REFUSAL CHANGES THE FUTURE: ROLLOUT-EQUALIZED CREDIT ASSIGNMENT FOR SEQUENTIAL MEMORY MAINTENANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

A refused merge is a state transition: preserving two records changes retrieval and every later maintenance opportunity. Rollout-Equalized Refusal (RER) measures this dependence by freezing the pre-mutation refusal transition in paired reruns; a second intervention applies one state-conditional candidate-construction rule to both controllers. Its operator-conditional estimand admits exact pathwise accounting, and a two-step construction shows how equal refusal coverage or logged filtering can fail to identify the rollout quantity. At its complete-policy 200-slot operating point, MERQ reaches 78.4% task success and 3.2% method-blind, policy-level per-decision contradiction burden across 8,400 episodes, gains 1.6 task points over Compressive Memory, and uses 22.1% fewer slots than saturated LRU/LFU. Under the frozen MERQ-associated operator, its advantage over same-action-space RL is $+0.7$ task and $+1.8$ favorable contradiction points on StreamMem-Bench and has the same direction on LongMemEval, LoCoMo, and AgentOps-90d. Under the common construction rule, $+0.6\,[0.31, 0.88]$ task and $1.3\,[1.02, 1.58]$ favorable contradiction points remain; against independently trained Selective PPO at nearly identical refusal coverage, the margins are $+0.4\,[0.12, 0.69]$ and $0.9\,[0.62, 1.18]$. A 3,600-decision blinded audit measures 87.5% safe-merge recall, 92.4% unsafe-merge recall, and Fleiss' κ=0.86. These results establish refusal as a rollout-level evaluation object and localize MERQ's retained advantage to post-generation control under the declared operator and candidate-construction rule.

1 INTRODUCTION

A memory controller proposes merging an earlier tool return with a later correction. Execution leaves one record; refusal preserves two, marks a conflict, redirects retrieval, and changes later proposals. This is not a prediction removed from a fixed test set, but a transition that creates the test cases encountered next. The evaluation error appears when later probes and merge opportunities consume records preserved by refusal, a routine pattern in assistants and tool agents with evolving preferences, observations, and task state (Park et al., 2023; Yao et al., 2023; Packer et al., 2023; Chhikara et al., 2025). Their complete-policy operating points entangle candidate generation, action scoring, refusal, and induced state; matching only marginal refusal leaves these sources unresolved.

RER turns the entanglement into three executable comparisons. Native rollouts measure complete policies. Common-transition rollouts impose one held-out-calibrated noop-plus-quarantine operator with paired episode order, reader samples, controller streams, and environment outcomes. The final comparison applies the same deterministic construction function H to each controller's own visited state before scoring and selection. The sequence removes one source of advantage at a time while retaining downstream consequences.

The middle comparison defines an operator-conditional residual. An exact identity relates the native gap to this residual, a one-step contrast at refusal disagreements, and the induced-state remainder. A two-step counterexample establishes that equal coverage or post-hoc deletion can fail to identify the residual when the unobserved next bank determines later reward.

At its complete-policy point, MERQ leads same-action-space RL by $+1.3$ task and $+0.7$ favorable per-decision contradiction points. The common MERQ-associated transition changes the StreamMem

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.

margins to $+0.7/+1.8$, with all eight intervals across four settings excluding zero. Applying H as well leaves $+0.6\,[0.31, 0.88]/+1.3\,[1.02, 1.58]$; AgentOps-90d retains $+0.9/+1.0$. This final comparison is post-generation: both controllers use the same state-conditional construction rule and refusal transition, while scoring, selection, visited states, and realized tuples can differ.

MERQ combines learned candidate ordering, a separately parameterized NLI risk gate, and quarantine. Its held-out $\tau_q^\star$=0.847 is frozen across five settings and three readers; AMA-Bench supplies a second natural-trajectory test (Zhao et al., 2026). The complete policy reaches 78.4% task, 3.2% method-blind per-decision contradiction burden, 155.8 slots, and 389 ms p95 latency. At nearly identical refusal coverage, independently trained Selective PPO trails by 0.4 task and 0.9 favorable contradiction points; a separate blinded audit directly measures gate behavior.

The paper makes three contributions:

- A rollout-level estimand for stateful refusal, an exact three-term pathwise accounting, and a counterexample showing when static rejection controls fail to identify it.
- A native-to-common-transition-to-common-construction protocol that separates proposal generation, scoring and selection, refusal, and induced-state effects without replacing rollouts by logged filtering.
- A complete-policy study across five memory settings, three readers, two natural-trajectory families, an independently learned reject policy, and 3,600 audited decisions with 10,800 blinded labels, localizing retained differences to post-generation control under the declared transition and construction rule.

Together, the analysis and experiments make a coupled reject transition measurable and show how evaluation conclusions change as the refusal transition and candidate-construction rule are held common.

2 Related Work

Selective prediction studies risk at a chosen coverage on a fixed sample (Bartlett & Wegkamp, 2008; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019). Its observational unit remains unchanged after rejection. In sequential memory, the rejected mutation determines the next memory bank; equal coverage can therefore conceal unequal trajectories. RER instead treats the refusal rule as part of the transition, reruns subsequent decisions, and reports the resulting difference in task and contradiction-burden units.

Off-policy evaluation estimates outcomes under alternative policies from logged or modeled data (Thomas & Brunskill, 2016); dynamic mediation separates immediate and delayed pathways under longitudinal causal assumptions (Ge et al., 2023). RER uses paired executable rollouts: the declared operator is directly imposed, the environment and random streams are coupled, and the later states are observed under the intervention. Its immediate/rollout split is an exact accounting under a specified branch probe, rather than an operator-free path-specific causal effect. Action-constrained RL and tool-agent alignment learn when to refuse (Schulman et al., 2017; Hung et al., 2025; Agarwal et al., 2026); Selective PPO provides the closest independently learned reject-policy comparison at matched coverage and compute.

External-memory architectures learn storage, retrieval, organization, or compression. Memory Networks, Differentiable Neural Computers, and memory-augmented meta-learning established trainable stores (Weston et al., 2015; Sukhbaatar et al., 2015; Graves et al., 2016; Santoro et al., 2016); RETRO, Memorizing Transformers, Compressive Memory, Infini-attention, MemLLM, MemGPT, MemoryBank, Mem0, and A-MEM extend retrieval and persistent agent memory (Borgeaud et al., 2022; Wu et al., 2022; Rae et al., 2020; Munkhdalai et al., 2024; Modarressi et al., 2025; Packer et al., 2023; Zhong et al., 2024; Chhikara et al., 2025; Xu et al., 2025). The intervention matrix uses controllers whose native signals can be mapped to one merge/delete/archive/split/noop boundary; broader end-to-end write interfaces such as Mem0 and A-MEM define complementary system-level comparisons. LongMemEval, LoCoMo, and AMA-Bench evaluate updates, multi-session dialogue, and agent trajectories (Wu et al., 2025; Maharana et al., 2024; Zhao et al., 2026). RER adds a common-transition view to outcomes on these trajectory families.

3 INTERVENTIONAL CREDIT ASSIGNMENT FOR STATEFUL REFUSAL

3.1 DECISION AND TRANSITION SEMANTICS

At maintenance decision t, the bank $\mathcal{M}_t$ contains typed entries with content, provenance, confidence, recency, read frequency, and contradiction markers. A controller g produces an ordered list of at most $K{=}8$ typed tuples from merge, delete, archive, split, and noop. Merge tuples name two entries; delete, archive, and split name one; noop has no target. A tuple is structurally feasible when its targets are live, its arity and entry types satisfy the action schema, and its execution satisfies the declared resource budget. Semantic merge risk is evaluated only after this structural check. Let S return the first feasible tuple and let $\mathrm{qnoop}(i,j)$ denote a noop that also quarantines the named pair, distinct from an ordinary noop. Candidate screening is not an action batch: each decision executes exactly one transition.

Let $U=(u_1,\dots,u_T)$ collect paired exogenous randomness, including episode order, reader sampling, controller random streams, and environment outcomes. Under refusal rule r, proposal, selection, executed action, and state evolve as

$$\mathcal{L}_t^g = g(\mathcal{M}_t^{g,r}, h_t), \quad \tilde{a}_t^g = S(\mathcal{M}_t^{g,r}, \mathcal{L}_t^g), \quad a_t^{g,r} = T_r(\mathcal{M}_t^{g,r}, \tilde{a}_t^g), \quad \mathcal{M}_{t+1}^{g,r} = P(\mathcal{M}_t^{g,r}, a_t^{g,r}, u_t), \tag{1}$$

where T_r maps a rejected $\mathrm{merge}(i,j)$ to $\mathrm{qnoop}(i,j)$ and otherwise returns $\tilde{a}_t^g$. The action-cost vector is $\mathbf{c}(a)=(c^{\mathrm{slot}},c^{\mathrm{byte}},c^{\mathrm{lat}})$. Main results use slot budget $C{=}200$; byte and latency scalarizations appear in Appendix J.

3.2 ROLLOUT-EQUALIZED REFUSAL

Let $e_\psi(i,j), c_\psi(i,j)\in[0,1]$ be the gate's entailment and contradiction softmax probabilities, and let $\mathrm{ProvOK}(i,j)$ indicate that the merge preserves source lineage. Define

$$\mathrm{HardOK}_\psi(i,j) = \mathbb{1}[e_\psi(i,j)\ge 0.85,\ c_\psi(i,j)\le 0.3,\ \mathrm{ProvOK}(i,j)].$$

For each pair, $q_\psi(m_i,m_j)=\sqrt{e_\psi(i,j)[1-c_\psi(i,j)]}$. Write $a_{\min}(i,j)=\min\{\mathrm{age}(m_i),\mathrm{age}(m_j)\}$ and

$$r_{\psi,\tau}(i,j) = \mathbb{1}[\mathrm{HardOK}_\psi(i,j)=0 \ \vee\ q_\psi(m_i,m_j)<\tau \ \vee\ a_{\min}(i,j)<24\mathrm{h}].$$

The shared operator acts on the selected feasible tuple when it is a merge:

$$\Omega_{\tau_q^\star}(a) = \begin{cases} r_{\psi,\tau_q^\star}(i,j), & a=\mathrm{merge}(i,j),\\ 0, & \text{otherwise.}\end{cases} \tag{2}$$

The wrapped controller $W_{\tau_q^\star}(g)$ uses $T_{r^\star}$ to execute the selected proposal when $\Omega_{\tau_q^\star}=0$ and $\mathrm{qnoop}(i,j)$ otherwise. Quarantine preserves both records and provenance, keeps them separately retrievable, excludes the pair from later merge proposals in that episode, and charges both records to the slot and byte footprint. No release occurs before episode end. Hard checks, thresholding, the 24-hour rule, termination, and this transition are identical for every wrapped controller. The wrapper therefore changes future eligibility, retrieval, and capacity rather than filtering a log. The declared $r^\star$ is MERQ-associated, so $G_{\mathrm{res}}(r^\star)$ measures controller behavior with that transition. Section 6.2 adds the common construction rule H.

For trajectory functional F (task success or negative per-decision post-state contradiction burden, evaluated separately), denote by $Z_g^r(U)=F(\mathcal{M}_{0:T}^{g,r},a_{0:T}^{g,r})$ the outcome of controller g under refusal rule r, with native rule r_g and shared rule $r^\star$. Define the operator-conditional residual

$$G_{\mathrm{res}}(r^\star) = \mathbb{E}_U[Z_g^{r^\star}(U) - Z_b^{r^\star}(U)]. \tag{3}$$

For controller c, branch once at every native-versus-shared refusal disagreement on the pre-action state visited by its shared-rule rollout. One branch applies r_c and the other $r^\star$, with identical u_t. For task, both post-decision states answer the same held-out probe scheduled at that decision and the local contrast is the difference in binary reader correctness. For contradiction, both branches score the same feasible merge target with the post-state incompatibility predicate used by the main endpoint. Each local task contrast is divided by the trajectory's number of probes and each contradiction contrast by its number of maintenance decisions; variable-length trajectories and trajectories without

Table 1: Interventions and supported interpretations. Every comparison shares paired exogenous randomness; "specific" components remain inside the reported difference.

Quantity	Held common	Controller-specific	Supported interpretation
G_{native}	Evaluation episodes and U	Proposals, scores, refusal, states	Complete-policy difference (Table 2)
$G_{\text{res}}(r^\star)$	Episodes, U, refusal transition	Proposals, scores, resulting states	Gap under the declared operator (Table 4)
$G_{\text{select}}, G_{\text{roll}}$	Shared-rule visited state and local probe	Native/shared branch action; later state	Immediate branch contrast and exact remainder (Table 3)
$G_{\text{res}}(r^\star, H)$	Episodes, U, operator, construction rule H	Realized tuples, scores, selection, states	Post-generation gap under common state-conditional construction (Table 5)

disagreements therefore retain the normalization of F. The branches are then discarded before the shared rollout continues. Let $I_c(U)$ be the sum of these signed native-minus-shared local contrasts. The induced-state remainder, including all later consequences and interactions among disagreements, is

$$R_c(U) = Z_c^{r_c}(U) - Z_c^{r^\star}(U) - I_c(U). \tag{4}$$

Define $G_{\text{native}} = \mathbb{E}_U[Z_g^{r_g}(U) - Z_b^{r_b}(U)]$. The native gap obeys the pathwise accounting identity

$$G_{\text{native}} = G_{\text{res}}(r^\star) + \underbrace{\mathbb{E}[I_g - I_b]}_{G_{\text{select}}} + \underbrace{\mathbb{E}[R_g - R_b]}_{G_{\text{roll}}}. \tag{5}$$

RER estimates $G_{\text{res}}(r^\star)$ by averaging paired rollout differences. Within each seed, probes or decisions are first aggregated over the locked evaluation corpus; reported intervals bootstrap the ten paired seed-level means. Identification requires intervention consistency, identical initial states, replayable U, no cross-rollout interference, and deterministic outcomes conditional on the coupled random streams. The one-step branches estimate G_{select} and subtraction gives G_{roll}. The latter is an exact remainder under this reference path and probe definition, rather than a branch-invariant mediation effect.

Proposition 1. *Equal marginal refusal coverage, or post-hoc removal of refused actions, can fail to identify $G_{\text{res}}(r^\star)$. In particular, there exist two-step processes with equal coverage and equal fixed-sample immediate risk but different rollout residuals because refusal changes a state variable that controls later reward. Static and rollout-equalized comparisons coincide when the execute/refuse choice leaves the transition and all downstream outcomes invariant; action-independent transitions with no immediate refusal effect are a sufficient case.*

Appendix A gives a two-step counterexample and proof.

4 MERQ UNDER ROLLOUT-EQUALIZED REFUSAL

4.1 PROPOSAL, REFUSAL, AND LEARNING

MERQ (Memory Editing with Refusal and Quarantine) encodes each typed tuple from the action type, target contents, provenance compatibility, confidence, age, read frequency, contradiction markers, history, and remaining budget. Its proposal network f_ϕ maps these features to one ordering logit per tuple; exact duplicates are removed before the top $K{=}8$ form $\mathcal{L}_t^g$, and S then applies the full feasibility check. Training uses a straight-through Gumbel–Softmax estimator for the discrete ordering under Equation 8; inference sorts logits deterministically. RL-Only uses the same state, action schema, and reward with PPO. Selective PPO adds a reject action with the identical terminating quarantine transition and selects its coverage multiplier on held-out decisions; neither comparator can query q_ψ. Appendix C gives the mappings and held-common training choices.

The separate gate q_ψ uses DeBERTa-v3-large (He et al., 2023) fine-tuned on MNLI (Williams et al., 2018). Equation 2 enforces the hard checks and thresholds $q_\psi = \sqrt{e_\psi(1 - c_\psi)}$; the age rule protects recent evidence. Thus f_ϕ chooses which merge reaches the gate, and q_ψ determines whether it mutates the bank.

4.2 FREEZE-ONCE CALIBRATION

On $\mathcal{D}_{\text{cal}}$ with n_{cal}=1,000, define

$$\widehat{\text{UAR}}(\tau) = \frac{\#\{\text{gold-unsafe proposals executed at } \tau\}}{\#\{\text{gold-unsafe proposals}\}}, \tag{6}$$

$$\tau_q^\star = \min\{\tau : \widehat{\text{UAR}}(\tau) \le 0.03\}. \tag{7}$$

At $\tau_q^\star$=0.847, the calibration split contains 7/238 executed unsafe proposals (unsafe-action rate 2.94%, 95% binomial interval [1.2,5.9]) and 670/762 executed safe proposals. The 1,000 proposal decisions are disjoint from reported test episodes and from the 3,600-decision mechanism audit. This proposal-conditional quantity differs from $\Pr(\text{unsafe} \mid \text{execute})$, per-decision contradictory post-state burden, and contradiction among accepted audited merges. Bootstrap resampling gives $[0.839, 0.855]$ for the selected threshold. We freeze this value for the five-setting primary matrix and all readers, then reuse it unchanged for the separate AMA-Bench replication.

5 EXPERIMENTAL DESIGN

StreamMem-Bench contains 8,400 test episodes: 3,200 extended PersonaChat sessions (Zhang et al., 2018), 2,800 API-Bank tool traces (Li et al., 2023), and 2,400 Natural Questions streams with temporal revisions (Kwiatkowski et al., 2019). Source identities are partitioned before stream construction. Each episode preserves source chronology, withholds probe turns from controller input, and scores the latest supported value after a correction together with unrevised rare details. Episodes span 50–200 interaction steps and banks grow from 50 to more than 500 entries. LongMemEval and LoCoMo process released sessions chronologically and retain their question categories as held-out probes. AgentOps-90d is a deterministic collection of 500 naturally long-running tool-agent sessions (Appendix G); the AMA-Bench real-trajectory split contributes six agent domains. MedNotes-Stream contains 1,200 held-out de-identified clinical-note episodes with an unseen memory taxonomy, and FinNotes-Stream contributes 400 held-out financial-note decisions to the domain-stratified audit. Appendix B gives the split and scoring map.

The complete-policy comparison covers LRU/LFU, periodic Llama-3.1-8B summarization, Compressive Memory, Infini-attention, MemGPT, MemoryBank, same-action-space PPO (RL-Only), and exhaustive search on episodes whose banks begin with at most 100 entries. Architecture baselines enter through a common maintenance interface: retention scores choose keep/archive, compression state proposes merge, and unsupported split actions map to noop. Intervention cells use the subset with explicit action scores at the shared boundary. RL+RER and Compressive+RER retain their native proposals and training objectives, then receive the operator in Equation 2, including identical termination and quarantine. All learned controllers use frozen readers, the same validation criterion, one reader pass, and a 900 ms/decision ceiling. The small-bank oracle is a quality reference outside that ceiling.

The primary reader is frozen Llama-3.1-70B-Instruct (Grattafiori et al., 2024) at bf16 tensor-parallel four; replications use Qwen2.5-72B-Instruct (Yang et al., 2024) under the same configuration and Llama-3.1-8B-Instruct on one A100-80GB. All systems share C=200, reader, retrieval schema, and maintenance budget. Runs use ten paired seeds $\{13, 42, 97, 128, 271, 314, 512, 777, 1024, 2024\}$. Task success averages binary correctness over held-out probes and then episodes; rare-detail retention is episode-level; quarantine coverage is per decision.

The automatic contradiction endpoint is a policy-level burden. A separate frozen, method-blind binary classifier receives the selected feasible pair when it is a merge, source provenance, and post-transition records; it is distinct from q_ψ and unavailable to every controller. The decision-level predicate is the union of two conditions: the source pair remains semantically incompatible, or an executed merge target contradicts either source. A rejected merge is scored on its preserved, quarantined sources; a decision whose selected feasible tuple is not a merge contributes zero. Thus the endpoint combines merge opportunity, acceptance, and handling quality. The prevalence-weighted human audit separately measures accepted-safe recall, rejected-unsafe recall, and residual contradiction conditional on accepted merges. The final intervention holds the state-conditional construction rule fixed while realized sets remain trajectory-dependent, and it yields nearly identical quarantine rates.

Table 2: Complete-policy StreamMem-Bench operating points under the shared maintenance interface at C=200 with Llama-3.1-70B-Instruct. Task, detail, and contradiction are means $\pm$ 95% CI over ten seeds; slots and p95 are descriptive resource summaries.

Method	Task ↑	Rare detail ↑	Merge-target contrad. ↓
LRU	73.1 ± 0.6	79.2 ± 0.8	0.0 (N/A)
LFU	73.8 ± 0.5	80.1 ± 0.7	0.0 (N/A)
Periodic summary	76.3 ± 0.5	83.4 ± 0.6	5.1 ± 0.3
MemGPT	75.9 ± 0.5	84.1 ± 0.6	4.5 ± 0.3
MemoryBank	76.1 ± 0.5	84.6 ± 0.6	4.2 ± 0.3
Infini-attention	76.6 ± 0.5	85.9 ± 0.5	4.9 ± 0.3
Compressive Memory	76.8 ± 0.5	85.2 ± 0.6	4.8 ± 0.3
RL-Only	77.1 ± 0.6	87.3 ± 0.5	3.9 ± 0.3
MERQ	78.4 ± 0.4	91.7 ± 0.3	3.2 ± 0.2
Oracle (small banks)	82.1 ± 0.3	94.8 ± 0.3	1.1 ± 0.1

Method	Slots ↓	p95 ms	Method	Slots ↓	p95 ms
LRU	200.0	2	LFU	200.0	3
Periodic summary	162.3	847	MemGPT	171.2	318
MemoryBank	169.4	276	Infini-attention	160.5	538
Compressive	158.7	612	RL-Only	164.1	423
MERQ	155.8	389	Oracle	148.2	12,400

LRU/LFU never define merge targets, so their zero is marked N/A rather than interpreted as merge safety.

For each seed, all systems see the same locked episodes and random streams. Paired 95% intervals bootstrap the ten seed-level differences and quantify seed variation for the finite, locked evaluation corpora. Conclusions across trajectory sources are based on replicated direction and effect size rather than a population-level episode claim. Human-audit rate intervals are Wilson intervals conditional on majority labels, and the κ interval uses an item bootstrap.

The audit is sampled in proportion to proposal prevalence across eight domains. All 3,600 items receive three independent safe/unsafe labels from domain-qualified annotators blinded to method identity, and majority labels are used for the confusion matrix; disagreements are adjudicated by a fourth senior annotator for the residual-outcome analysis. Medical items are labeled by registered nurses, all annotators pass a 40-item qualification at at least 85%, and compensation meets or exceeds the platform's posted regional median. The institutional review process classified this non-identifying text-labeling task as exempt minimal-risk research. Fleiss' κ=0.86 summarizes the three-primary-rater agreement.

The experiments answer three questions in sequence: (Q1) does MERQ define a competitive complete-policy operating point; (Q2) which task and contradiction differences survive a common refusal transition and then the common construction rule; and (Q3) do the retained differences persist across controller families, readers, calibration choices, trajectory sources, and mechanism ablations?

6 Results

6.1 From native performance to a common refusal transition

Table 2 establishes the first rung of the comparison. MERQ reduces footprint by 22.1% relative to saturated LRU/LFU, gains 1.6 task points over Compressive Memory, and leads same-action-space RL-Only by 1.3 task and 0.7 favorable per-decision contradiction points. It also improves rare-detail retention by 4.4 points over RL-Only at lower p95 maintenance latency. The following rungs test how much of this complete-policy operating point survives the common refusal transition and construction rule.

Under the declared shared operator and branch probe, Table 3 assigns 0.2 of the 1.3 native task points to the immediate disagreement branches and 0.4 to the exact rollout remainder. Under the favorable-reduction convention, the negative contradiction branch terms partially offset the $+1.8$

Table 3: Pathwise credit accounting on StreamMem-Bench. Task uses MERQ minus RL; contradiction uses the favorable reduction $C_{\mathrm{RL}} - C_{\mathrm{MERQ}}$, so positive values favor MERQ. Native equals residual plus immediate selection plus rollout remainder. Brackets are paired 95% CIs for the intervention terms.

Metric	Native	Residual	Immediate selection	Rollout remainder
Task points	$+1.3$	$+0.7$	$+0.2\,[0.07, 0.34]$	$+0.4\,[0.16, 0.65]$
Contradiction points	$+0.7$	$+1.8$	$-0.3\,[-0.47, -0.14]$	$-0.8\,[-1.03, -0.58]$

Table 4: Paired residuals under the frozen $\tau_q^\star{=}0.847$ transition. Each cell before the deltas is task success / per-decision contradiction (%). ΔT is MERQ minus RL+RER; ΔC is the favorable contradiction reduction. Brackets are paired 95% CIs over ten seeds. The wrapper reruns the rollout after each transition.

Setting	RL+RER	MERQ	ΔT	ΔC
StreamMem	77.7 / 5.0	78.4 / 3.2	$+0.7\,[0.70, 0.78]$	$+1.8\,[1.51, 2.08]$
LongMemEval	70.4 / 5.7	71.3 / 4.0	$+0.9\,[0.55, 1.24]$	$+1.7\,[1.41, 2.00]$
LoCoMo	68.3 / 5.9	69.3 / 4.1	$+1.0\,[0.63, 1.37]$	$+1.8\,[1.49, 2.10]$
AgentOps-90d	73.6 / 5.2	74.8 / 3.7	$+1.2\,[0.80, 1.60]$	$+1.5\,[1.17, 1.82]$

common-transition gap: -0.3 immediate and -0.8 remainder produce the $+0.7$ native gap. Opposite movement across outcomes is consistent with stateful refusal effects: preserved records can improve a later answer while exposing a different conflict, or prevent a conflict while redirecting retrieval.

The operator-conditional comparison repeats across all four settings in Table 4; all eight paired intervals exclude zero in MERQ's favorable direction. StreamMem's task difference contracts by 46% while its contradiction separation grows from 0.7 to 1.8 points. Because every row is a complete rerun, the table measures behavior under the modified transition rather than selecting favorable outcomes from native logs.

6.2 A COMMON STATE-CONDITIONAL CONSTRUCTION RULE

The final rung applies one state-conditional rule H without assuming identical visited banks or realized tuples. On each controller's current state, the frozen MERQ and RL generators each emit up to $K{=}8$ action tuples. For source generator s, define $\rho_s(a) = (\mathrm{rank}_s(a) - 1)/(K - 1)$ when s proposes a and $+\infty$ otherwise. $H(\mathcal{M}, h)$ deduplicates exact tuples, sorts by $\min_s \rho_s(a)$, breaks ties lexicographically by action type and entry identifiers, and retains the first eight. Each controller recomputes its own features and scores on the resulting set before selecting one feasible action; the common operator is applied only to the selected merge. Thus H is identical across controllers but its realized set can differ because their visited banks differ. Earlier scoring and selection still determine later state, which is the post-generation behavior this rung retains.

Let $Z_c^{r^\star, H}(U)$ apply H at every state visited by controller c before c scores and selects. The final estimand is

$$G_{\mathrm{res}}(r^\star, H) = \mathbb{E}_U[Z_g^{r^\star, H}(U) - Z_b^{r^\star, H}(U)].$$

StreamMem retains $+0.6$ task and $+1.3$ favorable contradiction points with nearly identical quarantine coverage (6.3% versus 6.4%); AgentOps-90d retains $+0.9$ and $+1.0$ points with 7.1% versus 7.2% quarantine. Relative to native-generator common-transition execution, applying H contracts StreamMem by only 0.1 task and 0.5 contradiction points. Generator access explains part of the operator-conditional separation; under the declared MERQ-associated operator and construction rule, the remaining difference comprises scoring, selection, and their induced-state effects on both controlled and natural trajectories.

6.3 CONTROLLER-FAMILY BEHAVIOR UNDER A COMMON TRANSITION

The common transition exposes heterogeneous controller–operator interactions. MemoryBank moves from a 0.7-point task deficit to Compressive Memory to a 0.1-point lead, while MemGPT moves

Table 5: Estimates of $G_{\mathrm{res}}(r^\star, H)$ under common state-conditional construction. Reader, budget, seed, refusal transition, and rule H are shared. Task and quarantine are percentages; contradiction is per decision. In paired rows, ΔT is MERQ minus RL+RER and ΔC is the favorable contradiction reduction.

Condition	Task	Contradiction	Quarantine
StreamMem-Bench			
Rule-H RL+RER	77.6 ± 0.5	4.8 ± 0.3	6.4 ± 0.3
Rule-H MERQ	78.2 ± 0.4	3.5 ± 0.2	6.3 ± 0.3
Paired $\Delta T/\Delta C$	$+0.6\,[0.31, 0.88]$	$+1.3\,[1.02, 1.58]$	—
Native-gen RL+RER	77.7	5.0	6.5 ± 0.3
Full MERQ	78.4	3.2	6.2
AgentOps-90d			
Rule-H RL+RER	73.5 ± 0.5	5.0 ± 0.3	7.2 ± 0.4
Rule-H MERQ	74.4 ± 0.5	4.0 ± 0.2	7.1 ± 0.4
Paired $\Delta T/\Delta C$	$+0.9\,[0.57, 1.22]$	$+1.0\,[0.72, 1.29]$	—
Native-gen RL+RER	73.6	5.2	7.4 ± 0.4
Full MERQ	74.8	3.7	7.0 ± 0.4

Table 6: Refusal regimes on StreamMem-Bench. Native and RER rows for each controller differ only in the declared refusal transition. Selective PPO learns reject actions independently; MERQ uses Equation 2. Contradiction is per decision.

Controller	Refusal regime	Task ↑	Contradiction ↓
MemGPT	native	75.9	4.5
MemGPT+RER	shared operator	76.9 ± 0.5	5.2 ± 0.3
MemoryBank	native	76.1	4.2
MemoryBank+RER	shared operator	77.6 ± 0.5	5.5 ± 0.3
Compressive	native	76.8	4.8
Compressive+RER	shared operator	77.5	5.8
RL-Only	native	77.1	3.9
RL+RER	shared operator	77.7	5.0
Selective PPO	learned reject	78.0 ± 0.5	4.1 ± 0.3
MERQ	declared operator	78.4 ± 0.4	3.2 ± 0.2

from a 0.3-point contradiction deficit to MemoryBank to a 0.3-point lead. The paired changes are $+0.8\,[0.55, 1.04]$ task points and $+0.6\,[0.31, 0.90]$ favorable contradiction points. Compressive–RL task separation contracts from 0.3 to 0.2 points. The effect is therefore not specific to the MERQ–RL pair: changing the transition reveals different compatibility across four controller families.

Selective PPO is the closest independently learned refusal policy. Both coverages are the fraction of maintenance decisions ending in a reject/quarantine transition: $6.1 \pm 0.3\%$ for Selective PPO and $6.2 \pm 0.3\%$ for MERQ, with 431 and 389 ms p95 latency. At these nearly identical rates, MERQ retains $+0.4\,[0.12, 0.69]$ task points and $+0.9\,[0.62, 1.18]$ favorable contradiction points. Appendix F reports all five settings, where Selective PPO reaches $70.8 \pm 0.6/4.8 \pm 0.3$ on LongMemEval, $68.7 \pm 0.6/4.9 \pm 0.3$ on LoCoMo, and $74.2 \pm 0.5/4.4 \pm 0.3$ on AgentOps-90d.

6.4 Stability across calibration, readers, and natural trajectories

Per-setting and per-reader recalibration moves absolute operating points slightly while preserving the comparison: the smallest task advantage is $+0.6\,[0.23, 0.96]$, and the smallest contradiction reduction is $1.3\,[0.96, 1.61]$. Freeze-once evaluation therefore prevents test-setting tuning without changing which controller has the favorable paired tradeoff.

On AMA-Bench's 600 real-trajectory sessions (6,000 probes; 84,672 decisions), Table 7 shows the same pattern. Conflicts arise from source corrections and later tool observations, without injected conflicts.

Table 7: AMA-Bench real-trajectory replication on 600 sessions, 6,000 held-out probes, and 84,672 maintenance decisions. Values are mean $\pm$ 95% CI over ten seeds; contradiction is per decision. The final column reports task gain / favorable contradiction reduction relative to RL+RER.

Controller	Task ↑	Contradiction ↓	Relation to RL+RER
RL+RER	72.9 ± 0.6	5.4 ± 0.3	reference
Selective PPO	73.3 ± 0.5	4.8 ± 0.3	$+0.4/+0.6$
MERQ	73.9 ± 0.5	4.0 ± 0.2	$+1.0\,[0.62, 1.39]/+1.4\,[1.08, 1.70]$

Table 8: Prevalence-weighted human audit and large-scale automatic screen. All 3,600 audited decisions receive three primary labels; bracketed rate intervals are Wilson intervals for majority-label outcomes. Residual contradiction is a post-execution accepted-merge outcome and can include interactions with other bank entries.

Quantity	Count	Rate (95% interval)
Safe proposals accepted	2,507/2,865	$87.5\,[86.2, 88.7]$
Unsafe proposals rejected	679/735	$92.4\,[90.2, 94.1]$
Accepted merges with residual contradiction	67/2,563	$2.6\,[2.1, 3.3]$
Primary labels	10,800	Fleiss' κ=0.86 $[0.85, 0.87]$
Automatic accepted-merge screen	1.02×10^6	2.6 contradiction (descriptive)

6.5 Mechanism reliability and transfer

The 3,600-decision audit yields 87.5% safe recall and 92.4% unsafe recall; its 2.6% residual rate matches the automatic screen. MedNotes is the hardest slice: task is 76.9% versus 78.4% on StreamMem and quarantine 11.8% versus 6.2%.

Mechanism ablations localize the retained difference: fixed τ_q=0.75, uniform proposal scoring, and removing the 24-hour rule cost 0.5/0.8/0.3 task points and add 1.1/0.9/0.6 contradiction points. Without the NLI gate, contradiction rises from 3.2% to 8.7% at nearly unchanged footprint (155.8 to 155.5 slots; Table 17).

7 Discussion

Under the declared MERQ-associated operator and common construction rule, MERQ retains 0.6 task and 1.3 favorable contradiction-burden points on StreamMem and 0.9/1.0 on AgentOps-90d at nearly identical quarantine rates. Because H acts on each visited state while scoring and selection still change later states, these estimands target operator compatibility rather than a controller-free effect. The accepted-merge audit complements the merge-propensity-sensitive endpoint; three readers, five settings, and two natural-trajectory families test stability.

8 Conclusion

Refusal changes the future state on which a memory controller is judged. RER measures that dependence through paired reruns, pathwise accounting, and a common state-conditional construction rule. Under the declared operator, MERQ retains favorable task and contradiction-burden differences while improving complete-policy utility, detail retention, and footprint. Sequential memory evaluation should report both the equalized transition and its rollout residual.

AI Use Statement

Generative AI tools were used to assist with editorial revision and presentation of this manuscript. The authors are responsible for the final submission.

REFERENCES

Aradhye Agarwal, Gurdit Siyan, Yash Pandya, Joykirat Singh, Akshay Nambi, and Ahmed Awadallah. Learning when to act or refuse: Guarding agentic reasoning models for safe multi-step tool use. *arXiv preprint arXiv:2603.03205*, 2026. doi: 10.48550/arXiv.2603.03205. URL https://arxiv.org/abs/2603.03205.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008. URL https://www.jmlr.org/papers/v9/bartlett08a.html.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2206–2240, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

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.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Lin Ge, Jitao Wang, Chengchun Shi, Zhenke Wu, and Rui Song. A reinforcement learning framework for dynamic mediation analysis. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11050–11097, 2023. URL https://proceedings.mlr.press/v202/ge23a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

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, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016. doi: 10.1038/nature20101. URL https://doi.org/10.1038/nature20101.

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.

Wei Hung, Shao-Hua Sun, and Ping-Chun Hsieh. Efficient action-constrained reinforcement learning via acceptance-rejection method and augmented MDPs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AgMpK7z4bz.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116, 2023. doi: 10.18653/v1/2023.emnlp-main.187. URL https://aclanthology.org/2023.emnlp-main.187/.

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, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Ali Modarressi, Abdullatif Köksal, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. MemLLM: Finetuning LLMs to use explicit read-write memory. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=dghM7sOudh.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. Leave no context behind: Efficient infinite context transformers with infini-attention. In *arXiv preprint arXiv:2404.07143*, 2024. doi: 10.48550/arXiv.2404.07143. URL https://arxiv.org/abs/2404.07143.

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 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*. ACM, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylKikSYDH.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1842–1850, 2016. URL https://proceedings.mlr.press/v48/santoro16.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2440–2448, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148, 2016. URL https://proceedings.mlr.press/v48/thomasa16.html.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1410.3916.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TrjbxzRcnf-.

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. 20004–20031, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, 2018. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205/.

Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, Yuandong Tian, and Jishen Zhao. AMA-Bench: Evaluating long-horizon memory for agentic applications. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GoSVL7mLcM.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

A Static Matching Counterexample

A.1 Proof of the proposition

Consider a two-step process with equiprobable exogenous episode types $u \in \{A, B\}$ and initial state $x_1 = 0$. At step one, each controller proposes the same merge. Executing it sets $x_2 = 1$; refusing sets $x_2 = 0$; both choices have zero immediate loss. Controller g refuses on type A and executes on type B, whereas controller b executes on type A and refuses on type B. Both therefore have one-half refusal coverage and identical fixed-sample immediate risk. At step two, the only available action gives reward 1 when $(u, x_2) \in \{(A, 0), (B, 1)\}$ and 0 otherwise. Controller g receives reward 1 on every episode and b receives 0 on every episode despite their matched marginal coverage. A logged post-hoc deletion cannot reconstruct the unobserved step-two state. Under one shared pre-mutation refusal map, however, both controllers reach the same x_2 for each paired episode, so their residual is zero. Thus static coverage leaves the rollout residual unidentified. If execute and refuse induce the same transition and immediate outcome, downstream trajectories are identical and the static and rollout constructions coincide.

B Split, Scoring, and Comparator Protocols

The empirical protocol separates learned-component fitting, operating-point selection, and final scoring by episode or source identity. Controller-training episodes, validation episodes, the 1,000-decision refusal calibration set, the 8,400 StreamMem test episodes, and the 3,600 audit decisions are mutually disjoint. Models and thresholds are frozen before test rollouts. Table 9 records which objects each stage can affect.

Table 9: Split and scoring map. "Unit" is the object kept intact across partitions; no held-out probe text is available to a controller before scoring.

Stage	Partition unit	Selected or measured object	Information boundary
Controller fitting	Source identity / episode	Proposal and PPO parameters	Training streams only
Model selection	Episode	Checkpoint, baseline settings, PPO multiplier	Held-out validation outcomes
Refusal calibration	Proposal decision	$\tau_q^\star$ under the unsafe-action constraint	1,000 labeled decisions only
Primary evaluation	Episode	Task, detail, contradiction, footprint, latency	8,400 locked StreamMem episodes
Transferred evaluation	Released session / episode	Same frozen policies and $\tau_q^\star$	Setting-specific held-out trajectories
Mechanism audit	Proposal decision	Safe/unsafe recall and accepted-merge outcome	Method identity hidden from raters

For StreamMem, source identities are partitioned before any session extension or temporal revision is created. PersonaChat probes target corrected preferences and unrevised rare attributes; API-Bank probes target later tool-return state; Natural Questions probes target the latest supported answer after a chronological revision. A probe is revealed only after its referenced observation and any scheduled correction, and its text never enters the maintained bank. Binary exact-reference correctness is averaged within episode and then across episodes. LongMemEval and LoCoMo preserve released session order and use their released question/reference pairs after the corresponding history prefix.

The automatic contradiction classifier receives the selected feasible pair when it is a merge, its provenance, and the post-transition records. Its binary decision is the union of mutually incompatible preserved sources and an executed merge target that contradicts either source. The evaluator is frozen before policy testing, does not receive method identity, and is not used by the controllers. Rejected merges are scored on their quarantined sources; only decisions whose selected feasible tuple is not a merge contribute zero. The independently labeled accepted-merge audit supplies the complementary conditional measurement.

C Comparator Parity

Every method uses the same frozen reader, retrieval schema, budget state, episode order, and one-reader-pass allowance. Table 10 makes explicit which native signal is mapped into the maintenance interface; wrapping changes only the post-proposal refusal transition.

Table 10: Comparator adaptation and selection protocol. Validation outcomes use the same task/contradiction criteria; the oracle is excluded because it exceeds the latency ceiling.

Method	Maintenance mapping	Learned or selected component	Test-time constraint
LRU / LFU	Recency/frequency chooses keep or archive	None	Metadata-only; no merge target
Periodic summary	Scheduled summary emits compression merge	Period selected on validation episodes	One Llama-3.1-8B summary pass
Compressive / Infini	Native retention state chooses archive; compression state proposes merge	Native scoring with validation-selected adapter settings	One proposal pass; shared action schema
MemGPT / MemoryBank	Native importance and retrieval scores choose keep/archive/merge	Native scoring with validation-selected interface settings	One proposal pass; unsupported split is noop
RL-Only	Scores merge/delete/archive/split/noop	PPO checkpoint on controller training/validation streams	Same action space and 900 ms ceiling
Selective PPO	Adds a learned reject action to the same space	PPO checkpoint and held-out coverage multiplier	No access to q_ψ; reject quarantines and terminates
MERQ	f_ϕ orders all five actions; q_ψ screens the selected merge	Proposal checkpoint plus separately calibrated $\tau_q^\star$	Deterministic order, $K{=}8$, one reader pass

D Complete Threshold and Audit Views

The following sweep exposes the complete utility, contradiction, detail, and quarantine envelope around the frozen operating point; the two audit tables then separate prevalence-weighted behavior from a balanced diagnostic.

Table 11: Seven-point MERQ sweep at $C{=}200$ (ten-seed mean $\pm$ 95% CI). The selected $\tau_q^\star{=}0.847$ is rounded to 0.85 in the row label. Rare-detail loss is per executed merge; contradiction is per maintenance decision.

τ_q	Task ↑	Contrad. ↓	Detail loss ↓	Quarantine
0.50	77.1 ± 0.6	8.3 ± 0.5	10.2 ± 0.4	2.1 ± 0.2
0.60	77.7 ± 0.5	5.9 ± 0.4	7.6 ± 0.4	3.1 ± 0.2
0.70	78.2 ± 0.4	4.5 ± 0.3	6.1 ± 0.3	4.4 ± 0.3
0.80	78.4 ± 0.4	3.6 ± 0.3	5.0 ± 0.3	5.7 ± 0.3
0.85	78.4 ± 0.4	3.2 ± 0.2	4.2 ± 0.2	6.2 ± 0.3
0.90	78.0 ± 0.5	2.8 ± 0.2	3.9 ± 0.2	8.9 ± 0.4
0.95	77.2 ± 0.6	2.5 ± 0.2	3.6 ± 0.2	14.3 ± 0.5

Table 12: Domain slices for the prevalence-weighted 3,600-decision audit. Residual contradiction is among accepted merges.

Slice	n	Unsafe %	Safe recall	Unsafe recall	Residual
PersonaChat-ext	560	16.7	88.3	92.5	2.1
API-Bank	480	17.5	88.2	93.6	2.0
NQ-stream	440	19.5	87.9	92.7	2.4
MedNotes-Stream	360	26.4	85.6	90.6	3.5
LongMemEval	400	19.3	87.6	92.6	2.3
LoCoMo	360	20.1	86.9	91.9	2.5
AgentOps-90d	600	23.1	87.8	92.9	2.8
FinNotes-Stream	400	22.0	86.8	91.6	3.1
All	3,600	20.4	87.5	92.4	2.6

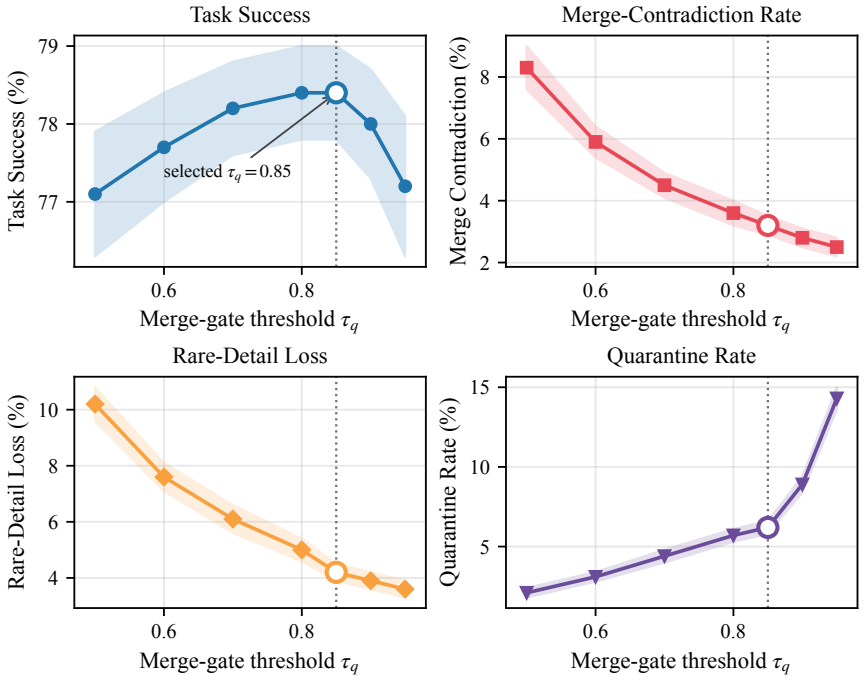


Figure 1: Full-rollout threshold sweep. Bands are ten-seed 95% CIs. The empirical task/contradiction/quarantine tradeoff has a stable knee around the frozen deployment threshold.

Table 13: Balanced 600-decision diagnostic at τ_q=0.85.

Category	Precision	Recall	Contradiction	Detail loss
Accepted safe merges	94.7	87.4	2.1	3.8
Rejected unsafe merges	88.3	93.8	—	—
Adversarial near-duplicates	91.2	94.1	1.4	2.9

At the selected operating region, the balanced diagnostic aligns with the prevalence-weighted audit: unsafe-merge recall remains above 90%, while accepted safe and adversarial near-duplicate cases retain high precision. The two views separate threshold behavior from domain prevalence.

E FROZEN VERSUS RECALIBRATED OPERATING POINTS

These 15 diagnostic cells refit only the scalar threshold on disjoint calibration labels. They leave proposal networks, controller parameters, readers, and evaluation episodes unchanged, isolating sensitivity to a deployment-specific operating point.

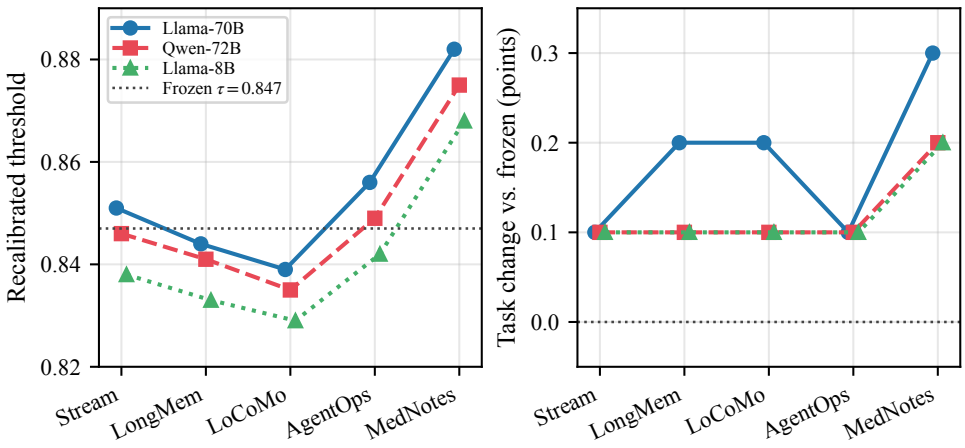


Figure 2: Frozen versus per-setting recalibration over five settings and three readers. Small horizontal offsets separate coincident reader series without changing values. Left: selected thresholds; right: recalibrated-minus-frozen MERQ task changes. Every value appears in Table 14; thresholds span 0.829–0.882 and task changes are at most 0.3 points.

Table 14: Per-setting recalibration with disjoint labels. Each row reports recalibrated τ, MERQ task/contradiction/quarantine, recalibrated RL+RER task/contradiction, and favorable paired Δ task/contradiction. All quantities are ten-seed means; the frozen comparison uses $\tau_q^\star$=0.847.

Reader	Setting	Recal. τ	MERQ T/C/Q	RL+RER T/C	Δ T/C
Llama-70B	Stream	0.851	78.5/3.1/6.0	77.8/4.9	+0.7/+1.8
	LongMemEval	0.844	71.5/3.8/6.7	70.6/5.5	+0.9/+1.7
	LoCoMo	0.839	69.5/3.9/7.1	68.5/5.7	+1.0/+1.8
	AgentOps	0.856	74.9/3.5/7.0	73.8/5.0	+1.1/+1.5
	MedNotes	0.882	77.2/3.8/11.2	76.0/6.0	+1.2/+2.2
Qwen-72B	Stream	0.846	78.0/3.3/6.3	77.3/4.9	+0.7/+1.6
	LongMemEval	0.841	70.7/4.1/6.9	70.0/5.6	+0.7/+1.5
	LoCoMo	0.835	68.7/4.2/7.4	67.9/5.9	+0.8/+1.7
	AgentOps	0.849	74.2/3.8/7.4	73.3/5.1	+0.9/+1.3
	MedNotes	0.875	76.5/4.1/11.4	75.4/6.1	+1.1/+2.0
Llama-8B	Stream	0.838	71.3/3.7/6.8	70.6/5.2	+0.7/+1.5
	LongMemEval	0.833	64.9/4.6/7.2	64.3/6.0	+0.6/+1.4
	LoCoMo	0.829	63.2/4.9/7.8	62.6/6.2	+0.6/+1.3
	AgentOps	0.842	68.0/4.3/7.7	67.4/5.6	+0.6/+1.3
	MedNotes	0.868	70.0/4.5/11.6	69.3/6.4	+0.7/+1.9

F CONTROLLER-FAMILY AND READER MATRICES

The family matrix applies one declared operator across proposal mechanisms; the reader matrix repeats the principal controller comparisons without changing memory policies or thresholds.

Table 15: Full RER controller-family matrix with Llama-3.1-70B. Cells are task/contradiction (%). Native StreamMem rows appear in Table 2; wrapped rows report ten-seed means under the common operator.

Method	Stream	LongMemEval	LoCoMo	AgentOps	MedNotes
MemGPT+RER	76.9/5.2	67.6/6.1	67.4/6.4	72.8/5.9	75.0/7.0
MemoryBank+RER	77.6/5.5	67.9/6.0	67.7/6.3	73.0/5.8	75.2/6.9
Compressive+RER	77.5/5.8	68.3/5.7	68.4/6.2	73.2/5.6	75.5/6.8
RL+RER	77.7/5.0	70.4/5.7	68.3/5.9	73.6/5.2	75.7/6.2
Selective PPO	78.0/4.1	70.8/4.8	68.7/4.9	74.2/4.4	76.1/5.0
MERQ	78.4/3.2	71.3/4.0	69.3/4.1	74.8/3.7	76.9/4.1

Table 16: Auxiliary-reader replication at $C{=}200$ and frozen $\tau_q^\star{=}0.847$. Cells are task/contradiction (%). A wrapper is a rollout intervention and may raise contradiction relative to its native row.

Reader	Method	Stream	LongMem	LoCoMo	AgentOps	MedNotes
Qwen-72B	Compressive	76.4/5.0	65.9/6.9	66.2/7.3	71.8/6.3	74.1/7.5
	Compressive+RER	77.0/5.9	67.7/5.9	67.8/6.3	72.7/5.7	75.0/6.9
	RL-Only	76.6/4.1	67.0/5.8	66.5/6.2	72.5/5.9	74.8/6.0
	RL+RER	77.2/5.1	69.9/5.8	67.7/6.1	73.1/5.3	75.2/6.3
	MERQ	77.9/3.4	70.6/4.2	68.6/4.3	74.1/3.9	76.3/4.3
Llama-8B	Compressive	69.7/5.4	60.5/7.3	60.9/7.6	65.9/6.6	67.9/7.8
	Compressive+RER	70.3/6.2	62.2/6.2	62.5/6.6	66.7/5.9	68.8/7.2
	RL-Only	69.9/4.4	61.6/6.1	61.2/6.5	66.5/6.2	68.6/6.2
	RL+RER	70.4/5.4	64.1/6.2	62.4/6.4	67.2/5.7	69.1/6.6
	MERQ	71.2/3.8	64.8/4.7	63.1/5.0	67.9/4.4	69.8/4.7

G Natural-Trajectory Protocols

AgentOps-90d is a deterministic slice of 500 long-running tool-agent sessions. The reduction funnel is $128{,}473 \to 71{,}864 \to 38{,}412 \to 24{,}091 \to 14{,}652 \to 9{,}318 \to 500$: duration/tool-call filters, explicit user-correction events, tool-return inconsistencies, bank-size bounds, and stratified sampling with seed 7,919. Selected sessions span a mean 14.3 days (median 11.6, maximum 87.4) and end at a mean 287 entries (median 264, maximum 612). Task success uses ten held-out follow-up probes per session. Table 4 reports per-decision merge-target post-state contradiction (3.7% for MERQ); Table 12 reports contradiction among accepted audited merges (2.8% for the AgentOps slice). The denominators answer different questions.

For AMA-Bench, we use only the released real-trajectory component (Zhao et al., 2026). Deterministic stratification across its six source domains yields the 600 sessions, 6,000 probes, and 84,672 decisions reported in Table 7. The selected trajectories have median/IQR duration $18.7\,[8.4, 41.6]$ hours and terminal bank size $312\,[146, 587]$ entries. Each of ten held-out probes per session asks for a fact or tool state that changed after its first observation. Contradiction is scored after every maintenance decision, and no conflict is added beyond source corrections and later tool/environment observations.

H Mechanism, Scaling, and Compute Controls

Component removal, proposal scaling, and compute matching test whether the retained gains can be attributed to a single threshold, additional reader calls, or unconstrained candidate search.

The no-gate variant bypasses HardOK_ψ and q_ψ for feasible merges while retaining proposal ordering and all non-merge transitions. Its StreamMem footprint is 155.5 slots versus 155.8 for full MERQ, so the 8.7% contradiction rate is not explained by additional compression.

The K and N columns are separate series and rows do not pair a recall value with the adjacent latency value. The $N{=}800$ and $N{=}1{,}600$ measurements exceed the 900 ms ceiling and do not enter quality comparisons; reported operating points satisfy the ceiling using end-to-end p95 decision latency.

Table 17: Mechanism isolation. Entries are task/contradiction on StreamMem and AgentOps-90d. Both delta columns use StreamMem relative to RL+RER; ΔC is the favorable contradiction reduction.

Variant	StreamMem	AgentOps	Δ task	ΔC
RL+RER	77.7/5.0	73.6/5.2	0.0	0.0
MERQ fixed $\tau = 0.75$	77.9/4.3	74.1/4.6	+0.2	+0.7
MERQ uniform proposals	77.6/4.1	73.9/4.5	-0.1	+0.9
MERQ without age rule	78.1/3.8	74.5/4.2	+0.4	+1.2
MERQ without NLI gate	—/8.7	—	—	—
Full MERQ	78.4/3.2	74.8/3.7	+0.7	+1.8

Table 18: Two independent diagnostics: proposal recall versus K (left, banks with $N \leq 100$) and median maintenance-call latency versus bank size N (right, 100 calls per N on one A100-80GB). These medians are distinct from the end-to-end p95 values in Tables 2 and 19.

K	Recall (%)	N	Latency (ms)
2	61.2	100	184
4	79.8	200	389
8	92.4	400	822
16	95.1	800	1,745
32	96.3	1,600	3,700

Table 19: Compute-matched comparison at a shared 900 ms/step ceiling. All methods receive one frozen Llama-3.1-70B reader pass.

Method	p95 ms	Passes	Task	Rare detail
Periodic summary	847	1	76.3 ± 0.5	83.4 ± 0.6
Compressive Memory	612	1	76.8 ± 0.5	85.2 ± 0.6
RL-Only	423	1	77.1 ± 0.6	87.3 ± 0.5
MERQ	389	1	78.4 ± 0.4	91.7 ± 0.3

I PER-SEED AND MULTI-BUDGET RESULTS

The per-seed row exposes the paired variation behind the central StreamMem task interval. The figures then place the same policy among slot, byte, latency, contradiction, detail-retention, and taxonomy-shift operating points.

Table 20: Per-seed StreamMem task-success differences at C=200. Each value is MERQ minus RL+RER in task points.

Seed	13	42	97	128	271	314	512	777	1024	2024	Mean
Δ task	+0.7	+0.9	+0.7	+0.7	+0.7	+0.7	+0.8	+0.7	+0.7	+0.7	+0.7

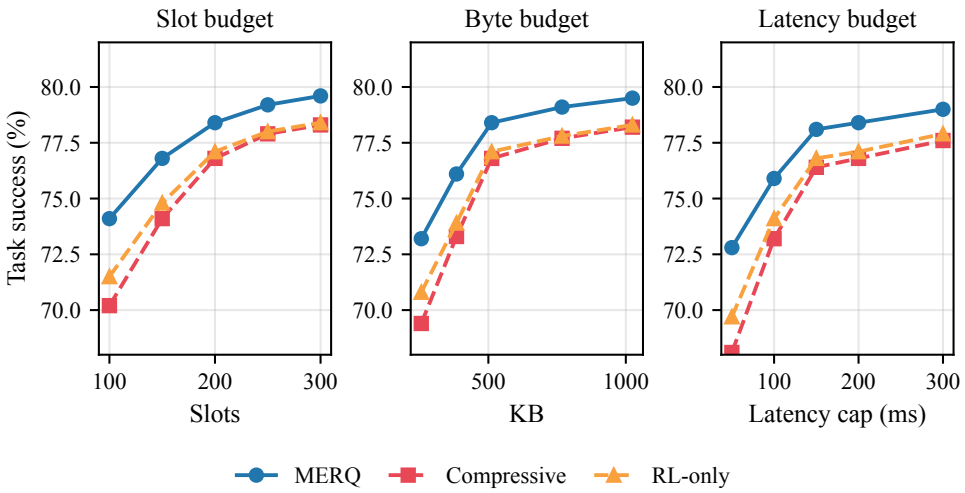


Figure 3: Task success under slot, serialized-byte, and retrieval-latency budgets. MERQ preserves the ordering at the default 200 slots, 512 KB, and 150 ms operating points.

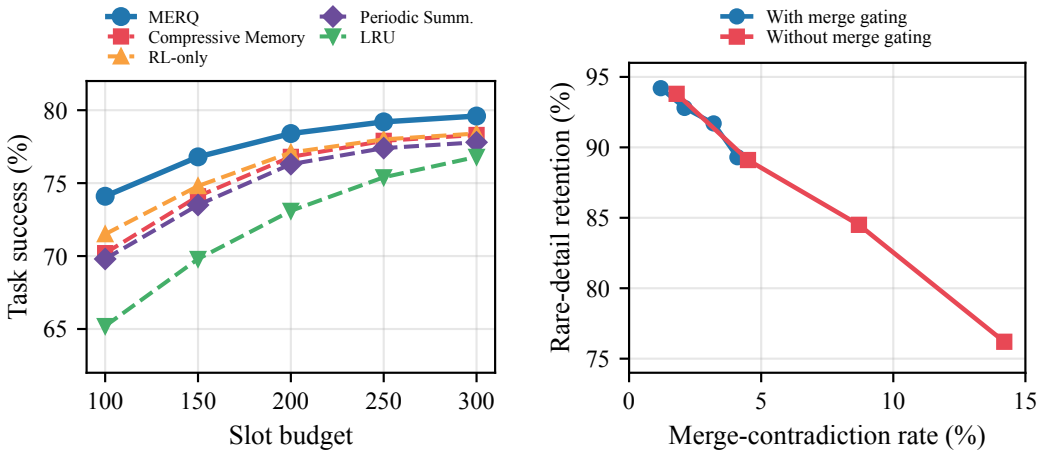


Figure 4: Left: task success across slot budgets. Right: contradiction versus rare-detail retention as compaction moves from conservative to very aggressive. MERQ occupies the non-oracle frontier; its largest task margin over Compressive Memory is 3.9 points at C=100 and narrows to 1.3 at C=300.

Across slot, byte, and latency constraints, MERQ preserves the task ordering rather than relying on one resource definition. The compaction path separately shows that lower contradiction can coincide with higher rare-detail retention, which is inconsistent with a simple indiscriminate-refusal account of the frontier.

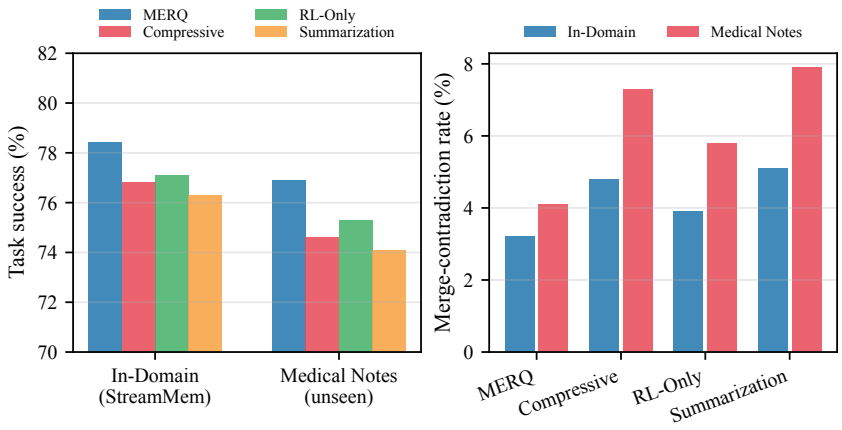


Figure 5: MedNotes-Stream transfer under the same frozen threshold. Task reaches 76.9% versus 78.4% on StreamMem, contradiction is 4.1% versus 3.2%, and quarantine is 11.8% versus 6.2%.

J Implementation and Objective

Experiments run on 8×NVIDIA A100-80GB GPUs with CUDA 12.4, driver 550.54.15, PyTorch 2.3.1+cu121, Python 3.10.14, DeepSpeed 0.14.0, and vLLM 0.5.4. The proposal network is a four-layer Transformer with width 384, six heads, and 4.2M trainable parameters. Adam uses learning rate 5×10^{-4}, β_1=0.9, β_2=0.999, weight decay 10^{-5}, batch size 32, gradient clipping 1.0, and 100,000 steps. Gumbel–Softmax temperature anneals from 1.0 to 0.1 over 50,000 steps; the forward pass takes the hard top-ranked tuple and the backward pass uses its relaxed sample. DeBERTa-v3-large is fine-tuned on MNLI for three epochs with seed 42. Llama-3.1-70B and Qwen2.5-72B use bf16 tensor-parallel four on four A100s; Llama-3.1-8B uses bf16 on one A100.

For action cost $\mathbf{c}(a) = (c^{\text{slot}}, c^{\text{byte}}, c^{\text{lat}})$ and cumulative footprint $\mathbf{C}_t$, deployment scalarization $\mathbf{w}$ requires $\mathbf{w}^\top \mathbf{C}_t \le B(\mathbf{w})$. The controller maximizes

$$J(\theta) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t \left(r_t^{\text{task}} - \lambda_1 c_t^{\text{corrupt}} - \lambda_2 c_t^{\text{maint}}\right)\right], \tag{8}$$

Here r_t^{task} is reader correctness on a scheduled probe (zero otherwise), c_t^{corrupt} is the training-time post-state incompatibility indicator for a selected merge, and $c_t^{\text{maint}} = \mathbf{w}^\top \mathbf{c}(a_t)$ is normalized by its declared budget. We use γ=0.99, λ_1=1.0, and λ_2=0.1; slot, byte, and latency points set $\mathbf{w}$ to $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$, respectively.

K MERQ Maintenance Algorithm

Algorithm 1 implements Equation 1: rejection executes qnoop and ends the decision rather than continuing down the proposal list.

Algorithm 1 MERQ maintenance decision

Require: Bank $\mathcal{M}_t$, budget C, proposal network f_ϕ, gate q_ψ, history h_t, threshold $\tau_q^\star$

Generate ordered proposal list $\mathcal{L}_t = \{a_k\}_{k=1}^{K} \leftarrow f_\phi(\mathcal{M}_t, h_t)$
for $k = 1$ to K **do**
  if targets are not live, schema/arity/type is invalid, or execution violates C **then**
    continue
  end if
  if $a_k = \mathrm{merge}(i, j)$ and $\Omega_{\tau_q^\star}(a_k) = 1$ **then**
    $a_t \leftarrow \mathrm{qnoop}(i, j)$; return $P(\mathcal{M}_t, a_t, u_t)$
  end if
  $a_t \leftarrow a_k$; return $P(\mathcal{M}_t, a_t, u_t)$
end for
$a_t \leftarrow \mathrm{noop}$; return $P(\mathcal{M}_t, a_t, u_t)$

L Compute Ledger

The ledger separates core quality studies from the interventions and sensitivity analyses added to identify refusal, proposal, reader, and trajectory effects.

Table 21: Compute accounting. The 11 core experiment families account for 880 runs and 942.0 A100-GPU-hours; the intervention studies extend the total to the reported aggregate. Per-run values are A100-GPU-hours, including tensor-parallel devices.

Experiment family	Runs	GPU-h/run	GPU-hours
Main table	100	1.90	190.0
Slot-budget Pareto sweep	200	0.50	100.0
Threshold envelope	70	0.70	49.0
Byte/latency budgets	80	0.60	48.0
MedNotes transfer	50	1.10	55.0
Scaling study	25	0.60	15.0
Human/automatic/mechanism audits	25	3.00	75.0
LongMemEval/LoCoMo transfer	100	1.50	150.0
Compressive/RL wrapper isolation	80	0.50	40.0
Auxiliary-reader replication	100	1.30	130.0
AgentOps-90d	50	1.80	90.0
Rule-H isolation	40	0.70	28.0
Controller-family RER audit	100	0.80	80.0
Selective PPO	50	1.20	60.0
Per-setting/reader recalibration	150	0.65	97.5
AMA-Bench real trajectories	30	1.60	48.0
Total	1,250		1,255.5