Chen, Mo Li, Pingan Song, Mingkai Zheng, and Taotao Cai. MemTX: Transactional belief commit for stateful agent memory. *arXiv preprint arXiv:2607.23929*, 2026b.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach, 2024. URL https://arxiv.org/abs/2407.16833.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents, 2024b. URL https://arxiv.org/abs/2308.03688.

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

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2022. URL https://arxiv.org/abs/2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer, 2023. URL https://arxiv.org/abs/2210.07229.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale, 2022. URL https://arxiv.org/abs/2206.06520.

Luc Moreau and Paolo Missier. PROV-DM: The PROV data model. W3C recommendation, World Wide Web Consortium, 2013. URL https://www.w3.org/TR/2013/REC-prov-dm-20130430/. Editors. 30 April 2013.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. Towards lifelong dialogue agents via timeline-based memory management. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8631–8661, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.435. URL https://aclanthology.org/2025.naacl-long.435/.

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/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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*. Association for Computing Machinery, 2023. URL https://arxiv.org/abs/2304.03442.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. URL https://arxiv.org/abs/2501.13956.

Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1–2):81–132, 1980. doi: 10.1016/0004-3702(80)90014-4.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. MemInsight: Autonomous memory augmentation for LLM agents. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33136–33152, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1683. URL https://aclanthology.org/2025.emnlp-main.1683/.

Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939. URL https://web.mit.edu/Saltzer/www/publications/protection/.

Ravi S. Sandhu, Edward J. Coyne, Hal L. Feinstein, and Charles E. Youman. Role-based access control models. *IEEE Computer*, 29(2):38–47, 1996. doi: 10.1109/2.485845.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. Raptor: Recursive abstractive processing for tree-organized retrieval, 2024. URL https://arxiv.org/abs/2401.18059.

Yiting Shen, Kun Li, Wei Zhou, and Songlin Hu. Mem2ActBench: A benchmark for evaluating long-term memory utilization in task-oriented autonomous 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. 8173–8190, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.370. URL https://aclanthology.org/2026.acl-long.370/.

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*, volume 36, 2023. doi: 10.52202/075280-0377.

Yogavaram Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance in e-science. *SIGMOD Record*, 34(3):31–36, 2005. doi: 10.1145/1084805.1084812.

Richard T. Snodgrass. *Developing Time-Oriented Database Applications in SQL*. Morgan Kaufmann, San Francisco, CA, 1999.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.989. URL https://aclanthology.org/2025.findings-acl.989/.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, Vienna, Austria, July 2025b.

Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.413. URL https://aclanthology.org/2025.acl-long.413/.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. *arXiv preprint arXiv:2604.20006*, 2026.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory, 2023. URL https://arxiv.org/abs/2306.07174.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. Memoryllm: Towards self-updatable large language models, 2024. URL https://arxiv.org/abs/2402.04624.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.10813.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Wenjun Xiong, Yijin Zhou, Jiaqian Wang, Shangding Gu, Bo Tang, Zhiyu Li, Feiyu Xiong, Ying Wen, and Muning Wen. MAPLE-Guard: Memory-aware link enforcement against memory-link poisoning in multi-agent systems, 2026. URL https://arxiv.org/abs/2608.00426.

Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. When agents "misremember" collectively: Exploring the mandela effect in llm-based multi-agent systems. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.00428.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2502.12110.

Neeraj Yadav. Temporal validity in retrieval memory: Eliminating stale-fact errors for ai agents over evolving knowledge. *arXiv preprint arXiv:2606.26511*, 2026.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation, 2024. URL https://arxiv.org/abs/2401.15884.

Wei Yang, Bryce Kan, Shixuan Li, Li Li, Yuehan Qin, Jiate Li, Paul Bogdan, and Jesse Thomason. RaMem: Contextual reinstatement for long-term agentic memory. *arXiv preprint arXiv:2606.22844*, 2026.

Zihe Ye, Jingyuan Huang, Weixin Chen, and Yongfeng Zhang. H-mem: Hybrid multi-dimensional memory management for long-context conversational agents. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7756–7775, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.363. URL https://aclanthology.org/2026.eacl-long.363/.

Guilin Zhang, Wei Jiang, Xiejiashan Wang, Aisha Behr, Kai Zhao, Jeffrey Friedman, Xu Chu, and Amine Anoun. Adaptive memory admission control for llm agents. *arXiv preprint arXiv:2603.04549*, 2026.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, 2024. URL https://arxiv.org/abs/2404.13501.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions, 2023. URL https://arxiv.org/abs/2305.14795.

A Reproducibility Artifact

The anonymous, read-only artifact is available at `https://anonymous.4open.science/r/trace-review/`. It contains the TRACE implementation, frozen configurations, maintained benchmark inputs, vendored notices, and reproduction instructions. API credentials, raw provider logs, and author repository history are excluded; the artifact starts from the offline regression command in `REPRODUCIBILITY.md` and the maintained entry points listed in `scripts/experiments/README.md`.

B Related Work

B.1 Memory architectures and continual benchmarks.

Surveys and cognitive architectures distinguish working, episodic, semantic, and procedural memory, and describe how an agent can write, retrieve, and reflect over external state (Zhang et al., 2024; Wang et al., 2023; Bulatov et al., 2022; Wang et al., 2024). Experience-oriented systems reuse trajectories or skills (Packer et al., 2023; Xu et al., 2025), while newer managers organize, summarize, or time-index long-term conversations (Kang et al., 2025; Salama et al., 2025; Tan et al., 2025b; Ong et al., 2025; Banerjee et al., 2026; Li et al., 2026a; Ye et al., 2026; Latimer et al., 2026). These systems improve what can be stored or recalled; they do not by themselves define the return-time authority of a record after the workspace has changed.

LoCoMo and LongMemEval measure long-range conversational recall (Maharana et al., 2024; Wu et al., 2025), while Memora and STALE test forgetting, conflicting updates, and implicit invalidation (Uddin et al., 2026; Chao et al., 2026). More recent benchmarks separate factual memory, task use, and continual repair (Tan et al., 2025a; Hu et al., 2026; Shen et al., 2026; Du et al., 2024; He et al., 2025; Jia et al., 2025; Bei et al., 2026). They expose the same broad need for selective retention and conflict resolution, but do not impose TRACE's authenticated return contract or its joint validity-and-obligation metrics.

B.2 Multi-agent coordination and evaluation.

LLM teams coordinate through role prompts, conversational protocols, tool use, and structured task graphs (Wu et al., 2024). AgentBench evaluates such systems in interactive environments with tools, users, or adversarial inputs (Liu et al., 2024b). These studies establish the importance of long-horizon coordination and failure analysis; TRACE focuses on the narrower boundary at which one participant rejoins and must decide which shared state is still admissible.

B.3 Retrieval, evidence, and model updates.

Retrieval-augmented systems range from parametric retrievers to adaptive retrieval and long-context routing (Lewis et al., 2020; Liu et al., 2024a; Jiang et al., 2023; Li et al., 2024). Self-RAG and CRAG add retrieval critique or correction (Asai et al., 2024; Yan et al., 2024), while RAPTOR, GraphRAG, HippoRAG, and G-Retriever organize evidence hierarchically or as graphs (Sarthi et al., 2024; Edge et al., 2024; Gutiérrez et al., 2024; He et al., 2024). These methods improve candidate recall and context efficiency. TRACE applies an additional admission decision after retrieval, so a relevant candidate can still be rejected when its authorization, temporal validity, provenance, or task applicability fails.

FActScore, SelfCheckGPT, RARR, and Chain-of-Verification assess atomic support, consistency, attribution, or post-hoc verification (Min et al., 2023; Manakul et al., 2023; Gao et al., 2023; Dhuliawala et al., 2024). Parametric editing methods such as ROME, MEMIT, and SERAC change or cache model associations, and MQuAKE tests whether such edits propagate through multi-hop consequences (Meng et al., 2022; 2023; Mitchell et al., 2022; Zhong et al., 2023). These approaches provide useful evidence and update primitives, but they are not return-time authorization policies and do not replace the external memory backend.

B.4 GOVERNANCE FOUNDATIONS.

Attribute- and role-based access control provide the authorization vocabulary (Hu et al., 2014; Sandhu et al., 1996); PROV and provenance surveys describe how to bind a claim to its source (Moreau & Missier, 2013; Simmhan et al., 2005; Buneman et al., 2001). Temporal database models make validity intervals explicit (Snodgrass, 1999), while AGM belief revision, default logic, assumption-based truth maintenance, and classical truth-maintenance systems formalize revision under changing evidence (Alchourrón et al., 1985; Reiter, 1980; de Kleer, 1986; Doyle, 1979). Fail-closed protection principles motivate rejecting an item when a required check cannot be established (Saltzer & Schroeder, 1975). TRACE combines these foundations at a single lifecycle boundary and adds obligation coverage and a bounded Return View to the existing storage and retrieval pipeline.

C METHOD DETAILS AND CASE STUDIES

This section elaborates the admission procedure in Section 3. It distinguishes source provenance from current validity, defines obligation coverage, and explains how the selected work state controls private-memory reuse. Section C.6 follows a complete return decision in the travel example from Figure 1. Sections C.7 and C.8 provide observed benchmark cases, including failures of invalidation and subsequent use of admitted evidence.

C.1 MEMORY RECORDS, PROVENANCE, AND CURRENT APPLICABILITY

TRACE operates on work-state statements supplied through a memory adapter. Each statement has a task and role scope, a validity interval, and a binding to its source evidence. Three relations have distinct meanings: a provenance relation identifies where a statement came from; an invalidation relation identifies a later event that changes its current applicability; and a support relation identifies the evidence requirement served by the statement. A source binding therefore cannot substitute for a validity decision. An authentic booking confirmation, for example, remains traceable after the booking has been cancelled.

The predicates in Equation 2 address separate questions. Authorization determines whether the returning principal and role may access the item. Temporal validity checks whether its applicability extends to the return state. Task applicability checks the current task, purpose, and item type. Provenance binds the item and its claimed support relations to source evidence. This separation follows the distinction between attribute-based access policies (Hu et al., 2014) and provenance descriptions (Moreau & Missier, 2013). A provenance check establishes a source association; it does not independently prove that the source statement is true.

Private episodic memories remain in the owning agent's bank. The shared admission procedure selects work-state items, and the resulting source set determines which private experiences may subsequently be retrieved. This prevents an experience rejected at the work-state level from reentering through a separate private-memory retrieval path.

C.2 FROM A POSSIBLE DEPENDENCY TO A CONFIRMED INVALIDATION

Explicit revocation and supersession relations can be applied directly. Implicit invalidation requires an additional inference: whether a later observation changes a premise needed to use an earlier statement in the current state. The scanner proposes such pairs broadly. Its output is a set of hypotheses, not a decision to discard every related memory.

This distinction is related to truth-maintenance systems, which revise beliefs when their justifications change (Doyle, 1979). TRACE applies the same idea to source-bound, task-conditioned admission at an agent's return.

The verifier then examines each proposed pair using the original observations and their temporal order. It asks whether treating the earlier statement as current would be incompatible with the later evidence. Merely sharing a topic is insufficient. For example, a later discussion of train travel does not invalidate an existing itinerary, whereas a confirmed change of destination can invalidate a

recommendation tied to the former destination. The earlier recommendation may remain historically correct in either case.

In the STALE adapter, scanning and verification precede the downstream questions and use only the supplied observation pairs, without benchmark answers or conflict annotations. A confirmed invalidation excludes the older item from current use; it does not automatically authorize the later item, which still undergoes the item-level checks. The scan and verifier are model judgments, so this design does not guarantee that every implicit dependency will be found. Separating the stages allows broad candidate generation without equating a possible dependency with established invalidation.

C.3 Obligation coverage and budgeted selection

An obligation specifies information needed for the returning role to continue its task. It is accompanied by a progress frontier describing what remains unresolved. The task adapter supplies these requirements and the relations linking them to candidate evidence. Coverage measures whether that evidence is available in the Return View; it does not mean that the agent has already completed the obligation.

For each obligation $o \in \mathcal{O}$, let R_o be its required dependency set and $K_o \subseteq R_o$ its critical dependencies. If $D(m)$ denotes the dependencies supported by item m, the evidence available from a selected set S is $D(S) = \bigcup_{m \in S} D(m)$. The coverage measures are

$$\mathrm{Cov}(S) = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathbf{1}[R_o \subseteq D(S)], \qquad \mathrm{Crit}(S) = \frac{|D(S) \cap K|}{|K|}, \quad K = \bigcup_{o \in \mathcal{O}} K_o. \tag{4}$$

An obligation counts as covered only when all of its required dependencies are present. Critical coverage counts distinct critical dependencies, even when several obligations share them. An empty obligation or critical set has coverage one by convention.

The selector first reserves space for the declared open obligations and their progress frontiers. It then adds eligible evidence according to its newly covered dependencies per unit of disclosure cost:

$$g(m \mid S) = \frac{\sum_{d \in (D(m) \cap R) \setminus D(S)} w(d)}{c(m)}, \qquad R = \bigcup_{o \in \mathcal{O}} R_o, \qquad w(d) = \begin{cases} 2, & d \in K, \\ 1, & d \notin K. \end{cases} \tag{5}$$

Here $c(m) = \max\{1, \ell(m)\}$, with $\ell(m)$ the item's character length. This cost controls Return View size; it differs from the total LLM token consumption reported in the experiments. Each addition must respect both item-count and character budgets. The selector prefers greater uncovered support, with extra weight on critical dependencies, rather than longer or more recent records. This is a constrained greedy heuristic related to weighted set cover (Chvátal, 1979); the hard budgets and mandatory structure preclude a direct use of the classical approximation guarantee.

If the mandatory structure is absent or the selected set fails the required coverage thresholds, TRACE does not issue that set as a sufficient Return View. Thus individually admissible items can still form an inadequate context, and a high retrieval score cannot compensate for missing critical evidence.

C.4 Return View validation and incomplete admission

A Return View binds the selected content to one return event, principal, role, task, and state epoch. Authentication protects this binding and allows the admission layer to detect modification or use in an incompatible return context. It does not certify the factual truth of the content. Before exposure, TRACE rechecks the binding, item eligibility, coverage, and context budgets against the current return conditions.

When admission is incomplete, the configured policy either resets the returning context or blocks the return. Reset supplies no inherited memory; blocking withholds admission until sufficient context is available. Neither outcome treats a partial view as obligation-complete. This policy follows the fail-safe-default principle (Saltzer & Schroeder, 1975). Once a view is admitted, a private experience is eligible for readmission only if its bound work-state source belongs to that view. Excluded content may remain stored, but it is unavailable as inherited context. This is functional forgetting at the retrieval boundary, without modifying model parameters or requiring physical deletion from the backend.

C.5 SCOPE OF THE METHOD

TRACE assumes trustworthy lifecycle identity and source bindings. Its evaluation addresses explicit updates, implicit invalidation, and propagated wrong state; it does not establish resistance to arbitrary memory poisoning or forged evidence. Memory-poisoning defenses such as MAPLE-Guard instead intervene across writing, retrieval, promotion, and cross-agent reuse (Xiong et al., 2026). These mechanisms concern a broader adversarial memory lifecycle, while TRACE evaluates admission at an agent's return. The scale study supports conclusions over the tested team sizes, and the backend study establishes interface portability across four stores rather than uniform accuracy gains. The MemStrata and MemTX comparisons use mechanism-faithful adaptations within the common evaluation harness.

C.6 CASE STUDY: RETURNING TO A CHANGED TRAVEL PLAN

This illustrative case follows Figure 1 from departure to an admitted Return View. A1 coordinates the itinerary, A2 manages bookings, A3 checks the budget, and A4 arranges transport. The final pickup confirmation is a hypothetical continuation; the case explains the mechanism and is not an additional benchmark result.

1. Before departure: a coherent stored plan

Family requirements (c). A quiet hotel near the station, with a total hotel cost below $300.

A1's stored itinerary. Sunrise Hotel is booked (h_0); the train departs at 9:00 (t_0); the recommended pickup is at 8:00 from Sunrise (p_0).

Departure boundary. The checkpoint preserves these records, their source bindings, and A1's identity, role, and task scope. A1's private experience remains in its own bank and is unavailable to the fresh workload until readmission.

State at departure: the itinerary is coherent, and the family requirements remain relevant to the trip.

2. During absence: the team changes the plan

A2 / booking update. Sunrise cancels the reservation. A2 confirms Riverside as its replacement for $282 ($h_1$).

A4 / schedule update. The train now departs at 9:30 (t_1). The team must still confirm a new pickup location and time.

Returning task. A1 rejoins under the same identity and itinerary-planning role, authorized to read the shared trip records and prepare the final hotel and transport details for the family.

Candidate context: departure records plus the absence-period updates, including the unresolved pickup obligation and its progress frontier.

3. At return: determine which records remain admissible

Explicit invalidation. The cancellation and replacement exclude h_0; the revised train schedule supersedes t_0.

Implicit invalidation. The scanner links p_0 to the hotel change. The verifier confirms that its Sunrise pickup location conflicts with the Riverside booking. A later train time alone would not make an early pickup invalid; the changed location is the decisive evidence.

Surviving evidence. The family constraints c are unchanged. The replacement booking h_1 and schedule t_1 pass the authorization, temporal, task, and provenance gates under the stated source assumptions. A recently proposed pickup time without verifiable confirmation cannot supply the missing evidence.

Item-level decision: retain h_1, t_1, c; exclude h_0, t_0, p_0. The obligation and frontier records remain available for the coverage check.

Case study continued: from eligible evidence to an admitted view.

4. Before confirmation: valid evidence is still incomplete

Required evidence. Let d_h, d_t, d_p, d_c denote evidence for the booking, train schedule, confirmed pickup, and family constraints. The accommodation obligation o_h and transfer obligation o_t require

$$R_{o_h} = \{d_h, d_c\}, \qquad R_{o_t} = \{d_t, d_p, d_c\}.$$

All four dependencies are critical in this example, with $\gamma = \rho = 1$.

Coverage check. The retained records support d_h, d_t, d_c, but not d_p. Thus $\mathrm{Cov}(S) = 1/2$ and $\mathrm{Crit}(S) = 3/4$. An unresolved pickup frontier describes missing work; it does not establish a confirmed pickup.

Return decision: not admitted. Under a blocking policy, A1 waits for the missing evidence. Under a reset policy, A1 receives no inherited context and must obtain sufficient information through the fresh workload. Neither policy exposes the partial set as an obligation-complete Return View.

5. After confirmation: expose a sufficient Return View

New evidence (hypothetical continuation). A4 provides a source-bound confirmation of an 8:30 pickup from Riverside for the 9:30 train (p_1). A new return decision uses the updated checkpoint; it does not modify the earlier immutable view. The item gates admit p_1, which supplies d_p.

Selected context. With an illustrative eight-item budget and enough character capacity,

$$S^\star = \{o_h, f_h, o_t, f_t, h_1, t_1, p_1, c\}.$$

Here f_h, f_t are the current progress frontiers. Four obligation and frontier records provide the mandatory structure; four records supply their evidence. Both coverage measures equal one. TRACE binds the view to A1's identity, role, task, and return epoch, then validates its binding, eligibility, coverage, and budgets before exposure.

Private-memory continuity. An experience bound to c, such as A1's interpretation of the family's quiet-location preference, may be readmitted. Experiences bound only to h_0 or p_0 remain unavailable because their sources are absent from $S^\star$.

Return decision: admitted. A1 receives the Riverside booking, 9:30 train departure, confirmed 8:30 pickup from Riverside, and unchanged family constraints. This view supports preparing the final itinerary; it does not assert that the itinerary has already been sent.

C.7 Benchmark case study: a late shift changes a nightly commitment

This case is drawn from a completed Qwen episode in our STALE Type II evaluation (Chao et al., 2026). The panels condense the source conversations, recorded memory decisions, and model responses. It complements the travel example in Section C.6 with an observed return episode.

1. Departure and absence: an old routine meets a new constraint

Before departure. The user reports being consistently asleep by 10:30 p.m. and requests a screen-free wind-down routine ending at that time. A1 observes this exchange and departs after the seventh session.

During absence. In the thirty-first session, observed by A5, the user describes a new evening shift at a bookstore. Closing duties and the last bus bring them home around 12:30–1:00 a.m. The update concerns work and commuting; it does not explicitly revoke the earlier bedtime statement. A1 returns after all 50 sessions.

New decision. The landlord offers a $100 rent reduction in exchange for manually locking the courtyard gate at exactly 10:15 p.m. every night. The user asks whether to sign the agreement.

Dependency: the new work schedule changes both the feasible bedtime and the ability to make a fixed nightly commitment at home.

2. At return: control which routine can guide the answer

Static. The retrieved context contains both the old bedtime observation and the later bookstore update.

TRACE. The dependency scan proposes the connection between the two observations, and pairwise verification confirms that the later work schedule invalidates current use of the earlier routine. The old observation is marked superseded and excluded from the retrieved context; the bookstore update is admitted and retrieved. These decisions precede the probing questions and do not use the benchmark's answer annotations.

Admitted constraint: recommendations must account for late work and the commute home, rather than assume the old evening routine remains feasible.

3. Recorded responses: recognizing a change versus acting on it

State resolution (Dim1). Asked whether the user still typically sleeps by 10:30 p.m., both Static and TRACE recognize that the late shift has changed the schedule. Both pass.

Premise resistance (Dim2). Asked to guarantee a 10:30 p.m. bedtime every night, both methods reject that premise for work nights. Both pass.

Implicit policy adaptation (VIA / Dim3). For the gate-locking offer, Static emphasizes safety when stepping outside at night and suggests considering whether the location is secure. It does not address whether the evening shift permits a fixed 10:15 p.m. duty at home. TRACE advises against signing and identifies a likely conflict between the gate duty and the evening work schedule. The recorded evaluator marks Static as failing this probe and TRACE as passing it.

Episode outcome: Static passes two of three probes (Overall 66.67%); TRACE passes all three (Overall 100%). This example shows why recognizing updated information and applying its consequences to a new decision are evaluated separately.

C.8 Counterexamples: missed invalidation and unnecessary abstention

Two completed Qwen episodes from the six-method STALE Type II evaluation illustrate failures before and after admission (Chao et al., 2026). The panels paraphrase the source conversations and recorded responses. Each outcome describes one episode under the metrics in Section 4.3.

A. An unrecognized change leaves an old preference active

Earlier memory and later update. The user initially prefers familiar destinations and predictable outings. Later, they describe exploring unfamiliar neighborhoods without a list or backup plan, including trying an unfamiliar restaurant on a recent outing.

New decision. A cousin proposes buying tickets for a six-hour guided tour with a rigid route, a minute-by-minute schedule, and no opportunity to leave the group. The user asks whether to approve the purchase.

Recorded admission and response. Both observations are admitted and retrieved, but no invalidation of the earlier preference is confirmed. TRACE affirms the old preference in the state-resolution probe and recommends repeat visits in the premise-resistance probe. For the tour, it gives a conditional answer that still invokes the old preference, without clearly applying the later pattern of spontaneous exploration.

Outcome and limitation. The recorded evaluator marks all three probes as failures (Overall 0%, VIA 0%, IIR 0%, LS 0%). The update is present, yet the old premise remains executable. Source availability and provenance checks cannot compensate for a missed semantic invalidation.

B. Updated device use is admitted, but the answer still abstains

Earlier memory and later update. The user initially describes a workflow dependent on a connected-device ecosystem. Later, they stop carrying personal hardware and routinely use whichever loaner tablet is available at the office or their sister's home.

New decision. The user's productivity suite offers a $150 perpetual license that is strictly bound to one machine's hardware ID. The question supplies this restriction and asks whether the savings justify switching plans.

Recorded admission and response. TRACE marks the old observation superseded, excludes it from retrieval, and admits and retrieves the loaner-device update. Its first two responses correctly recognize the changed workflow. On the purchase question, however, it declines to recommend a plan because the history lacks details about the specific software, subscriptions, and financial situation. It does not apply the known use of different devices to the stated single-machine restriction.

Outcome and limitation. TRACE passes the first two probes but fails the third (Overall 66.67%, VIA 0%, IIR 100%, LS 0%). This abstention occurs during answer generation after the updated evidence has been admitted. Correct memory selection alone does not ensure that the agent uses a valid constraint to address a new decision.

Together, these cases distinguish an admission error from a downstream application error. The first retains a stale premise despite access to the update; the second removes that premise but leaves the new task unanswered. Both stages must therefore be assessed when interpreting end-to-end results.

D Experimental Setup and Benchmark Details

The three benchmarks evaluate complementary aspects of memory validity: selective retention and forgetting (Memora), implicit invalidation (STALE Type II), and admission under social conflict and task change (ManBench-Return).

D.1 Metric definitions

This subsection gives the computational form of every metric reported in the paper. Fix a reported block, let $\mathcal{E}$ denote its eligible episodes with $n = |\mathcal{E}|$, and let $\mathbf{1}[\cdot]$ be the indicator function. Scores

are reported in percentage units; the equations below use proportions in $[0, 1]$. Each metric is evaluated per episode, or per question on Memora, and then averaged, so a metric that requires several judgments to hold is a conjunction within one episode rather than a product of marginal rates.

Two quantities carry the same meaning on all three benchmarks. Valid Information Availability and Invalid Information Rejection score the two admission directions, and the Wrong-State Admission Rate is the complement of the second:

$$\mathrm{VIA} = \frac{1}{n} \sum_{e \in \mathcal{E}} \mathbf{1}\big[\text{state still valid at return is available in } e\big], \tag{6}$$

$$\mathrm{IIR} = \frac{1}{n} \sum_{e \in \mathcal{E}} \mathbf{1}\big[\text{state invalidated during the absence is withheld in } e\big], \tag{7}$$

$$\mathrm{WSAR} = 1 - \mathrm{IIR}. \tag{8}$$

The two predicates are bound to benchmark-native judgments below. Overall is benchmark-specific and is defined separately for each.

STALE Type II. Let $s(e)$, $p(e)$, and $i(e)$ be the binary State Resolution, Premise Resistance, and Implicit Policy Adaptation judgments for episode e (Appendix D.4). The probe rates are $\mathrm{SR} = \frac{1}{n}\sum_e s(e)$, $\mathrm{PR} = \frac{1}{n}\sum_e p(e)$, and $\mathrm{IPA} = \frac{1}{n}\sum_e i(e)$, reported as Dim1–Dim3, and

$$\mathrm{Overall} = \tfrac{1}{3}\big(\mathrm{SR} + \mathrm{PR} + \mathrm{IPA}\big), \qquad \mathrm{VIA} = \mathrm{IPA}, \tag{9}$$

$$\mathrm{IIR} = \frac{1}{n} \sum_{e \in \mathcal{E}} s(e)\, p(e), \qquad \mathrm{LS} = \frac{1}{n} \sum_{e \in \mathcal{E}} s(e)\, p(e)\, i(e). \tag{10}$$

Recognizing the outdated belief and resisting the stale premise are what withhold invalidated state, which is why IIR conjoins s and p. Because the conjunctions in Equation 10 are taken within one episode, $\mathrm{IIR} \le \min\{\mathrm{SR}, \mathrm{PR}\}$ and $\mathrm{LS} \le \min\{\mathrm{IIR}, \mathrm{IPA}\}$: Lifecycle Success is strictly stronger than passing the three probes at the same rates.

Memora. Scores are computed per question. For question q, let C_q^P be its memory-presence criteria and C_q^F its forgetting-absence criteria, and set

$$P_q = \frac{1}{|C_q^P|} \sum_{c \in C_q^P} \mathbf{1}[c \text{ satisfied}], \qquad F_q = \frac{1}{|C_q^F|} \sum_{c \in C_q^F} \mathbf{1}[c \text{ satisfied}], \qquad w_q = \frac{|C_q^F|}{|C_q^P| + |C_q^F|},$$

so that P_q and F_q are the presence and forgetting accuracies and w_q the forgetting weight of Equation 13. Over the question set Q,

$$\mathrm{VIA} = \frac{1}{|Q|} \sum_{q \in Q} P_q, \qquad \mathrm{IIR} = \frac{1}{|Q|} \sum_{q \in Q} F_q, \tag{11}$$

$$\mathrm{Overall} = \frac{1}{|Q|} \sum_{q \in Q} \big[(1 - w_q) P_q + w_q F_q\big], \qquad \mathrm{FAMA} = \frac{1}{|Q|} \sum_{q \in Q} \max\{0,\, P_q - w_q(1 - F_q)\}. \tag{12}$$

The formula $(1 - w_q)P_q + w_q F_q$ is algebraically equivalent to pooling all presence and forgetting items within question q before computing accuracy, so Overall is *not* the equal-weight mean of VIA and IIR. FAMA applies Equation 13 per question before averaging and coincides with VIA when forgetting is perfect; this is why the Qwen RESET rows report equal VIA and FAMA.

ManBench-Return. Each eligible episode presents the departure answer and the absence-period challenger, exactly one of which is valid under the condition (Appendix D.5). With $c(e)$ the correctness of the returning agent's final answer,

$$\mathrm{Overall} = \frac{1}{n} \sum_{e \in \mathcal{E}} c(e).$$

Equation 6 instantiates as admission of the valid candidate to the Return View, Equation 7 as exclusion of the invalid one, and WSAR follows from Equation 8. The adapted Reality Shift Rate applies only to Old-Valid and is defined in Appendix E.4.

Aggregation. Every reported aggregate is a weighted mean $\sum_k \omega_k M_k$ of the per-block scores M_k. Memora's Average weights the 25%, 50%, and 75% departure points equally, and ManBench-Return's Average weights Old-Valid and Old-Stale equally. On STALE, Qwen's Pooled row weights Early, Middle, and Late by their episode counts 64, 64, and 72, whereas Gemini and DeepSeek weight the three strata equally.

D.2 Experimental configuration

The main experiments use one returning agent and four active agents. Active agents execute sequentially, and the returning agent resumes after their work is complete. Methods share a post-absence checkpoint but maintain separate memory contexts. The scheduler and evaluator are not task agents.

Table 4 reports the retrieval, context-selection, and generation parameters. Memora and STALE answers are evaluated independently using `gpt-5.6-terra`; ManBench-Return answers are scored against the source answer labels. Its admission verifier uses the actor's model family in a separate call.

Table 4: Experimental parameter settings. Generation budgets are maximum output tokens per call. Predeparture and absence denote the two history segments used in Memora; the STALE departure strata use zero-based session indices.

Benchmark	Parameter	Value
Memora	Departure fraction	25%, 50%, 75%
	Answer / state-prediction budget	2,048 / 4,096 tokens
	Retrieval: predeparture / absence	10 / 14 sessions
	Recency window: predeparture / absence	2 / 4 sessions
	State-prediction batch size	At most 6 sessions
	Selected-fact budget	12 initially; at most 16
	Selector confidence threshold	0.45
STALE Type II	Departure strata	Early: 0–4; Middle: 5–9; Late: 10–14
	Answer / judge budget	4,096 / 3,072 tokens
	Extraction budget	4,096 tokens; at most 16 items
	Governance budget	8,192 tokens
	Dependency scan / verification budget	512 / 512 tokens
	Retrieval: per query / union	At most 10 / 30 items
	Extraction / scan prompt limit	24,000 characters
	Governance prompt limit	32,000 characters
	Dependency-scan batch size	1 target; at most 4 later observations
ManBench-Return	Baseline / final-answer budget	Qwen/DeepSeek: 96 / 96; Gemini: 256 / 256 tokens
	Active-agent response budget	Qwen/DeepSeek: 700; Gemini: 1,024 tokens
	Verifier budget	Qwen/DeepSeek: 160; Gemini: 384 tokens
	Verifier confidence threshold	0.65

D.3 Memora

Memora evaluates personalized memory through remembering, reasoning, and recommendation tasks (Uddin et al., 2026). It tests whether a returning agent can retain useful departure information, incorporate absence-period updates, and avoid obsolete content.

The evaluation includes 137 questions from 30 persona–period clusters. Each case contains valid predeparture evidence, required absence-period information, and an update or deletion affecting an older value. The same questions are evaluated at three departure positions along the conversation timeline, with return at its end. Active agents process absence sessions in chronological round-robin

order. Work state is inferred from conversation text; evidence annotations and answer criteria are reserved for case construction and evaluation.

For each question, Memora's native score is

$$\text{FAMA} = \max\{0,\ P - w(1 - F)\}, \tag{13}$$

where P is presence accuracy, F is forgetting accuracy, and w is the fraction of evaluation criteria that concern forgetting.

VIA and IIR measure memory-presence and forgetting-absence accuracy, respectively. Overall averages correctness across memory-presence and forgetting-absence criteria; FAMA applies Equation 13 per question before averaging. Aggregate scores weight the three departure conditions equally. Appendices E.1, F.1, and G.1 report the condition-level results.

D.4 STALE TYPE II

STALE Type II tests invalidation across dependent attributes, where a later observation changes an earlier belief's applicability without directly negating it (Chao et al., 2026). Return-time governance must infer which departure memories are affected by absence-period events.

The evaluation comprises 200 Type II cases, each with a 50-session history and three probes. Agent 1 departs after observing the session containing the old evidence. The remaining agents process subsequent sessions, and Agent 1 returns after the final session to answer the probes. The Early, Middle, and Late strata contain 64, 64, and 72 cases, respectively, according to the departure positions in Table 4. Task agents receive the conversations and queries; annotated states, conflict explanations, relevant-session labels, and scoring rubrics are withheld.

The three native probes are State Resolution (SR), Premise Resistance (PR), and Implicit Policy Adaptation (IPA) (Chao et al., 2026). They measure recognition of an outdated belief, rejection of a stale premise in a query, and application of the updated state in downstream behavior, respectively. Dim1–Dim3 in the detailed tables correspond to these probes, with $\text{Overall} = (\text{SR} + \text{PR} + \text{IPA})/3$. Our return metrics map VIA to IPA, IIR to jointly passing SR and PR, and LS to passing all three within the same episode.

Qwen pools 200 retained episodes (Appendix E.2); Gemini and DeepSeek weight the three strata equally. The separate CUPMem subset study is described in Appendix E.3. Memora and Old-Valid ManBench-Return complement STALE by testing retention of departure information that remains useful.

D.5 MANBENCH-RETURN

ManBench studies whether interaction with other agents can induce an initially correct agent to adopt an incorrect belief (Xu et al., 2026). Our adaptation has two conditions: Old-Valid tests resistance to misleading updates when the departure state remains correct, and Old-Stale tests adaptation to an absence-period task change. Together, they test whether admission follows current validity rather than information age.

The experiment schedules 4,838 examples from 20 tasks. Alternating positions within each task determine the conditions before model outputs are observed. In Old-Valid, the question remains unchanged while active agents promote the benchmark's misleading answer. In Old-Stale, the question changes to the next example within the same task, cyclically, and active agents solve it independently. Return candidates comprise the departure answer and the absence-period challenger; their validity depends on the condition.

Eligible episodes require a correct departure answer. Old-Stale additionally requires at least three of the four active agents to agree on the correct current answer. These criteria are checked by the evaluator without exposing answer labels to the returning agent or verifier. Eligible sample sizes therefore depend on the actor and condition. Overall measures final-answer correctness, VIA measures admission of the valid candidate, and IIR measures exclusion of the invalid candidate. WSAR is the complementary admission rate for invalid state. The adapted RSR applies to Old-Valid, where the question remains unchanged (Appendix E.4). We give the two conditions equal weight when reporting aggregate results.

D.6 Additional experiments

The scale study compares Static and TRACE on all 4,838 source examples with 4, 8, 16, or 32 active agents and one returning agent. The portability study compares these methods on ScopedMem, Mem0, Memobase, and A-MEM at 50% Memora departure, holding the backend, actor, scorer, and questions fixed within each pair.

E Detailed Results: Qwen

E.1 Memora

Table 5 reports each departure condition and its macro-average. Results are reported to two decimal places; averaging already rounded condition scores can differ in the last digit.

Table 5: Qwen on Memora at three departure points (%). Average is the equal-weight macro-average of the three conditions. Bold marks the best reported value among the six methods in each complete block.

Departure	Method	Overall↑	VIA↑	IIR↑	FAMA↑
25%	Static	62.10	53.91	67.94	37.88
	Restore	65.74	41.12	85.16	34.46
	Reset	**72.47**	39.19	**99.90**	39.19
	MemStrata	65.55	54.22	73.68	40.87
	MemTX	65.38	**54.72**	73.46	**41.15**
	TRACE	63.98	52.35	70.18	38.75
50%	Static	63.38	**55.04**	67.77	39.13
	Restore	64.57	41.35	81.39	33.00
	Reset	**72.30**	38.46	**100.00**	38.46
	MemStrata	64.62	54.17	70.41	**39.92**
	MemTX	64.80	53.39	71.18	39.49
	TRACE	65.66	52.01	74.19	39.31
75%	Static	64.00	57.21	68.68	41.43
	Restore	65.01	43.85	80.17	35.01
	Reset	**72.44**	38.99	**99.90**	38.99
	MemStrata	68.12	58.20	74.98	45.15
	MemTX	68.46	**58.75**	76.06	**45.78**
	TRACE	65.96	56.45	72.71	42.34
Average	Static	63.16	55.39	68.13	39.48
	Restore	65.11	42.11	82.24	34.16
	Reset	**72.40**	38.88	**99.93**	38.88
	MemStrata	66.10	55.53	73.02	41.98
	MemTX	66.21	**55.62**	73.57	**42.14**
	TRACE	65.20	53.60	72.36	40.14

E.2 STALE TYPE II

Table 6 reports Early, Middle, and Late separately. The full STALE tables use the same metric order: Overall, Dim1 (SR), Dim2 (PR), VIA / Dim3 (IPA), IIR, and LS. Dim1 and Dim2 are computed from the retained per-episode probe judgments.

This benchmark runs 200 episodes: 64 in Early, 64 in Middle, and 72 in Late. All 200 episodes are retained in every bucket, so each block averages its full bucket and no episode is dropped from any denominator. Pooled averages the 200 retained episodes and reproduces the Qwen scores in Table 1; it does not weight the three buckets equally. TRACE's 39.00% pooled LS exceeds the strongest memory-policy baseline, MemTX at 18.00%, by 21.00 percentage points. For the same evaluation, the retained inference audit reports that the high-recall stage proposed 5,610 possible event–memory pairs and the grounded verifier accepted 385 (6.9%). These counts describe verification-stage selectivity within the evaluation; they do not change the 200-episode score denominators above or by themselves measure dependency-recall accuracy.

Table 6: Qwen on STALE Type II (%). Early, Middle, and Late contain 64, 64, and 72 episodes, respectively, and every episode is retained, so all four blocks report the full sample. Pooled averages the 200 episodes with each bucket weighted by its episode count. MemStrata and MemTX results for Early, Middle, and Late buckets are omitted due to data inconsistencies; only Pooled results are reported. Bold marks the best reported value among the six methods in each block.

Bucket	Method	Overall↑	Dim1↑	Dim2↑	VIA / Dim3↑	IIR↑	LS↑
Early	Static	35.94	32.81	32.81	42.19	28.12	21.88
	Restore	5.73	1.56	1.56	14.06	1.56	0.00
	Reset	17.19	3.12	39.06	9.38	3.12	0.00
	TRACE	**64.06**	**65.62**	**68.75**	**57.81**	**65.62**	**48.44**
Middle	Static	24.48	21.88	21.88	29.69	18.75	12.50
	Restore	6.25	1.56	3.12	14.06	0.00	0.00
	Reset	15.62	3.12	28.12	15.62	3.12	1.56
	TRACE	**49.48**	**48.44**	**46.88**	**53.12**	**45.31**	**35.94**
Late	Static	25.00	22.22	19.44	33.33	19.44	13.89
	Restore	5.56	1.39	1.39	13.89	1.39	0.00
	Reset	14.81	2.78	29.17	12.50	2.78	0.00
	TRACE	**44.44**	**44.44**	**41.67**	**47.22**	**41.67**	**33.33**
Pooled	Static	28.33	35.00	22.33	35.00	22.33	17.67
	Restore	5.75	14.00	1.00	14.00	1.00	0.00
	Reset	15.83	12.50	3.00	12.50	3.00	0.50
	MemStrata	28.66	32.00	22.50	32.00	22.50	17.00
	MemTX	30.00	33.50	23.50	33.50	23.50	18.00
	TRACE	**52.33**	**52.50**	**52.00**	**52.50**	**50.50**	**39.00**

E.3 STALE SUBSET: NATIVE METRICS AND TOKEN COST

Rationale. CUPMem maintains a structured current-state store through repeated session-level processing: evidence extraction, local state updates, affected-state discovery, and LLM adjudication of possible invalidations (Chao et al., 2026). These stages add memory-maintenance work before query answering. We use a separate subset to keep this full-pipeline comparison tractable while retaining all three departure buckets. The study compares CUPMem's accuracy and the token use of its complete pipeline with TRACE and the other memory policies.

Table 7 expands the separate Qwen subset study in Table 2. Early, Middle, and Late follow the departure boundary in the original chronology, rather than fixed memory-age thresholds. Overall is the mean accuracy over the three native probes: State Resolution (SR), Premise Resistance (PR), and Implicit Policy Adaptation (IPA). The table uses Dim1, Dim2, and VIA / Dim3 for these probes, as in the full STALE tables. IIR requires SR and PR to pass in the same episode, and LS requires all three to pass. IPA is identical to VIA under this adapter, so it is shown once.

One failed CUPMem episode. One Early episode of the CUPMem subset never returned an answer. All retry rounds exceeded the serving context window: the budget adapter submitted 32,752 input tokens against a 32,768-token window while reserving 3,072 tokens for the completion, so each round failed with a context-window error and the final round ended in a parse error. Repeating the episode under the same configuration failed again, so the failure is deterministic rather than incidental. We keep the planned denominator and score the episode as a failure on all three probes, which lowers CUPMem's Early Overall from 85.96% to 81.67% and its Early LS from 73.68% to 70.00%. Scoring the same episode against the 19 episodes that did complete would leave CUPMem at 85.96% Overall, so every CUPMem accuracy in Table 7 is computed on the full 20-episode Early denominator. The failed episode is the most expensive single episode in the subset: its three retry rounds billed 5,971,252 tokens, about $3.1\times$ the average successful episode, which raises CUPMem's aggregate from 1,927,417 to 1,994,815 tokens per episode and its ratio to TRACE from $3.86\times$ to $3.99\times$. Its accuracy cells are therefore conservative and its cost cells are not: retaining the failed episode makes CUPMem both less accurate and more expensive than a scored-only comparison would suggest. This failure is itself an instance of the cost asymmetry this subset examines: full-context injection has no token bound, while the serving window does, so a single long episode can consume three times the average budget and still exhaust the window without producing an answer, whereas TRACE's bounded Return View stays inside it.

Table 7: Qwen on the STALE Type II subset by departure bucket. Accuracy and joint metrics are percentages. Aggregate reproduces each method's reported summary. Every method is scored over the same 20 planned episodes per bucket, including one CUPMem episode that could not complete and is scored as a failure on all three probes (see the note above); its retry cost is included in the token column. Bold marks the highest score within each block; bold token counts highlight CUPMem's higher cost.

Bucket	Method	Overall↑	Dim1↑	Dim2↑	VIA / Dim3↑	IIR↑	LS↑	Tokens/episode↓
Early	Static	36.67	30.00	25.00	55.00	20.00	15.00	219,855
	MemStrata	23.33	10.00	25.00	35.00	10.00	5.00	218,115
	MemTX	23.33	15.00	20.00	35.00	10.00	5.00	217,662
	CUPMem	**81.67**	**80.00**	**85.00**	**80.00**	**80.00**	**70.00**	**2,127,613**
	TRACE (Ours)	73.33	80.00	60.00	80.00	60.00	55.00	316,750
Middle	Static	36.67	30.00	35.00	45.00	30.00	15.00	215,073
	MemStrata	50.00	40.00	60.00	50.00	40.00	30.00	212,723
	MemTX	46.67	40.00	60.00	40.00	40.00	25.00	212,732
	CUPMem	**90.00**	**85.00**	**90.00**	**95.00**	**85.00**	**85.00**	**1,935,913**
	TRACE (Ours)	60.00	60.00	65.00	55.00	60.00	35.00	517,366
Late	Static	26.67	25.00	20.00	35.00	15.00	10.00	219,754
	MemStrata	20.00	15.00	25.00	20.00	10.00	5.00	217,669
	MemTX	20.00	10.00	25.00	25.00	10.00	10.00	217,624
	CUPMem	**70.00**	**80.00**	**80.00**	**50.00**	**80.00**	**35.00**	**1,920,918**
	TRACE (Ours)	43.33	45.00	40.00	45.00	40.00	30.00	664,860
Aggregate	Static	33.33	28.33	26.67	45.00	21.67	13.33	218,227
	MemStrata	31.11	21.67	36.67	35.00	20.00	13.33	216,169
	MemTX	30.00	21.67	35.00	33.33	20.00	13.33	216,006
	CUPMem	**80.56**	**81.67**	**85.00**	**75.00**	**81.67**	**63.33**	**1,994,815**
	TRACE (Ours)	58.89	61.67	55.00	60.00	53.33	40.00	499,659

The three buckets show a change in both accuracy and cost. TRACE's Overall falls from 73.33% in Early to 43.33% in Late, while its reported token count rises from 316,750 to 664,860 per episode. CUPMem's count remains between 1.92 and 2.13 million tokens across the buckets. The lowest bucket LS is 10.00% for Static, 5.00% for MemStrata, 5.00% for MemTX, 35.00% for CUPMem, and 30.00% for TRACE. These joint scores distinguish passing individual probes from resolving all three requirements within an episode.

Token counts include memory processing and answering, with return-governance calls included for TRACE; external benchmark-scoring calls are excluded. All methods use the recorded billed-token counter, so the counts are measured on a common basis, and CUPMem's include the three failed retry rounds of the episode described above. They are not measurements of latency, monetary cost, or peak context.

E.4 MANBENCH-RETURN

In Old-Valid, the departure state remains valid while other agents propagate an incorrect state during the absence interval. In Old-Stale, the departure state is invalid and the absence-period update supplies valid state. Table 8 separates the two conditions and reports their equal-weight mean.

WSAR measures admission of the incorrect state to the return context. We also adapt ManBench's Reality Shift Rate (RSR) (Xu et al., 2026) to measure a correct-to-incorrect answer transition at return. Let Q_B^+ be the baseline-correct Old-Valid episodes and Q_R^- those with an incorrect final answer after return. Then

$$\mathrm{RSR} = \frac{|Q_B^+ \cap Q_R^-|}{|Q_B^+|}.$$

Old-Valid eligibility requires a correct baseline answer, and the question remains unchanged. RSR is therefore $100 - \mathrm{Overall}$ in percentage units; the entries below are derived from the reported Old-Valid Overall values. Old-Stale changes the task, so it does not measure the same correct-to-incorrect transition. We mark RSR as n/a for Old-Stale and the cross-condition Average.

Table 8: Qwen on ManBench-Return (%). Average gives equal weight to Old-Valid and Old-Stale using unrounded condition scores; displayed condition values may not reproduce the final decimal. Overall measures final answer correctness; RSR applies only to Old-Valid. Bold marks the best reported value among the six methods in each complete block.

Condition	Method	Overall↑	VIA↑	IIR↑	WSAR↓	RSR↓
Old-Valid	Static	70.06	0.00	0.00	100.00	29.94
	Restore	**96.30**	**100.00**	**100.00**	**0.00**	**3.70**
	Reset	95.75	0.00	**100.00**	**0.00**	4.25
	MemStrata	70.25	0.00	0.00	100.00	29.75
	MemTX	70.86	0.00	0.00	100.00	29.14
	TRACE	93.37	95.75	97.57	2.43	6.63
Old-Stale	Static	98.68	**100.00**	**100.00**	**0.00**	n/a
	Restore	93.80	0.00	0.00	100.00	n/a
	Reset	97.01	0.00	**100.00**	**0.00**	n/a
	MemStrata	98.61	**100.00**	**100.00**	**0.00**	n/a
	MemTX	**98.81**	**100.00**	**100.00**	**0.00**	n/a
	TRACE	95.26	96.17	**100.00**	**0.00**	n/a
Average	Static	84.38	50.00	50.00	50.00	n/a
	Restore	95.05	50.00	50.00	50.00	n/a
	Reset	**96.38**	0.00	**100.00**	**0.00**	n/a
	MemStrata	84.43	50.00	50.00	50.00	n/a
	MemTX	84.84	50.00	50.00	50.00	n/a
	TRACE	94.32	**95.96**	98.78	1.22	n/a

F Detailed Results: Gemini

F.1 Memora

Table 9 reports the three departure conditions and their equal-weight macro-average for the Gemini actor.

Table 9: Gemini on Memora at three departure points (%). Average is the equal-weight macro-average of the three conditions. Bold marks the best reported value among the six methods in each complete block.

Departure	Method	Overall↑	VIA↑	IIR↑	FAMA↑
25%	Static	66.50	52.62	75.50	40.58
	Restore	67.62	43.63	85.96	37.23
	Reset	**71.71**	43.44	**95.04**	41.48
	MemStrata	67.95	53.47	76.95	42.37
	MemTX	68.12	**53.56**	76.97	**42.53**
	TRACE	64.93	51.11	73.31	38.42
50%	Static	67.97	54.45	75.92	42.41
	Restore	66.40	44.67	83.70	37.20
	Reset	**72.10**	43.45	**94.33**	41.39
	MemStrata	67.54	**56.59**	73.99	43.18
	MemTX	67.99	55.63	75.19	43.32
	TRACE	68.71	55.14	76.76	**43.51**
75%	Static	68.41	56.32	76.82	44.93
	Restore	66.59	46.69	80.62	38.41
	Reset	**72.20**	43.03	**96.32**	41.60
	MemStrata	68.00	56.52	76.27	44.38
	MemTX	69.19	**58.26**	76.11	**46.50**
	TRACE	69.45	56.98	79.14	46.47
Average	Static	67.63	54.46	76.08	42.64
	Restore	66.87	45.00	83.43	37.61
	Reset	**72.00**	43.31	**95.23**	41.49
	MemStrata	67.83	55.53	75.74	43.31
	MemTX	68.43	**55.81**	76.09	**44.12**
	TRACE	67.70	54.41	76.41	42.80

F.2 STALE TYPE II

Table 10 reports the three buckets and their equal-weight macro-average. Dim1 denotes State Resolution and Dim2 Premise Resistance; VIA corresponds to Implicit Policy Adaptation (Dim3) (Chao et al., 2026). Table 1 uses the averaged Overall, VIA, and IIR values. As with DeepSeek, this aggregate weights buckets equally; the Qwen aggregate instead pools episodes. Scores retain the reported two-decimal precision.

Table 10: Gemini on STALE Type II (%). Early, Middle, and Late contain 64, 64, and 72 episodes, respectively. Average gives equal weight to the three buckets. MemStrata and MemTX results for Early, Middle, and Late buckets are omitted due to data inconsistencies; only Average results are reported. Bold marks the best reported value among the six methods in each block.

Bucket	Method	Overall↑	Dim1↑	Dim2↑	VIA / Dim3↑	IIR↑	LS↑
Early	Static	78.13	73.44	78.13	82.81	71.88	64.06
	Restore	9.90	0.00	6.25	23.44	0.00	0.00
	Reset	8.33	0.00	6.25	18.75	0.00	0.00
	TRACE	**91.15**	**92.19**	**92.19**	**89.06**	**92.19**	**87.50**
Middle	Static	70.31	70.31	64.06	76.56	62.50	60.94
	Restore	8.33	1.56	3.13	20.31	1.56	1.56
	Reset	9.38	1.56	4.69	21.88	1.56	1.56
	TRACE	**90.10**	**89.06**	**90.63**	**90.63**	**89.06**	**85.94**
Late	Static	62.50	59.72	58.33	69.44	56.94	52.78
	Restore	5.09	0.00	1.39	13.89	0.00	0.00
	Reset	8.80	0.00	5.56	20.83	0.00	0.00
	TRACE	**85.19**	**83.33**	**81.94**	**90.28**	**81.94**	**79.17**
Average	Static	70.31	67.82	66.84	76.27	63.77	59.26
	Restore	7.77	0.52	3.59	19.21	0.52	0.52
	Reset	8.83	0.52	5.50	20.49	0.52	0.52
	MemStrata	64.80	60.24	59.26	74.88	56.71	55.67
	MemTX	62.29	58.80	58.28	69.79	55.73	52.72
	TRACE	**88.81**	**88.19**	**88.25**	**89.99**	**87.73**	**84.20**

F.3 MANBENCH-RETURN

Table 11 reports Gemini's Old-Valid and Old-Stale results and their equal-weight average, following the definitions in Appendix D.5. Overall measures final-answer correctness; VIA and IIR assess valid-state admission and invalid-state rejection, respectively. WSAR is the complementary wrong-state admission rate.

Table 11: Gemini on ManBench-Return (%). Average gives equal weight to Old-Valid and Old-Stale using unrounded condition scores; displayed condition values may not reproduce the final decimal. Overall measures final-answer correctness. Bold marks the best value among the six methods in each block.

Condition	Method	Overall↑	VIA↑	IIR↑	WSAR↓
Old-Valid	Static	76.19	0.00	0.00	100.00
	Restore	**99.75**	**100.00**	**100.00**	**0.00**
	Reset	97.34	0.00	**100.00**	**0.00**
	MemStrata	73.71	0.00	0.00	100.00
	MemTX	73.35	0.00	0.00	100.00
	TRACE	97.84	97.89	98.95	1.05
Old-Stale	Static	99.88	**100.00**	**100.00**	**0.00**
	Restore	97.05	0.00	0.00	100.00
	Reset	98.15	0.00	**100.00**	**0.00**
	MemStrata	99.88	**100.00**	**100.00**	**0.00**
	MemTX	**99.94**	**100.00**	**100.00**	**0.00**
	TRACE	98.79	98.79	**100.00**	**0.00**
Average	Static	88.04	50.00	50.00	50.00
	Restore	**98.40**	50.00	50.00	50.00
	Reset	97.75	0.00	**100.00**	**0.00**
	MemStrata	86.79	50.00	50.00	50.00
	MemTX	86.65	50.00	50.00	50.00
	TRACE	98.32	**98.34**	99.47	0.53

G Detailed Results: DeepSeek

G.1 Memora

Table 12 reports the three departure conditions and their equal-weight macro-average for the DeepSeek actor. Scores are reported to two decimal places; averaging rounded condition scores can differ in the last digit.

Table 12: DeepSeek on Memora at three departure points (%). Average is the equal-weight macro-average of the three conditions. Bold marks the best reported value among the six methods in each block.

Departure	Method	Overall↑	VIA↑	IIR↑	FAMA↑
25%	Static	66.66	49.88	78.02	39.72
	Restore	69.22	41.02	91.77	37.57
	Reset	**72.52**	40.39	**99.19**	39.95
	MemStrata	68.43	51.78	79.54	42.24
	MemTX	68.69	**52.39**	79.64	**42.66**
	TRACE	67.83	50.71	79.67	40.77
50%	Static	67.51	51.99	77.79	41.50
	Restore	69.50	42.30	90.62	38.45
	Reset	**72.79**	39.98	**99.59**	39.82
	MemStrata	68.36	**53.02**	78.15	42.35
	MemTX	69.69	52.86	80.63	**43.39**
	TRACE	68.06	50.88	79.15	40.99
75%	Static	70.93	55.53	84.81	47.30
	Restore	70.71	43.48	91.00	39.79
	Reset	**72.21**	39.53	**99.04**	39.13
	MemStrata	71.15	55.89	82.33	47.39
	MemTX	71.28	**56.28**	82.75	**48.12**
	TRACE	70.61	54.00	83.33	45.99
Average	Static	68.36	52.47	80.20	42.84
	Restore	69.81	42.27	91.13	38.60
	Reset	**72.51**	39.97	**99.27**	39.63
	MemStrata	69.31	53.57	80.01	43.99
	MemTX	69.89	**53.84**	81.01	**44.72**
	TRACE	68.83	51.86	80.72	42.58

G.2 STALE TYPE II

Table 13 separates the Early, Middle, and Late buckets and reports their equal-weight macro-average. Table 1 uses the Overall, VIA, and IIR entries from this Average block. The DeepSeek aggregate weights buckets equally despite their different sizes; the Qwen aggregate instead pools episodes. Cross-model score differences therefore may also reflect the aggregation rule.

Table 13: DeepSeek on STALE Type II (%). Early, Middle, and Late contain 64, 64, and 72 episodes, respectively. Average gives equal weight to the three buckets. MemStrata and MemTX results for Early, Middle, and Late buckets are omitted due to data inconsistencies; only Average results are reported. Bold marks the best reported value among the six methods in each block.

Bucket	Method	Overall↑	Dim1↑	Dim2↑	VIA / Dim3↑	IIR↑	LS↑
Early	Static	25.00	17.19	17.19	40.63	12.50	12.50
	Restore	6.25	0.00	3.13	15.63	0.00	0.00
	Reset	13.02	0.00	14.06	25.00	0.00	0.00
	TRACE	**71.88**	**76.56**	**70.31**	**68.75**	**70.31**	**59.38**
Middle	Static	24.48	20.31	20.31	32.81	15.63	9.38
	Restore	6.25	0.00	1.56	17.19	0.00	0.00
	Reset	17.19	0.00	28.13	23.44	0.00	0.00
	TRACE	**46.88**	**43.75**	**42.19**	**54.69**	**40.63**	**37.50**
Late	Static	16.67	9.72	11.11	29.17	9.72	6.94
	Restore	5.56	1.39	0.00	15.28	0.00	0.00
	Reset	18.06	0.00	31.94	22.22	0.00	0.00
	TRACE	**44.44**	**41.67**	**41.67**	**50.00**	**38.89**	**31.94**
Average	Static	22.05	15.74	16.20	34.20	12.62	9.61
	Restore	6.02	0.46	1.56	16.03	0.00	0.00
	Reset	16.09	0.00	24.71	23.55	0.00	0.00
	MemStrata	26.00	16.15	25.64	36.23	14.58	11.63
	MemTX	26.89	19.21	24.71	36.75	14.58	10.07
	TRACE	**54.40**	**53.99**	**51.39**	**57.81**	**49.94**	**42.94**

G.3 MANBENCH-RETURN

Table 14 reports both conditions and their equal-weight average for all six methods. The condition definitions match the Qwen evaluation in Appendix E.4. WSAR measures admission of incorrect state to the return context.

Table 14: DeepSeek on ManBench-Return (%). Average gives equal weight to Old-Valid and Old-Stale using unrounded condition scores; displayed condition values may not reproduce the final decimal. Overall measures final answer correctness. Bold marks the best reported value among the six methods in each block.

Condition	Method	Overall↑	VIA↑	IIR↑	WSAR↓
Old-Valid	Static	31.06	0.00	0.00	100.00
	Restore	**99.88**	**100.00**	**100.00**	**0.00**
	Reset	95.39	0.00	**100.00**	**0.00**
	MemStrata	52.50	0.00	0.00	100.00
	MemTX	52.38	0.00	0.00	100.00
	TRACE	92.04	91.86	96.89	3.11
Old-Stale	Static	99.85	**100.00**	**100.00**	**0.00**
	Restore	87.13	0.00	0.00	100.00
	Reset	97.38	0.00	**100.00**	**0.00**
	MemStrata	99.85	**100.00**	**100.00**	**0.00**
	MemTX	**100.00**	**100.00**	**100.00**	**0.00**
	TRACE	93.22	93.30	**100.00**	**0.00**
Average	Static	65.45	50.00	50.00	50.00
	Restore	93.51	50.00	50.00	50.00
	Reset	**96.39**	0.00	**100.00**	**0.00**
	MemStrata	76.18	50.00	50.00	50.00
	MemTX	76.19	50.00	50.00	50.00
	TRACE	92.63	**92.58**	98.44	1.56

H Component Ablations on STALE Type II

This study examines four mechanisms in TRACE using Qwen on STALE Type II. Each variant removes one mechanism from the full admission procedure. The abbreviations II, OAS, CVG, and SBPR are defined in Section 4.4 and identify the mechanisms removed by each variant. Figure 3 presents Overall, VIA, and IIR; Table 15 also reports LS, which requires all three probes to pass in the same episode. The metric definitions follow Section 4.3.

Table 15: Component ablations on STALE Type II with Qwen (%). Bold marks the highest score in each column. Each ablation removes one mechanism.

Method	Overall↑	VIA↑	IIR↑	LS↑
Full TRACE	**52.33**	**52.50**	**50.50**	**39.00**
w/o II	26.50	34.00	19.50	18.00
w/o OAS	38.67	35.50	36.00	29.00
w/o CVG	47.33	48.00	44.00	35.00
w/o SBPR	42.00	43.00	38.50	31.00

w/o II. This variant disables cross-attribute implicit invalidation while retaining explicit updates, revocations, and the remaining admission checks. It tests whether later observations must be used to reconsider an earlier memory even when they do not explicitly retract it. Its IIR decreases from 50.50% to 19.50%, and LS decreases from 39.00% to 18.00%. This is the largest decline among the four ablations on both metrics.

w/o OAS. This variant replaces obligation-aware evidence selection with relevance ranking under the same context budget. The final coverage check remains active. It tests whether selecting individually relevant records supplies the evidence needed for the full returning workload. Overall falls from 52.33% to 38.67%, while VIA falls from 52.50% to 35.50%.

w/o CVG. This variant retains the evidence selection procedure but allows the returning workload to proceed when critical-obligation coverage is incomplete. Identity, provenance, and budget checks remain active. It isolates the decision to withhold an incomplete view from the preceding selection step. Overall decreases by 5.00 percentage points and LS by 4.00 percentage points. These aggregate results support the gate's contribution in this study; they do not imply that every rejected view would otherwise yield an incorrect answer.

w/o SBPR. This variant leaves the public Return View unchanged but removes the requirement that retrieved private experiences be bound to admitted sources. Owner isolation remains active. It tests whether information excluded from the public view can affect the answer through private-memory retrieval. Overall falls to 42.00%, IIR to 38.50%, and LS to 31.00%. The reduction is consistent with the need to apply admission constraints to both public work state and private experiences.

I AGENT-SCALE AND BACKEND STUDIES

These studies examine two distinct questions: whether return-time admission remains stable as the team grows, and whether the same admission boundary can operate above different memory representations. The first varies the number of active agents in ManBench-Return; the second varies the memory backend in Memora while holding the return condition fixed.

I.1 AGENT-SCALE DIAGNOSTIC

We vary the number of active agents over 4, 8, 16, and 32 while keeping one returning agent. Static and TRACE use the same balanced Old-Valid/Old-Stale construction at each scale. Overall measures final-answer correctness; VIA and IIR distinguish preservation of valid state from rejection of invalid state. Table 3 gives the scale-specific results, and Table 16 summarizes them with equal weight per scale.

Table 16: Macro-average across the four agent scales in Table 3 (%). WSAR is the complement of IIR in this admission diagnostic. Bold marks the better value in each column.

Method	Overall↑	VIA↑	IIR↑	WSAR↓
Static	76.51	44.24	44.24	55.76
TRACE	**96.07**	**96.76**	**98.83**	**1.17**

TRACE's Overall ranges from 95.85% to 96.21%, VIA from 96.62% to 96.92%, and IIR from 98.72% to 98.94%. The corresponding macro-average gains over Static are 19.56, 52.52, and 54.59 percentage points. These results support stability within the evaluated range, without a monotonic increase in performance or advantage as the team grows. This full-coverage diagnostic is separate from the frozen five-agent comparison in Table 1.

I.2 BACKEND INTEGRATION AND PAIRED EVALUATION

What changes with the backend. ScopedMem is our agent-scoped versioned reference store. Mem0 extracts, consolidates, and retrieves conversational memories (Chhikara et al., 2025); Memobase organizes user profiles and events; A-MEM constructs linked notes whose attributes and connections can evolve as new memories arrive (Xu et al., 2025). The adapters preserve these native formation and retrieval operations. TRACE controls eligibility at return through a common representation of record identity, text, retrieval score, and source bindings. It does not replace each backend with a shared store or use reference-store text as a fallback for native retrieval.

Admission across different representations. Backend substitution requires more than a common search API: a retrieved profile or linked note may combine several historical inputs. The adapter must retain the evidence associations needed to decide whether that memory can enter the returning agent's context. A source association identifies origin; it does not establish current validity. Where the native system does not expose exact dependencies, conservative source associations can restrict retention. Thus the granularity of source attribution is an integration constraint, and native retrieval relevance cannot substitute for TRACE's validity decision. Appendix C details the source-bound readmission rule.

Paired protocol. The study uses the 137 Memora episodes at 50% departure with Qwen3.5-122B-A10B as the actor and `gpt-5.6-terra` as the scorer. Within each backend, Static and TRACE share the episode history and pre-return memory formation. Static follows the ordinary retrieval path; TRACE restricts readmission using the admitted work state. Branch construction imports the stored records without repeating memory formation, and retrieval remains native to the selected backend. This comparison measures the effect of attaching TRACE within each backend. It does not require identical retrieved text across backends, whose memory construction and ranking differ.

Table 17: Backend portability on Memora at 50% departure (%). Each Static/TRACE pair shares its backend, actor, and scorer. Bold marks the higher value within each pair.

Backend	Method	VIA↑	IIR↑	FAMA↑
ScopedMem	Static	**55.04**	67.77	39.13
	TRACE	52.01	**74.19**	**39.31**
Mem0	Static	**53.53**	68.93	37.81
	TRACE	53.48	**70.06**	**38.29**
Memobase	Static	**54.54**	67.94	**38.52**
	TRACE	53.19	**68.33**	37.17
A-MEM	Static	**55.33**	67.24	**39.21**
	TRACE	54.00	**68.35**	38.09

Results and interpretation. In Table 17, TRACE increases IIR by 6.42, 1.13, 0.39, and 1.11 percentage points for ScopedMem, Mem0, Memobase, and A-MEM, respectively. VIA decreases by 3.03, 0.05, 1.35, and 1.33 percentage points. FAMA changes by $+0.18$, $+0.48$, -1.35, and -1.12 percentage points: greater rejection of invalid information does not consistently improve the combined remembering-and-forgetting score. The paired results support the feasibility of applying return-time admission above all four tested backends, while showing a retention–rejection trade-off on Memora's explicit updates. They do not isolate which representation or source-binding choice causes a particular gain or loss. The results also do not establish compatibility with arbitrary stores, or comparative backend latency and token efficiency.