PERSISTENT-STATE ACTIVATION BEYOND UTILITY: SCALE REPRESENTATION AND EVIDENCE FOR NON-COMPENSATORY CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent memory creates a control problem that retrieval relevance alone does not solve: a memory can remain highly useful after an update, deletion, or revocation makes it inadmissible for the current answer. We formalize this as a separation between *utility* and *authority*. A fixed finite penalty applied to an unnormalized utility score cannot guarantee exclusion under arbitrary positive-affine reparameterization of that score; by contrast, rank-normalized compensation is scale-invariant and therefore forms a stronger empirical comparator. Our prospectively frozen TIDE/LongMemEval primary was quarantined before a valid HELDOUT comparison because the materialized TIDE adapter conflated historical age with query-relative inadmissibility and the aligned LongMemEval split left no DEV set for the predeclared penalty selection. We therefore report a post-primary replacement diagnostic on Memora Remembering, where update/delete operations provide item-level forgetting state. On Qwen3-8B, DEV selected $\lambda = 0.6$ from a ten-point normalized SOFT family. Across 185 HELDOUT units in 28 dependency clusters, HARD exclusion yields 4.04% balanced construct error versus 19.66% for locked SOFT, a paired difference of 15.61 points with a 20,000-replicate cluster-bootstrap 95% interval of [13.07,18.76]. The effect is driven primarily by forgotten-value leakage while current-value recall is preserved. This is same-Q operator-comparison evidence, not a universal claim that scalar control fails, not an evaluation of learned authority inference, and not an independent downstream-harm endpoint.

1 INTRODUCTION

Long-lived language-model agents increasingly externalize memory: past exchanges are stored, retrieved, summarized, ranked, and reinserted into later prompts. This architecture appears in retrieval-augmented generation (Lewis et al., 2020), nearest-neighbor language models (Khandelwal et al., 2020), retrieval-enhanced transformers (Borgeaud et al., 2022), memorizing transformers (Wu et al., 2022), explicit long-term memory modules (Wang et al., 2023), and interactive agents with persistent experience stores (Park et al., 2023; Zhong et al., 2024). Long-context models do not eliminate the control problem: even when information can be placed in context, models use long contexts unevenly and are sensitive to position and distraction (Liu et al., 2024).

The usual objective in these systems is to recover information that is relevant, similar, recent, useful, or predictive. That objective is incomplete when memory changes over time. A deleted task, superseded preference, corrected address, revoked instruction, or obsolete value may remain semantically close to the query and highly predictive of the user's history while no longer being permitted to determine the current answer. Long-term-memory benchmarks increasingly expose this update problem (Maharana et al., 2024; Wu et al., 2025; Uddin et al., 2026; Patel, 2026; Tao et al., 2026). The central distinction in this paper is therefore between *utility*: how useful a memory would be if used, and *authority*: whether it is currently admissible to use.

This separation has deep precedents outside LLM memory. Security architectures distinguish possession of information from permission to use it, from classic protection and information-flow models (Saltzer and Schroeder, 1975; Denning, 1976) through role-based and usage control (Sandhu et al.,

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.

1996; Ferraiolo et al., 2001; Park and Sandhu, 2004; Zhang et al., 2005). In particular, usage control emphasizes decision continuity, mutability, and immediate revocation rather than a one-time relevance judgment. Decision theory likewise distinguishes compensatory aggregation from lexicographic or otherwise non-compensatory preference structures (Fishburn, 1974; 1976). Measurement theory warns that conclusions can depend on which transformations of a numerical scale are meaningful (Stevens, 1946; Narens, 1981). Constrained optimization and safe control provide a complementary lesson: constraints and objectives need not be represented by the same mechanism (Han and Mangasarian, 1979; Di Pillo and Grippo, 1989; Garcia and Fernandez, 2015; Achiam et al., 2017; Alshiekh et al., 2018; Brunke et al., 2022).

Our question is deliberately narrower than "hard gating is better." We ask: *holding the authority state Q fixed, does non-compensatory exclusion change forgetting-sensitive answer reliability relative to a strong finite compensatory controller?* The qualification matters. Exact penalties can enforce constraints under problem-dependent conditions, and a scalar controller with an explicit threshold can reproduce hard feasibility. We therefore compare HARD exclusion to a scale-normalized finite SOFT family rather than to an obviously fragile raw-score penalty.

The execution history is itself part of the scientific result. We prospectively froze a TIDE (Sobhani et al., 2026)/LongMemEval (Wu et al., 2025) matrix over Qwen3, Llama-3.1, and Gemma-3 (Yang et al., 2025; Grattafiori et al., 2024; Gemma Team et al., 2025). Before a defensible HELDOUT comparison, construct audits showed that our TIDE adapter confused historical age with query-relative inadmissibility, while the aligned LongMemEval split left no DEV set for the predeclared λ selection. We quarantined those cells rather than convert a protocol failure into a nominal positive result.

We then evaluated a replacement Memora Remembering cell. Memora explicitly includes updates/deletions and scores forgetting-aware behavior (Uddin et al., 2026). The replacement result is large, but its evidence class is intentionally limited because the utility proxy and deterministic scorer were finalized during replacement development. The paper therefore makes three contributions: (1) a scale-representation argument explaining why fixed raw finite penalties do not define representation-independent exclusion; (2) a cross-literature formulation of persistent-memory authority as an admission problem distinct from retrieval utility; and (3) an auditable replacement experiment showing a substantial same-Q difference between HARD exclusion and the DEV-selected member of a normalized finite-SOFT family.

2 Authority Before Utility

Let a query expose candidate memory items $i \in \mathcal{M}$. Each item has a raw utility score $U_i \in \mathbb{R}$ produced by some retrieval or ranking mechanism and an authority state $m_Q(i) \in \{0, 1\}$ supplied by a frozen object Q, where $m_Q(i) = 1$ means that item i is inadmissible for the present answer. We condition on Q throughout. Learning or inferring Q from raw dialogue is a separate problem.

Hard admission. HARD defines the feasible set

$$\mathcal{F}(Q) = \{i \in \mathcal{M} : m_Q(i) = 0\}$$

and ranks only within $\mathcal{F}(Q)$. This makes admissibility logically prior to utility ranking.

Finite compensation. A common alternative is to retain all candidates and subtract a finite penalty,

$$S_\lambda(i) = U_i - \lambda m_Q(i), \qquad \lambda < \infty.$$

An inadmissible item can then survive whenever its utility advantage exceeds the penalty. The issue is not merely whether one particular λ is numerically large enough; it is whether the exclusion claim survives admissible reparameterizations of the utility representation.

Proposition 1 (Raw finite-penalty rescaling)**.** *Fix any finite λ. Suppose an inadmissible candidate j has positive raw-utility advantage $U_j - U_k > 0$ over some admissible candidate k. Then there exists a positive-affine transformation $U'_i = aU_i + b$, $a > 0$, under which j outranks k according to $U'_i - \lambda m_Q(i)$.*

Proof. The transformed score difference is $a(U_j - U_k) - \lambda$. Choosing $a > \lambda/(U_j - U_k)$ makes this quantity positive; the additive constant b cancels. Hence a fixed raw penalty cannot guarantee

exclusion across positive-affine representations whenever an inadmissible item has a positive utility advantage. □

The proposition is a representation statement, not an impossibility theorem for scalar control. Measurement theory distinguishes scales by transformations under which empirical statements remain meaningful (Stevens, 1946; Narens, 1981). If only ordering matters, any increasing transformation preserves ranks, whereas a fixed additive penalty has units tied to the chosen numerical representation. This is why the paper does not treat an arbitrary raw-score penalty as a strong baseline.

Scale-normalized compensation. Let $R_i \in [0, 1]$ be a rank-normalized utility proxy. We test

$$S_\lambda^{\mathrm{norm}}(i) = R_i - \lambda m_Q(i).$$

Because R_i depends only on order, positive-affine transformations of raw U do not change the normalized ranking. The finite-SOFT family is therefore invariant to the specific positive-affine parameterization attacked by Proposition 1. HARD remains distinct because infeasible items are removed rather than merely displaced in a shared score.

Relation to exact penalties. Exact-penalty theory establishes that finite penalties can recover constrained optima under regularity conditions and problem-dependent parameter bounds (Han and Mangasarian, 1979; Di Pillo and Grippo, 1989). Our claim is compatible with that literature. A thresholded scalar scheme can also reproduce the same feasible set as HARD. The empirical question here is only HARD versus the explicitly tested normalized finite family under common Q, candidates, model, prompt, and scorer.

3 Related Work

3.1 External memory, long context, and persistent agents

Retrieval-augmented language modeling makes external state directly useful at inference time. kNN-LM interpolates a neural language model with nearest-neighbor retrieval (Khandelwal et al., 2020); RAG combines parametric generation with a non-parametric document index (Lewis et al., 2020); RETRO retrieves from a very large text database during generation (Borgeaud et al., 2022); and Memorizing Transformers add approximate nearest-neighbor access to past internal representations (Wu et al., 2022). LongMem extends this line with a decoupled long-term memory network (Wang et al., 2023). These systems establish the value of externally accessible state, but their main objective is to find useful information, not to represent whether useful information remains authorized after mutation.

Persistent agents make the temporal problem unavoidable. Generative Agents store experiences, synthesize reflections, and retrieve memories to guide later behavior (Park et al., 2023). MemoryBank explicitly adds long-term user memory and time-sensitive memory updating (Zhong et al., 2024). LoCoMo evaluates conversations spanning many sessions and long-range temporal/causal dependencies (Maharana et al., 2024); LongMemEval isolates extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention (Wu et al., 2025). Memora goes further by explicitly evaluating remembering and forgetting under updates/deletions (Uddin et al., 2026). Recent diagnostics such as Supersede and MemConflict target supersession and conflict handling (Patel, 2026; Tao et al., 2026), while authority-oriented retrieval work asks which source remains controlling rather than merely similar (Bacellar, 2026). This progression motivates our focus on current admissibility rather than memory capacity alone.

Long context is not a substitute for admission control. Liu et al. (2024) show that models can use the same relevant evidence differently depending on its position in a long context. In our setting, this makes context occupancy a potential mechanism: excluding inadmissible items changes not only their direct availability but also which admissible items occupy limited prompt space. We therefore treat context reshaping as an unresolved alternative explanation rather than attributing the full observed effect to a pure logical-gating mechanism.

3.2 Forgetting, updating, and deletion

The continual-learning literature studies a different but adjacent failure: new learning can destroy performance on old tasks. Elastic weight consolidation made catastrophic forgetting a prominent modern benchmark problem (Kirkpatrick et al., 2017), and major reviews organize the resulting methods and scenarios (Parisi et al., 2019; De Lange et al., 2022; van de Ven et al., 2022). Our problem is almost the inverse: some old information should remain retrievable, while other old information should cease to control the answer after a state change. Thus "remember everything" and "forget obsolete state" are separate requirements.

Machine unlearning provides a second neighboring literature. Early work framed system forgetting as removal of data and its lineage (Cao and Yang, 2015); later surveys distinguish deletion from removal of a training sample's influence on a learned model (Xu et al., 2024). Persistent external memory is operationally simpler in one respect—an item can be withheld without retraining model weights—but harder in another: deleted or superseded content may still be present in stores, summaries, replicas, or retrieved context. Our experiment does not claim certified unlearning. It tests answer-time admission under supplied item-level authority labels.

3.3 Authorization, non-compensation, and runtime constraints

Computer-security research has long treated access as a permission relation rather than a relevance score. Classic protection design separates authorization mechanisms from application utility (Saltzer and Schroeder, 1975); lattice information-flow models impose structural constraints on permissible flows (Denning, 1976). RBAC organizes permissions through roles (Sandhu et al., 1996; Ferraiolo et al., 2001). Most directly relevant, UCON extends access control with mutable attributes, obligations, conditions, decision continuity, and ongoing revocation (Park and Sandhu, 2004; Zhang et al., 2005). Persistent LLM memory differs in object and semantics, but the architectural lesson is close: a long-lived system needs a control state that can change after information was once legitimate.

Non-compensatory choice has an established decision-theoretic lineage. Lexicographic rules give priority to one criterion such that gains elsewhere cannot compensate for failure on the higher-priority criterion (Fishburn, 1974); Fishburn separately formalized noncompensatory preferences (Fishburn, 1976). Safe reinforcement learning likewise distinguishes reward optimization from safety constraints (Garcia and Fernandez, 2015; Achiam et al., 2017); shielding enforces specifications by filtering actions around a learned policy (Alshiekh et al., 2018); and modern safe-learning surveys treat certified or runtime constraint enforcement as a distinct design axis (Brunke et al., 2022). Our contribution is not the generic discovery that constraints can dominate objectives. It is the scale-representation argument plus a same-Q persistent-memory operator comparison under forgetting-aware evaluation.

4 Execution Audit and Replacement Design

4.1 Why the frozen primary was quarantined

The prospectively frozen primary crossed TIDE and LongMemEval with Qwen3, Llama-3.1, and Gemma-3, plus a frontier-model stress test. The matrix was intended to test the same control distinction across temporal-version and conversational-memory settings. It did not survive construct/procedure audit.

TIDE is a version-resolution benchmark in which earlier official text can remain correct for a historically scoped query (Sobhani et al., 2026). Under our materialized date adapter, answer-bearing historical evidence was labeled inadmissible in units from both DEV and HELDOUT. The adapter therefore partially encoded "old" as "unauthorized," which is not the construct we intend. For aligned LongMemEval, the frozen dependency split placed all aligned units in HELDOUT, leaving no DEV examples for the predeclared per-cell λ selection. Running those cells anyway would have mixed a construct error with a protocol deviation. We therefore report neither benchmark as a completed primary cell.

This quarantine matters for interpretation. The replacement analysis below is not presented as if it were the originally preregistered six-cell experiment. It is a post-primary diagnostic developed after the primary failed its own validity checks.

4.2 MEMORA MATERIALIZATION AND AUTHORITY STATE

We pinned the public Memora repository to the commit identified in the evidence object (short form `a6493188efc8...87da30ff`; Appendix A). Memora's Remembering task includes current information and explicit forgetting evidence created by update/delete operations (Uddin et al., 2026). We materialize explicit current evidence as $q = 0$ and forgotten evidence as $q = 1$. The adapter yields 200 units, 998 current candidates, and 1,476 forgotten candidates.

Dependency clusters are defined as period×persona before outcome scoring. Hashing these clusters produces 15 DEV and 185 HELDOUT units with no cluster leakage. This split is fixed for the replacement experiment.

The source-accounting edge case is deterministic. Of 200 Remembering questions, 196 have nonempty source `forgotten_items`. Exactly 180 contain value-bearing forgotten items that the evidence-value adapter can materialize; 16 contain session identifiers without a forgotten value string. Four further questions have empty forgotten-item lists. Because the scorer tests textual presence/absence of values, the 16 session-only cases cannot contribute a value-level forgetting target. The evidence object enumerates all 196 IDs and their classification. A post-hoc sensitivity restricted to HELDOUT units with materialized forgotten values (167 units) yields a 17.30-point paired separation: HARD 1.78% versus SOFT 19.08% error.

4.3 CONTROLLERS AND LOCKED MODEL

HARD removes every $q = 1$ candidate before ranking. SOFT retains all candidates and ranks

$$R_i - \lambda q_i, \qquad \lambda \in \{0, .05, .10, .20, .40, .60, .80, .90, .95, .99\},$$

where $R_i \in [0, 1]$ is a rank-normalized session-recency proxy. At most 12 candidates enter the prompt. The executed checkpoint is Qwen3-8B (Yang et al., 2025), run deterministically with thinking disabled and a 256-token output ceiling.

DEV contains 15 units. The complete SOFT balanced-error surface is

$$.2783, .2650, .2550, .2325, .2134, \mathbf{.1912}, .1981, .2118, .2118, .2118,$$

so the locked selection rule chooses $\lambda = .6$; ties favor the smaller λ. HARD DEV error is .0733. The small DEV set and ten-way selection are important limitations rather than incidental details.

4.4 MEASUREMENT AND ESTIMAND

The deterministic scorer extracts target values that should be current and target values that should be forgotten. Current recall is the fraction of current target values appearing in the answer. Forgetting absence is the fraction of forgotten target values absent from the answer. We report balanced construct error

$$E = \tfrac{1}{2}[(1 - \text{current recall}) + (1 - \text{forgetting absence})].$$

This endpoint deliberately couples answer content to the forgetting construct. It is not an independent measure of downstream harm, user satisfaction, factuality in general, or policy compliance outside the supplied Q.

The causal object is correspondingly narrow. HARD and SOFT share Q, candidate pool, checkpoint, prompt template, generation settings, and scorer, and differ in how Q enters admission/ranking. But HARD can change prompt occupancy by deleting candidates, so the locked experiment does not isolate a pure "logical non-compensation" mechanism from benefits of freeing context slots or changing distractor composition.

The replacement pipeline is also not fully prospective. The session-recency utility proxy was materialized during replacement execution, and the deterministic scorer was finalized after DEV outputs existed. HELDOUT was not used to design either component, but this DEV adaptivity is sufficient to classify the result as operator-comparison evidence rather than prospective causal confirmation.

Table 1: Locked Qwen3-8B HELDOUT replacement cell (185 units; 28 period×persona dependency clusters).

Metric	HARD	SOFT ($\lambda = .6$)
Balanced construct error ↓	4.04%	19.66%
Current recall ↑	91.92%	89.52%
Forgetting absence ↑	100.00%	71.17%

5 Results

Table 1 reports the locked HELDOUT comparison. Across 185 units in 28 dependency clusters, HARD balanced construct error is 4.04% and SOFT($\lambda = .6$) error is 19.66%, giving a paired SOFT-minus-HARD difference of 15.61 percentage points. A 20,000-replicate percentile bootstrap over the 28 clusters gives a 95% interval of [13.07,18.76] points.

The decomposition is important. Current recall differs by 2.40 points, whereas forgetting absence differs by 28.83 points. Thus the aggregate gap is primarily a forgetting-sensitive answer/admission effect rather than a large gain in recalling current values. At the unit level HARD is better on 95 units and tied on 90; SOFT is better on none under this scorer.

Cluster robustness. Cluster-mean paired effects range from 6.67 to 40.00 points. Leave-one-cluster-out aggregate effects range from 14.94 to 16.13 points, so no single period×persona cluster explains the aggregate result. Because personas recur across periods, we also performed post-primary re-clustering sensitivities. A 20,000-replicate persona-only bootstrap (10 clusters) gives an approximate 95% interval of [13.80,18.01] points; a period-only bootstrap (3 clusters) gives approximately [10.76,22.66]. Both preserve the sign but are descriptive robustness checks, not a redefinition of the primary inferential unit.

6 Interpretation

The result is most naturally read as evidence about control architecture under known state. Once an item is labeled inadmissible, a controller can either remove it from the candidate set or continue to place it on a common utility scale and attempt to suppress it. The latter remains a legitimate design when the scale and admission threshold are explicitly engineered; exact-penalty theory makes that clear. What the present experiment shows is narrower: under the tested normalized compensatory family, the best DEV-selected finite penalty still admits substantially more forgotten-value evidence into answers than hard exclusion, without buying higher current-value recall.

This distinction helps connect persistent memory to older control literatures. In access control, permission is not generally treated as another feature whose weight can be compensated by task utility (Saltzer and Schroeder, 1975; Park and Sandhu, 2004). In safe control, a runtime shield or constrained optimizer can keep feasibility separate from reward maximization (Achiam et al., 2017; Alshiekh et al., 2018; Brunke et al., 2022). And in noncompensatory decision rules, a higher-priority condition may block a choice regardless of improvements on lower-priority attributes (Fishburn, 1974; 1976). Persistent LLM memory creates an analogous systems question: should evidence that was once legitimate remain eligible merely because it is useful now?

The answer is not automatically "always hard-delete." Authority can be uncertain, contested, hierarchical, or time-scoped. A system may need provenance, appeal, expiry, revocation, and conflict-resolution mechanisms rather than a single Boolean bit. Our known-Q experiment intentionally avoids that upstream problem. Its role is to isolate what happens after an authority state has already been supplied.

7 Limitations and Claim Boundary

The central quantitative claim is local. It applies to one Qwen3-8B Memora Remembering replacement cell, the supplied item-level Q, a 12-candidate context budget, a session-recency rank-

ing proxy, and a deterministic evidence-value scorer. It does not establish generalization across models, benchmarks, learned authority estimators, prompt budgets, retrieval systems, or alternative scalar/admission controllers.

The DEV set is small: 15 units select among ten λ values. The HELDOUT bootstrap conditions on the selected λ and therefore does not propagate DEV-selection uncertainty. A repeated-split or nested-selection design would be stronger.

The endpoint is construct-coupled. When HARD has oracle-known $q = 1$ labels, 100% forgotten-value absence is partly expected because those values are removed from the prompt. The informative accompanying observation is that current recall is not reduced in this cell. Even so, balanced construct error is not an independent downstream-harm measure, and the experiment does not quantify user-level consequences of a forgotten value reappearing.

HARD also changes context composition. Candidate deletion can free prompt slots and reduce distraction. Because long-context utilization is position- and context-sensitive (Liu et al., 2024), the experiment cannot identify how much of the effect arises from logical non-compensation versus context reshaping. A stronger factorial design would hold context occupancy fixed by replacing excluded items with matched admissible distractors or explicit null slots.

The experiment assumes correct authority labels. Inferring whether a memory is actually superseded, deleted, revoked, jurisdictionally controlling, or merely historically scoped is a separate learning and governance problem. Our failed TIDE adapter demonstrates why this distinction matters: historical age is not itself inadmissibility. Future work should evaluate learned or uncertain Q against independently adjudicated authority state rather than folding authority inference into the controller and calling the result a gating test.

Finally, the replacement result is not the prospectively frozen primary. The original TIDE/LongMemEval matrix failed construct/split checks and remains quarantined. The Memora utility proxy and scorer were finalized during replacement development. This lowers the evidentiary status of the replacement cell even though HELDOUT was preserved during that development.

8 Conclusion

Persistent memory creates a control problem that memory capacity and retrieval quality alone cannot solve. Information can remain relevant after it ceases to be admissible. Separating authority from utility makes that failure mode explicit and connects LLM memory to mature ideas in measurement theory, noncompensatory decision rules, access control, exact penalties, and safe runtime constraint enforcement.

A fixed finite penalty on an unnormalized utility representation cannot guarantee exclusion across positive-affine rescalings; rank normalization removes that specific representation defect and therefore supplies a stronger compensatory baseline. In our post-primary Memora/Qwen3-8B replacement cell, HARD exclusion substantially reduces forgetting-sensitive answer/admission error relative to the DEV-selected normalized finite-SOFT controller while preserving current-value recall. The result supports further study of authority-aware memory control, but it does not establish that every scalar controller is inferior, that authority can be inferred reliably, or that the measured gap is itself a downstream-harm estimate.

AI Use Statement

Generative AI tools were used in this work to help develop conceptual frameworks; formulate and refine mathematical claims; assist with proof development; propose or refine hypotheses; design and critique research methodology and experiments; implement methods; clean or reformat data and benchmark artifacts; interpret experimental results; draft and edit manuscript text; search, summarize, and organize literature; identify research gaps; improve readability; format references; and suggest manuscript structure. They were not used to generate synthetic research datasets, transcribe research recordings, formulate survey/interview questions, or generate scientific figures used as evidence.

All AI-assisted research outputs were subjected to artifact-level or author review as appropriate. Executed benchmark results were checked against materialized call plans, row-level outputs, deterministic scorers, hashes, and independent aggregate recomputation. Literature and citation claims were checked against source records. The authors take responsibility for the final manuscript, claims, code, analyses, citations, and artifacts.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *Proceedings of ICML*, PMLR 70:22–31, 2017.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):2669–2678, 2018. doi:10.1609/aaai.v32i1.11797.

Andre Bacellar. Controlling Authority Retrieval: A Missing Retrieval Objective for Authority-Governed Knowledge. arXiv:2604.14488, 2026.

Sebastian Borgeaud et al. Improving language models by retrieving from trillions of tokens. *Proceedings of ICML*, PMLR 162:2206–2240, 2022.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022. doi:10.1146/annurev-control-042920-020211.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. *IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi:10.1109/SP.2015.35.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi:10.1109/TPAMI.2021.3057446.

Dorothy E. Denning. A lattice model of secure information flow. *Communications of the ACM*, 19(5):236–243, 1976. doi:10.1145/360051.360056.

Gianni Di Pillo and Luigi Grippo. Exact penalty functions in constrained optimization. *SIAM Journal on Control and Optimization*, 27(6):1333–1360, 1989.

David F. Ferraiolo, Ravi Sandhu, Serban Gavrila, D. Richard Kuhn, and Ramaswamy Chandramouli. Proposed NIST standard for role-based access control. *ACM Transactions on Information and System Security*, 4(3):224–274, 2001. doi:10.1145/501978.501980.

Peter C. Fishburn. Lexicographic orders, utilities and decision rules: A survey. *Management Science*, 20(11):1442–1471, 1974. doi:10.1287/mnsc.20.11.1442.

Peter C. Fishburn. Noncompensatory preferences. *Synthese*, 33:393–403, 1976. doi:10.1007/BF00485453.

Javier Garcia and Fernando Fernandez. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

Gemma Team et al. Gemma 3 technical report. arXiv:2503.19786, 2025.

Aaron Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

S.-P. Han and O. L. Mangasarian. Exact penalty functions in nonlinear programming. *Mathematical Programming*, 17:251–269, 1979. doi:10.1007/BF01588250.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *ICLR*, 2020.

James Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi:10.1073/pnas.1611835114.

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

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

Adyasha Maharana et al. Evaluating very long-term conversational memory of LLM agents. *Proceedings of ACL*, pp. 13851–13870, 2024. doi:10.18653/v1/2024.acl-long.747.

Louis Narens. On the scales of measurement. *Journal of Mathematical Psychology*, 24(3):249–275, 1981. doi:10.1016/0022-2496(81)90045-6.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. doi:10.1016/j.neunet.2019.01.012.

Jaehong Park and Ravi Sandhu. The UCONABC usage control model. *ACM Transactions on Information and System Security*, 7(1):128–174, 2004. doi:10.1145/984334.984339.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *Proceedings of UIST*, Article 2, 2023. doi:10.1145/3586183.3606763.

Vedant Patel. Supersede: Diagnosing and training the memory-update gap in LLM agents. arXiv:2606.27472, 2026.

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

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.

Mahbub E. Sobhani, Md. Faiyaz Abdullah Sayeedi, Fahmid Hasan Chowdhury, Md. Adnan Arefeen, Farig Sadeque, Md. Faizul Bari, and Swakkhar Shatabda. Time present and time past: Benchmarking large language models on temporally evolving document understanding. arXiv:2608.08512, 2026.

S. S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946. doi:10.1126/science.103.2684.677.

Zhen Tao et al. MemConflict: Evaluating long-term memory systems under memory conflicts. arXiv:2605.20926, 2026.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. *Findings of ACL*, pp. 26814–26841, 2026. doi:10.18653/v1/2026.findings-acl.1337.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4:1185–1197, 2022. doi:10.1038/s42256-022-00568-3.

Weizhi Wang et al. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36, 2023.

Yuhuai Wu, Markus N. Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *ICLR*, 2022.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. *ICLR*, 2025.

Heng Xu et al. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1), Article 9, 2024. doi:10.1145/3603620.

An Yang et al. Qwen3 technical report. arXiv:2505.09388, 2025.

Xinwen Zhang, Francesco Parisi-Presicce, Ravi Sandhu, and Jaehong Park. Formal model and policy specification of usage control. *ACM Transactions on Information and System Security*, 8(4):351–387, 2005. doi:10.1145/1108906.1108908.

Wanjun Zhong et al. 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.

A Replacement Evidence Object

The replacement evidence object is included with the submission supplementary material as `ICLR2_MEMORA_QWEN_EVIDENCE_OBJECT_20260917.zip`, SHA-256 `5f04ba5e682ea4ab83e2d172dbb9c3f0dedd8ce453a716e3978b596c3aeb3981`. It contains the exact Memora commit binding, adapter and plan builders, split/cluster map, item-level materialized units and Q labels, complete DEV curve, scorer implementations, call plans, raw local-execution rows, row-level scores, HELDOUT aggregate, 196-vs-180 reconciliation, materialized-forgetting sensitivity, cluster robustness, and a per-file SHA-256 manifest. The executed model checkpoint is Qwen3-8B; the original Llama/Gemma planned cells were not executed and are not claimed.

B Exact Model and Runtime Binding

The executed local checkpoint is Qwen3-8B, Hugging Face revision `b968826d9c46dd6066d109eabc6255188de91218`. The audited runtime used Python 3.12.3, PyTorch 2.14.0+cpu, Transformers 4.57.6, FastAPI 0.141.1, Uvicorn 0.53.0, and OpenAI Python 3.14.1. Generation used BF16 model weights, `do_sample=False`, thinking disabled, and a 256-token output ceiling. Exact same-unit HARD/SOFT requests were deterministically deduplicated; no cross-unit response reuse was observed.

C DEV Surface and Accounting Reconciliation

The ten SOFT DEV errors for $\lambda = (0, .05, .10, .20, .40, .60, .80, .90, .95, .99)$ are $(.2783, .2650, .2550, .2325, .2134, .1912, .1981, .2118, .2118, .2118)$. The selected value is .6. HARD DEV error is .0733.

Among 200 Remembering questions, 196 have nonempty source `forgotten_items`. Exactly 180 contain value-bearing forgotten items; 16 contain session IDs without value strings. The supplementary evidence object enumerates all 196 IDs and source session IDs. The executed adapter cannot invent missing target strings, so those 16 are marked `SESSION_ONLY_NO_VALUE`. Four additional questions have empty forgotten-item lists.

D Cluster Robustness

Across the 28 HELDOUT dependency clusters, cluster-mean paired SOFT-minus-HARD error differences range from .0667 to .4000. Leave-one-cluster-out aggregate differences range from .1494 to .1613. At unit level HARD is better on 95 units and tied on 90; SOFT is better on none under the deterministic evidence-value scorer. These diagnostics were computed after the locked primary replacement result and are descriptive robustness checks.