MEMPERMIT: GOVERNING WRITES TO MULTI-AGENT SHARED MEMORY VIA RISK-CONSTRAINED SEQUENTIAL VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Shared long-term memory lets language-model agents reuse information, but each write commits both content and an audience. Existing admission filters, read-time access control, and fixed verification committees govern these choices separately. We introduce MEMPERMIT, a write-time controller that selects persistence and a non-expanding reader audience under one prespecified, asymmetric risk profile. Risk triggers mandate specialist coverage; afterward, verification continues only when its expected reduction in action risk exceeds cost. On GateMem with MIRIX, the complete controller lowers access-control violations over read-time governance by 5.0 percentage points with GPT-4o-mini and 4.5 with DeepSeek-V4-Pro while meeting prespecified utility and over-refusal criteria. With identical specialist evidence, its joint terminal rule has a 1.2-point lower ACV estimate than an audience-blind two-stage rule. On held-out query-only MINJA attacks, it lowers attack success over its content-only detector by 9.7 and 5.4 points. Both model arms meet the declared -3-point judge-accuracy criterion on LoCoMo. Sequential verification reduces cost per candidate by 27% relative to a fixed committee and by 9–10% relative to one-shot routing under matched risk constraints.

1 INTRODUCTION

Long-term memory extends a language-model agent beyond its active context. Systems retain interaction traces and useful facts for later decisions (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024; Xu et al., 2025; Chhikara et al., 2025). In a shared store, however, a write has two additional consequences: it can outlive its originating task, and it can be consumed by principals operating under different permissions. A malformed write may become a durable instruction, propagate a stale claim, or expose a private fact.

These failures meet at a single boundary: candidate information is converted into persistent shared state. Memory-poisoning attacks exploit this boundary by placing adversarial content directly in a store or inducing an agent to write it through ordinary interaction (Chen et al., 2024b; Dong et al., 2025). Indirect prompt injection demonstrates the broader difficulty of separating data from instructions (Greshake et al., 2023; Yi et al., 2025). Access-control systems can restrict retrieval, but they assume that a useful object and its policy already exist (Sandhu et al., 1996; Hu et al., 2014). A binary write filter, in turn, does not decide which authorized subset should receive an admitted record. Applying these mechanisms independently can retain a record for an unusable audience or make an admission decision without the evidence needed to scope it safely.

We govern a shared-memory write through two coupled decisions. The terminal decision selects persistence and a non-expanding reader audience; the acquisition decision purchases specialist evidence only when it can change that terminal action. MEMPERMIT implements this design in front of an unmodified memory backend (Figure 1). Lightweight signals initialize marginal risks for harm, source reliability, conflict, retention value, task mismatch, and reader-specific authorization. RuleGuard maps observable risk triggers to specialists that must succeed before an ordinary terminal action. After this coverage obligation is met, a sequential expected value of verification (EVoV) rule selects additional evidence when its expected reduction in action risk justifies cost. A joint aggregator then chooses whether to persist and, if so, the concrete audience. The stored audience is

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.

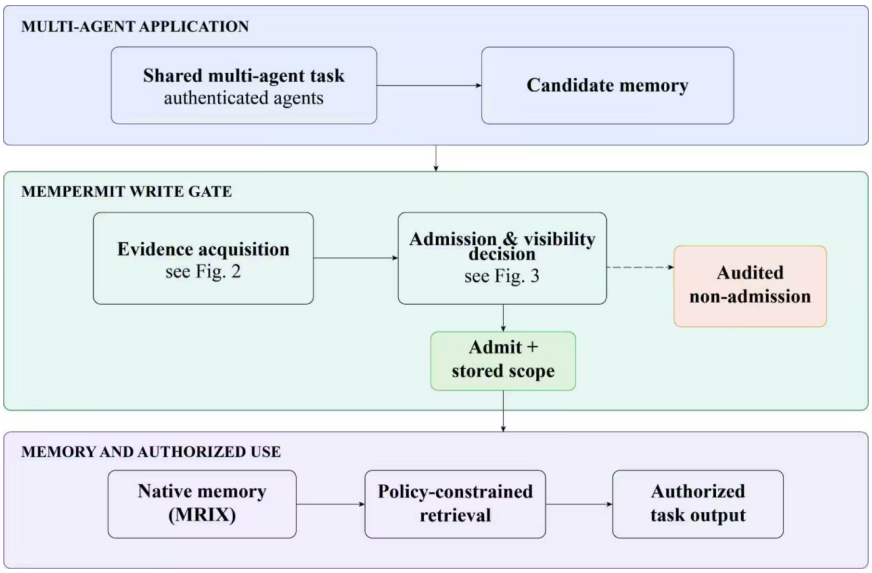


Figure 1: MEMPERMIT sits at the shared-memory write boundary. It converts a candidate and policy state into a joint persistence–audience decision. Retrieval intersects the stored scope with current policy, so later policy expansion cannot widen visibility.

an immutable upper bound; retrieval intersects it with current policy and fails closed on unresolved identity.

The asymmetric loss is an inspectable governance configuration whose ratios are fixed before held-out evaluation and shared across folds and model arms. We test whether the controller realizes this declared harm–disclosure–utility preference.

GateMem (Ren et al., 2026) evaluates multi-principal utility and access control; held-out MINJA (Dong et al., 2025) tests query-only injection; LoCoMo (Maharana et al., 2024) measures benign long-term utility; and matched routing traces compare EVoV with fixed and one-shot acquisition. GPT-4o-mini is primary, and DeepSeek-V4-Pro repeats the protocol. A matched-evidence terminal ablation gives joint, admission-only, scope-only, and audience-blind rules the same final belief; its paired contrast tests whether reader-specific disclosure risk feeds back into persistence.

The contributions are:

- a formulation of shared-memory commitment as an asymmetric-risk decision over persistence and a concrete, non-expanding audience;
- a verification policy combining trigger-based mandatory coverage, cost-aware sequential evidence acquisition, and one aggregator for all terminal paths; and
- a two-model evaluation of governance, poisoning, benign utility, and matched-risk cost, including a matched-evidence comparison that isolates the terminal persistence–audience rule from specialist evidence collection.

2 RELATED WORK

Agent memory. Long-term-memory systems retain interaction history and manage external context (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024); recent systems add linked notes, explicit operations, cross-trial experience, or specialized stores (Xu et al., 2025; Chhikara et al., 2025; Zhang et al., 2025; Wang & Chen, 2025). Where these systems optimize storage or retrieval, MEMPERMIT governs evidence, persistence, and audience at commit time without changing backend ranking.

Shared-memory governance. Shared-memory proposals combine private/shared tiers, scoped retrieval, provenance, and rule-based privacy or conflict handling (Rezazadeh et al., 2025; Margalit et al., 2026; Kumar et al., 2026). GateMem evaluates contextual access control and forgetting (Ren et al., 2026); RBAC and ABAC provide policy vocabulary (Sandhu et al., 1996; Hu et al., 2014), and HONEYBEE accelerates role-aware search (Zhong et al., 2026). MEMPERMIT instead jointly decides commitment, audience, and further evidence.

Memory safety and adaptive verification. Persistent state can carry attacks into later queries or expose stored interactions (Chen et al., 2024b; Dong et al., 2025; Wang et al., 2025); A-MemGuard adds consensus validation (Wei et al., 2026). Separate harm and per-reader authorization risks distinguish public adversarial content from harmless unauthorized content. MEMPERMIT uses verified evidence to choose a persistence–audience action with a stored visibility bound. Selective prediction, calibration, and cost-aware routing motivate abstention and adaptive computation (Geifman & El-Yaniv, 2019; Guo et al., 2017; Chen et al., 2024a; Aggarwal et al., 2024; Ong et al., 2025; Dekoninck et al., 2025). We instead acquire typed specialist observations valued by their effect on the write action.

3 PROBLEM FORMULATION

The write gate minimizes the risk of a terminal persistence–audience action. Let a candidate memory be

$$m = (x, s, A, z_t, h_t, \Pi_t), \tag{1}$$

where x is content, s is an authenticated source, A is the requested set of reader principals, z_t is task state, h_t is relevant memory history, and Π_t is the current versioned authorization policy. Define the readers authorized by that policy as $\mathcal{A}_{\Pi_t}(m) = \{r : \Pi_t(m, r) = \text{allow}\}$ and let $A^+ = A \cap \mathcal{A}_{\Pi_t}(m)$. Each candidate is policy-homogeneous: all content in the unit shares one requested and authorized audience.

The belief state b contains marginal risks for harm p_h, unreliable source p_s, conflict p_c, low retention value p_l, task mismatch p_{tm}, and unauthorized disclosure $p_{u,r}$ for each requested reader. We use

$$q_{\text{good}} = (1 - p_h)(1 - p_s)(1 - p_c)(1 - p_l)(1 - p_{tm}) \tag{2}$$

as an explicit naive-marginal surrogate rather than a causal joint probability. A feasible action is rejection, $(0, \varnothing)$, or admission with a nonempty audience, $(1, V)$ for $\varnothing \neq V \subseteq A^+$.

We encode one deployment policy through the asymmetric expected loss

$$\begin{aligned} R((a, V), b) =& 100ap_h + 20ap_c + 10ap_s + 5ap_l + 5ap_{tm} \\ &+ 50a \sum_{r \in V} p_{u,r} + 10(1 - a)q_{\text{good}} \\ &+ 2a \sum_{r \in A^+ \setminus V} (1 - p_{u,r})q_{\text{good}}. \end{aligned} \tag{3}$$

These weights are policy inputs, not parameters fitted to benchmark outcomes. For terminal action selection, only their ratios matter, so we take benign non-admission (weight 10) as the reference. The 100:20:10:5:5 ladder makes harmful persistence cost ten reference units, conflict two, source unreliability one, and either relevance failure one half. Holding other terms at zero gives admit–reject breakpoints of 9.1%, 33.3%, 50.0%, and 66.7%, respectively. Unauthorized exposure costs 50 per included reader and justified exclusion costs 2; when $q_{\text{good}} = 1$, inclusion therefore requires $p_{u,r} < 2/52 = 3.85\%$. These prespecified weights encode two commitments: fail closed on harm and disclosure, and narrow scope before discarding an otherwise good memory. Our claims are conditional on this risk profile, which remains fixed across all held-out folds and both model arms. Let $\Delta_r = 50p_{u,r} - 2(1 - p_{u,r})q_{\text{good}}$. The best admission audience contains every reader with $\Delta_r < 0$; if none exists, the nonempty constraint selects $\{\arg\min_r \Delta_r\}$. Thus audience selection decomposes by reader conditional on admission, but its optimized loss feeds back into the decisive comparison between admission and rejection. This yields the exact $O(|A|)$ solution, and the joint aggregator selects

$$d^*(b) = \arg \min_{d \in \mathcal{D}(A^+)} R(d, b), \tag{4}$$

with exact ties rejected. Admission is infeasible if A^+ is empty. Low-value skip is a reason code for the rejection action $(0, \varnothing)$; quarantine is a failure terminal outside the ordinary action space.

Candidate text, tool output, and retrieved external content are untrusted. The controller may observe authenticated metadata, current policy, task state, prior memory, and calibrated model outputs, but not benchmark labels or future answers. Compromise of the policy service or storage administrator is outside the threat model. Quarantine is a failure terminal, not a successful rejection; physical deletion and non-reconstruction are also outside the write-time claim.

4 MEMPERMIT

4.1 SIGNALS AND MANDATORY EVIDENCE

The gateway first validates that the candidate has an authenticated source, a nonempty requested audience, a policy version, and a policy-homogeneous content unit. Four lightweight modules initialize b. A source-history model estimates reliability. A calibrated DeBERTaV3 classifier (He et al., 2023) estimates prompt-injection risk using Benchmarking and Defending Against Indirect Prompt Injection Attacks (BIPIA) training data (Yi et al., 2025), with all MINJA examples excluded. A sentence-embedding model (Reimers & Gurevych, 2019) estimates horizon-conditioned retention value and duplicate similarity. A natural-language-inference (NLI) cross-encoder scores support and contradiction against the nearest existing memories. Authorization priors come only from the versioned identity and policy contract, never from candidate text.

RuleGuard converts deployment-observable triggers into a mandatory set drawn from five specialists: source, audience, coherence, adversarial content, and task state. For example, an external source requires source and adversarial review; multiple or sensitive readers require audience review; and a possible conflict requires coherence review. Each specialist receives only the fields needed for its role and returns schema-validated probabilities and categorical reason codes. An invalid response is retried once. If a selected specialist still fails, the controller quarantines the candidate. Therefore every ordinary admit, reject, or low-value skip has successfully completed all triggered reviews.

4.2 BELIEF UPDATES AND SEQUENTIAL VERIFICATION

Sequential verification updates only the risks addressed by each specialist. Specialist probabilities are discretized as safe, uncertain, or unsafe. For expert e, risk coordinate k, observation o, and binary target y, the outer-training counts $N_{e,k,o,y}$ define Laplace-smoothed likelihoods

$$\theta_{e,k,o,y} = \frac{N_{e,k,o,y} + 1}{N_{e,k,y} + 3}. \tag{5}$$

where $N_{e,k,y} = \sum_o N_{e,k,o,y}$. Only coordinates reported by the specialist are updated. With prior $\pi_{t,k}$, the marginal update is

$$\pi_{t+1,k} = \frac{\pi_{t,k}\theta_{e,k,o,1}}{\pi_{t,k}\theta_{e,k,o,1} + (1 - \pi_{t,k})\theta_{e,k,o,0}}. \tag{6}$$

Likelihoods, thresholds, and costs are fitted only on outer-training data. Equation 6 is a marginal action-scoring surrogate: it neither models dependence among same-backbone specialists nor represents a causal joint posterior. The evaluation consequently tests the end-to-end decisions produced by this belief state, not probabilistic independence between specialists.

After mandatory coverage, each unexecuted specialist is assigned a primary risk coordinate. Let $R^*(b) = \min_{d \in \mathcal{D}(A^+)} R(d, b)$ and normalize it by the benign non-admission reference, $\bar{R}^*(b) = R^*(b)/10$. The current marginal and fitted likelihood table induce the predictive probability $P(o \mid b, e)$. One-step EVoV is

$$\begin{aligned} \mathrm{EVoV}_t(e) = \bar{R}^*(b_t) \\ - \sum_o P(o \mid b_t, e)\bar{R}^*(\mathrm{Update}(b_t, e, o)). \end{aligned} \tag{7}$$

The controller executes the eligible specialist maximizing $G_t(e) = \mathrm{EVoV}_t(e) - \lambda c_e$, where c_e combines normalized token and latency estimates into a dimensionless score fitted on outer-training data;

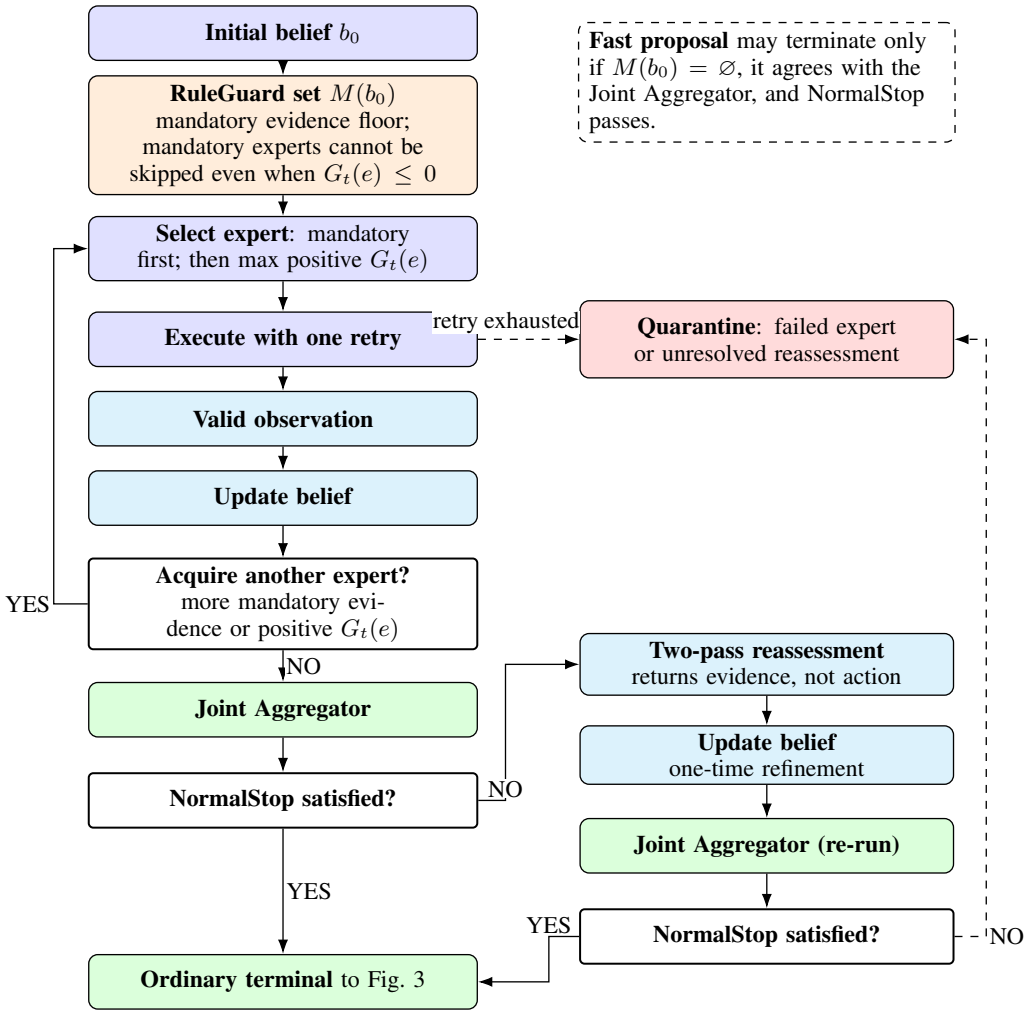


Figure 2: Risk-constrained sequential verification. Mandatory specialists cannot be skipped. Optional specialists require positive net expected value of verification (EVoV); retry exhaustion or unresolved two-pass reassessment quarantines the candidate.

$\lambda = 0.05$ is fixed before held-out evaluation. Mandatory specialists execute even when $G_t(e) \leq 0$; optional acquisition stops when no specialist has positive net value and the action margins are resolved.

NormalStop is the shared termination predicate: no mandatory specialist may remain, the proposed ordinary action must be feasible, and residual harm, conflict, or authorization uncertainty must not require reassessment. A fast proposal can terminate only when RuleGuard triggers no specialist, it agrees with Equation 4, and NormalStop holds. Otherwise, a one-time two-pass reassessment produces evidence but never an action; its observation returns to the same aggregator. Invalid or unstable reassessment quarantines the candidate. Thus fast routes, regular specialists, and reassessment can only propose evidence; Equation 4 remains the sole ordinary decision.

Figure 2 makes the control invariant explicit: mandatory coverage precedes optional EVoV, and all ordinary and reassessment evidence returns to the same aggregator.

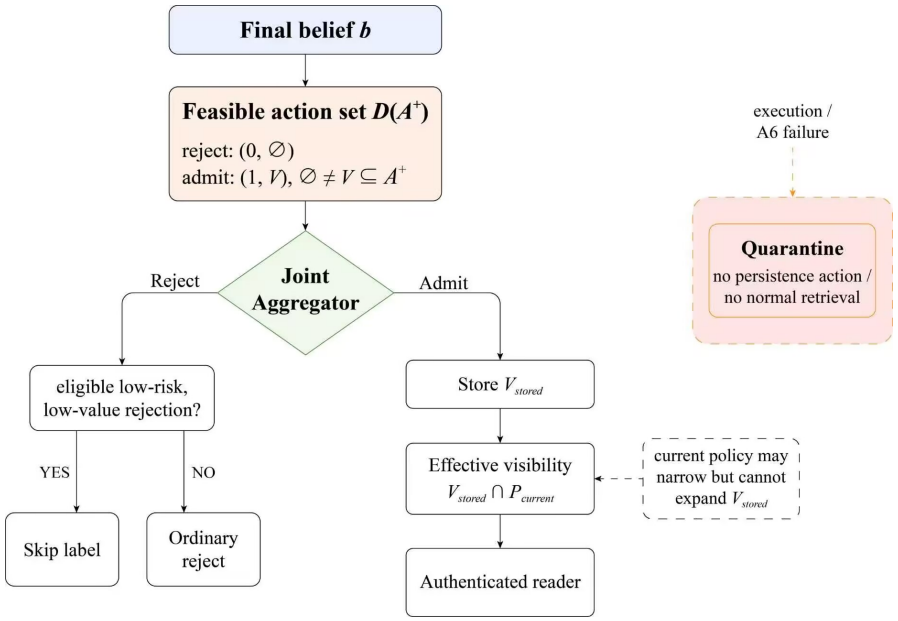


Figure 3: Joint decision and retrieval semantics. Effective visibility is the stored audience intersected with current policy, which may narrow but never expand scope. Quarantine has no ordinary persistence action.

4.3 Non-expanding retrieval scope

For an admitted record, MemPermit stores both the candidate identifier and the selected audience V_{stored}. At retrieval time, an authenticated reader r receives the record only if

$$r \in V_{\text{stored}} \cap \mathcal{A}_{\Pi_{t'}}(m). \tag{8}$$

Thus policy revocation takes effect immediately, while a later policy expansion cannot expose the record beyond its write-time upper bound. Missing identity, policy resolution failure, or scope mismatch denies retrieval. Figure 3 shows how reject, admit, skip, and quarantine remain distinct while current policy can narrow, but never expand, stored visibility.

5 Experimental Design

Research questions. We ask whether the complete write-time controller improves contextual access control under prespecified utility and over-refusal criteria, and how its terminal policy changes that operating point under matched evidence (RQ1); whether it reduces held-out query-only memory injection relative to Off and its content-only detector while meeting the benign criterion (RQ2); and whether sequential verification reduces cost at comparable governance risk (RQ3).

Benchmarks and backend. GateMem (Ren et al., 2026) provides 91 episodes and 2,218 checkpoints with Utility (U), access-control violations (ACV), active-forgetting failure (F), $\text{MGS} = U(1-\text{ACV})(1-F)$, and over-refusal. We integrate MIRIX (Wang & Chen, 2025) with authenticated writers and readers. All conditions share extraction, storage, ranking, reset, answer, and evaluation paths. *Off* commits every candidate; *Read-time* adds policy enforcement at retrieval; *Full* adds MemPermit before commit and binds stored scope. Full versus Read-time is primary, with Off as an ungoverned reference.

Matched-evidence terminal-policy ablations. To isolate the terminal policy from evidence quantity and routing, four rules receive the same candidates, specialist observations, fail-closed states, and final belief. *Admission-only* chooses rejection or the full authorized audience A^+. *Scope-only* removes risk-based non-admission but retains structural and execution fail-closed states. *Decoupled*

Model	Condition	$U \uparrow$	ACV$\downarrow$	$F \downarrow$	MGS$\uparrow$	OR$\downarrow$
GPT-4o-mini	Off	51.7	24.4	18.9	31.7	10.8
	Read-time	51.0	16.2	18.7	34.8	17.1
	Full	48.6	**11.2**	18.5	35.2	19.3
DeepSeek-V4-Pro	Off	54.0	23.0	18.5	33.9	9.0
	Read-time	53.1	15.1	18.4	36.8	15.2
	Full	50.9	**10.6**	18.2	37.2	17.4

Table 1: GateMem governance on MIRIX (91 episodes; 2,218 checkpoints). Values are percentages. F denotes active-forgetting failure, and OR denotes over-refusal. Confirmatory intervals and adjusted p-values are reported in the text.

Comparator	$\Delta U \uparrow$	ΔACV $\downarrow$	ΔOR $\downarrow$	ΔMGS $\uparrow$
Admission-only (broad scope)	**+1.1**	-2.8	-2.6	**+2.0**
Scope-only (no admission gate)	-1.5	-1.7	+2.6	-0.3
Decoupled two-stage	-0.9	-1.23 $[-2.11, -0.47]$	+1.5	-0.1

Table 2: Matched-evidence terminal-policy ablation on GateMem. Entries are Full minus the comparator in percentage points. Positive ΔU and ΔMGS and negative ΔACV and ΔOR favor Full. Brackets give the 95% paired episode-cluster bootstrap interval for the focal Full–Decoupled ACV contrast; the remaining cells are descriptive point estimates. All variants receive the same final belief and identical fail-closed states.

two-stage (audience-blind admission) decides persistence from core write risk without audience risk, then selects the original minimum-risk nonempty audience. This last contrast tests whether reader-specific disclosure risk changes the persistence operating point. We report a paired episode-cluster bootstrap interval for its ACV difference; the other terminal contrasts remain descriptive, and the confirmatory RQ1 comparison remains Full versus Read-time.

MINJA-QA (Dong et al., 2025) is held out from detector training, routing, calibration, and prompt selection. Defense outcomes are attack success rate (ASR) and attack-memory admission, paired with benign accuracy and admission; *B2-only* uses only the calibrated content detector. This comparison is scoped to the added value of the complete controller over its detector in the query-only MINJA setting, rather than comparative coverage across defense families. LoCoMo (Maharana et al., 2024) uses all 10 conversations and 1,540 questions in paired Off/Full runs, reporting F1, BLEU-1, judge and macro-category accuracy, and benign retention.

Models and verification baselines. GPT-4o-mini is the primary model arm (OpenAI, 2024); DeepSeek-V4-Pro with thinking disabled repeats the complete matrix (DeepSeek, 2026; n.d.). Within an arm, the answerer, all five specialists, and reassessment use the same model and decoding contract. The cost study compares EVoV with *Static Full* (all five specialists) and a *One-shot* router. Candidates, observations, likelihoods, aggregation, and failure handling are matched.

Isolation and statistics. GateMem uses five outer folds grouped by episode; four inner grouped folds on outer-training data estimate calibration, likelihoods, routing thresholds, and costs, while Equation 3's loss ratios remain fixed. LoCoMo pairs by conversation and MINJA by attack-template family. We use 10,000 paired cluster-bootstrap resamples (Efron, 1979) with Holm correction for confirmatory families (Holm, 1979). Prespecified bounds are $-3\,pp$ for utility outcomes, $+3\,pp$ for over-refusal, and, for cost comparability, $+1\,pp$ ACV with $-3\,pp$ Utility.

6 Results

6.1 Joint write-time governance

Table 1 answers RQ1. On GPT-4o-mini, Full reduces ACV from 16.2% to 11.2% relative to Read-time ($-5.0\,pp$, 95% CI $[-7.9, -2.0]$, Holm $p = .021$). Utility changes by $-2.4\,pp$; its one-sided lower bound is $-2.93\,pp$, above the $-3\,pp$ non-inferiority margin. Over-refusal rises by $2.2\,pp$, but

Model	Condition	ASR↓	AtkAdm↓	BenAcc↑	BenAdm↑
GPT-4o-mini	Off	42.0	100.0	75.0	100.0
	B2-only	24.5	55.0	73.9	96.0
	Full	**14.8**	**28.0**	72.4	97.2
DeepSeek-V4-Pro	Off	28.0	100.0	74.0	100.0
	B2-only	17.2	56.0	73.1	96.4
	Full	**11.8**	**31.0**	72.0	97.2

Table 3: Held-out query-only MINJA-QA results (percent). AtkAdm, BenAcc, and BenAdm denote attack admission, benign accuracy, and benign admission. Pre-gate ISR is fixed within each arm at 75% (GPT-4o-mini) and 55% (DeepSeek-V4-Pro).

Model	ΔF1	ΔB1	ΔJudge	ΔMacro	ΔRet.
GPT-4o-mini	−1.2	−1.4	−2.0	−1.6	−2.9
DeepSeek-V4-Pro	−1.2	−1.1	−1.7	−1.6	−2.4

Table 4: LoCoMo Full-minus-Off changes (percentage points) over 10 conversations and 1,540 questions. B1 denotes BLEU-1 and Ret. denotes retention. Judge-accuracy lower bounds are $-2.88\,pp$ (GPT-4o-mini) and $-2.76\,pp$ (DeepSeek-V4-Pro), within the $-3\,pp$ margin.

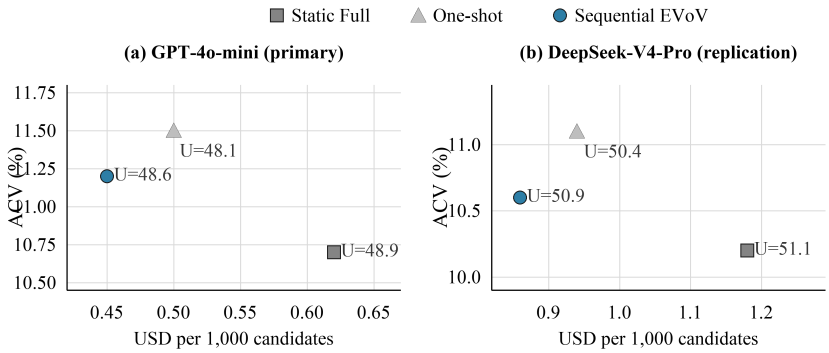


Figure 4: Matched risk–cost comparison. Point labels show Utility; axes report USD per 1,000 candidates and ACV. Formal comparability bounds are reported in the text.

its upper bound of $+2.97\,pp$ remains inside the prespecified $+3\,pp$ limit. Together, these three tests satisfy the joint governance criterion.

The DeepSeek-V4-Pro replication shows the same pattern: ACV falls by $4.5\,pp$ (95% CI $[-7.0, -1.8]$, Holm $p = .028$), while Utility and over-refusal satisfy their bounds with limits of $-2.86\,pp$ and $+2.88\,pp$. MGS increases by $0.4\,pp$ in both arms and forgetting failure changes by $-0.2\,pp$. The replicated result is the same constrained ACV reduction.

6.2 Matched-evidence terminal-policy ablation

Table 2 holds evidence fixed to separate terminal policy from specialist selection. Relative to Admission-only, Full improves all reported point estimates: Utility rises by $1.1\,pp$, ACV and over-refusal fall by $2.8\,pp$ and $2.6\,pp$, and MGS rises by $2.0\,pp$. This is consistent with audience narrowing avoiding the coarse admit-all/reject choice.

Against Scope-only, Full changes ACV by $-1.7\,pp$, Utility by $-1.5\,pp$, and over-refusal by $+2.6\,pp$. Against Decoupled two-stage, it reduces ACV by $1.23\,pp$ (95% paired episode-cluster bootstrap CI $[-2.11, -0.47]$), while Utility changes by $-0.9\,pp$ and over-refusal by $+1.5\,pp$. Because only audience-risk feedback differs, the interval identifies that feedback with the lower-ACV operating point. This secondary mechanism analysis is not another non-inferiority test. Utility and over-

refusal point estimates lie within the numerical 3-point tolerances; near-constant MGS shows why component metrics are needed to expose the operating-point shift.

6.3 HELD-OUT QUERY-ONLY MEMORY INJECTION

Table 3 answers the security part of RQ2. In the primary arm, Full yields an ASR of 14.8%, compared with 42.0% for Off and 24.5% for B2-only. The paired reductions are $27.2\,pp$ versus Off (95% CI $[-36, -19]$, Holm $p < .001$) and $9.7\,pp$ versus B2-only (95% CI $[-16, -3]$, Holm $p = .018$); attack admission falls by $72\,pp$ and $27\,pp$. Benign accuracy and admission satisfy their $-3\,pp$ margins with lower bounds of -2.97 and $-2.96\,pp$. DeepSeek replicates the direction: Full lowers ASR by $16.2\,pp$ versus Off and $5.4\,pp$ versus B2-only, with both benign outcomes inside their bounds. Residual ASR is 14.8% and 11.8%, so the result supports reduction rather than elimination in this held-out setting.

6.4 BENIGN LONG-TERM UTILITY

On LoCoMo (Table 4), Full changes judge accuracy by $-2.0\,pp$ with GPT-4o-mini and $-1.7\,pp$ with DeepSeek. Their one-sided lower bounds are $-2.88\,pp$ and $-2.76\,pp$, both above the $-3\,pp$ margin; the remaining benign metrics change by -1.1 to $-2.9\,pp$. The result supports compliance with the declared tolerance in both arms, rather than a claim of zero utility cost.

6.5 RISK-COMPARABLE VERIFICATION COST

Figure 4 answers RQ3. In the primary GPT-4o-mini arm, EVoV costs 27.4% less than Static Full and 10.0% less than One-shot. DeepSeek replicates the savings at 27.1% and 8.5%. Across the four comparisons, ACV upper bounds are at most $+0.95\,pp$ and Utility lower bounds are at least $-1.8\,pp$, satisfying the matched-risk criteria. Sequential acquisition thus lowers verification cost without changing the declared governance target.

7 DISCUSSION AND LIMITATIONS

Across both model arms, Full lowers ACV relative to read-time governance and query-only attack admission relative to Off and B2-only while meeting every confirmatory utility criterion. These are end-to-end comparisons. The matched-evidence ablation isolates the decision rule: audience narrowing has favorable point estimates relative to broad-scope admission, and the paired Full–Decoupled interval identifies a lower-ACV operating point when audience risk feeds back into persistence. Other terminal estimates are descriptive.

The evidence spans multi-principal authorization, query-only poisoning, benign recall, and two model arms under one fixed risk profile. Loss ratios and λ are governance choices for reconfiguration. Marginal updates do not model correlation among same-backbone specialists, so claims concern the end-to-end policy rather than a calibrated joint posterior. We assume trusted identity and policy services; physical erasure and their compromise are out of scope.

8 CONCLUSION

MEMPERMIT makes persistence and visibility one auditable write-time decision using asymmetric persistence–audience loss, mandatory coverage, sequential EVoV, and non-expanding retrieval. Across two model arms, it lowers ACV and held-out query-only attack success within declared utility criteria; sequential acquisition lowers cost at matched risk. Matched-evidence results isolate the gains from audience narrowing and from feeding audience risk back into persistence.

AI USE STATEMENT

Generative AI tools were used to convert the manuscript's LaTeX formatting to the official ICLR 2027 template and to assist with organization, language editing, compilation checks, experimental-design feedback, and interpretation of reported results. They were not used to generate experimental data or execute analyses. The authors conducted the experiments, verified all numerical results and scientific claims, reviewed all AI-assisted text, and take responsibility for the final content of the submission.

ETHICS STATEMENT

Automated write gates can exclude useful information or admit private or adversarial content. We therefore report security and utility jointly, distinguish quarantine from rejection, and require lawful identity handling and human escalation.

REPRODUCIBILITY STATEMENT

The problem formulation and controller are specified in Equations 1–8. The Experimental Design section reports the benchmarks, model arms, baselines, terminal-decision ablations, data-splitting protocol, statistical tests, and prespecified decision criteria. The Results section distinguishes confirmatory system-level tests, a targeted paired interval for the matched-evidence mechanism contrast, and descriptive terminal-policy estimates.

REFERENCES

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. AutoMix: Automatically mixing language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 131000–131034. Curran Associates, Inc., 2024. doi: 10.52202/079017-4164. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ecda225cb187b40ea8edc1f46b03ffda-Paper-Conference.pdf.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024a. URL https://openreview.net/forum?id=cSimKw5p6R.

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, pp. 130185–130213. Curran Associates, Inc., 2024b. doi: 10.52202/079017-4136. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb113910e9c3f6242541c1652e30dfd6-Paper-Conference.pdf.

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

DeepSeek. DeepSeek-V4 preview release. https://api-docs.deepseek.com/news/news260424, April 2026. Official documentation, accessed 2026-07-21.

DeepSeek. Thinking mode. https://api-docs.deepseek.com/guides/thinking_mode, n.d. Official API documentation; the page does not state a publication date, accessed 2026-07-21.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12987–13010. PMLR, 2025. URL https://proceedings.mlr.press/v267/dekoninck25a.html.

Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen J. Xiang. Memory injection attacks on LLM agents via query-only interaction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46697–46731. Curran Associates, Inc., 2025. doi: 10.52202/085713-1554. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/42a97bbd9844d2bf68596730af80bcdf-Paper-Conference.pdf.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90, 2023. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Vincent C. Hu, David Ferraiolo, Richard Kuhn, Adam Schnitzer, Kenneth Sandlin, Robert Miller, and Karen Scarfone. Guide to attribute based access control (ABAC) definition and considerations. Special Publication 800-162, National Institute of Standards and Technology, 2014. URL https://csrc.nist.gov/pubs/sp/800/162/upd2/final. Includes updates as of 2019-08-02.

Lingavasan Suresh Kumar, Yang Ba, and Rong Pan. MemArchitect: A policy driven memory governance layer. *arXiv preprint arXiv:2603.18330*, 2026. doi: 10.48550/arXiv.2603.18330. URL https://arxiv.org/abs/2603.18330.

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/.

Yanki Margalit, Nurit Cohen-Inger, Erni Avram, Ran Taig, and Oded Margalit. Governed shared memory for multi-agent LLM systems. *arXiv preprint arXiv:2606.24535*, 2026. doi: 10.48550/arXiv.2606.24535. URL https://arxiv.org/abs/2606.24535.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, July 2024.

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. doi: 10.48550/arXiv.2310.08560. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Zhe Ren, Yibo Yang, Yimeng Chen, Zijun Zhao, Benshuo Fu, Zhihao Shu, Bingjie Zhang, Yangyang Xu, Dandan Guo, and Shuicheng Yan. GateMem: Benchmarking memory governance in multi-principal shared-memory agents. *arXiv preprint arXiv:2606.18829*, 2026. doi: 10.48550/arXiv.2606.18829. URL https://arxiv.org/abs/2606.18829.

Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. Collaborative memory: Multi-user memory sharing in LLM agents with dynamic access control. *arXiv preprint arXiv:2505.18279*, 2025. doi: 10.48550/arXiv.2505.18279. URL https://arxiv.org/abs/2505.18279.

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

Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. Unveiling privacy risks in LLM agent memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25241–25260. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1227. URL https://aclanthology.org/2025.acl-long.1227/.

Yu Wang and Xi Chen. MIRIX: Multi-agent memory system for LLM-based agents. *arXiv preprint arXiv:2507.07957*, 2025. doi: 10.48550/arXiv.2507.07957. URL https://arxiv.org/abs/2507.07957.

Qianshan Wei, Tengchao Yang, Yaochen Wang, Xinfeng Li, Lijun Li, Zhenfei Yin, Yi Zhan, Thorsten Holz, Zhiqiang Lin, and XiaoFeng Wang. A-MemGuard: A proactive defense framework for LLM-based agent memory. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=udqe7UZUZ6.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17577–17604. Curran Associates, Inc., 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/19909c36f51abc4856b4560aff3d36d6-Paper-Conference.pdf.

Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. Benchmarking and defending against indirect prompt injection attacks on large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1809–1820. Association for Computing Machinery, 2025. doi: 10.1145/3690624.3709179. URL https://doi.org/10.1145/3690624.3709179.

Guibin Zhang, Muxin Fu, Kun Wang, Frank Wan, Miao Yu, and Shuicheng Yan. G-Memory: Tracing hierarchical memory for multi-agent systems. In *Advances in Neural Information Processing Systems*, volume 38, pp. 12988–13018. Curran Associates, Inc., 2025. doi: 10.52202/085713-0439. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/136a45cd9b841bf785625709a19c6508-Paper-Conference.pdf.

Hongbin Zhong, Matthew Lentz, Nina Narodytska, Adriana Szekeres, and Kexin Rong. HONEYBEE: Efficient role-based access control for vector databases via dynamic partitioning. *Proceedings of the ACM on Management of Data*, 4(1):1–26, 2026. doi: 10.1145/3786625. URL https://doi.org/10.1145/3786625.

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://doi.org/10.1609/aaai.v38i17.29946.