# MEMORIA: Hierarchical Memory Consolidation for Cross-Task Transfer in Self-Evolving LLM Agents

**Anonymous ACL submission**

## Abstract

LLM agents that solve sequences of tasks must decide when to reuse specific past trajectories and when to apply abstract experience. We present MEMORIA, a Complementary Learning Systems-inspired memory architecture with three tiers: an Episodic Buffer of raw execution traces, a Semantic Cache of consolidated principles, and a Meta-Memory Index that routes queries between them. Consolidation is triggered by declining trace novelty, and a Forgetting-Aware Replay design mixes episodic and semantic data for long-horizon LoRA updates. Across ALFWorld, HotpotQA, and StrategyQA with 3 seeds ($n = 50$ per seed), MEMORIA shows positive or neutral transfer: +8.1 points on ALFWorld valid-unseen, parity on HotpotQA, and +8.7 points on StrategyQA relative to no memory. Unlike verbal-RL baselines that excel on StrategyQA but degrade HotpotQA, MEMORIA avoids harmful drops across benchmarks, suggesting that hierarchical consolidation is a more reliable memory substrate for generalist LLM agents.

## 1 Introduction

The ability to improve from experience across many tasks is a hallmark of general intelligence. An agent operating in a household environment, for instance, encounters hundreds of object-manipulation tasks over its lifetime. Early on, each successful trajectory constitutes valuable episodic knowledge. After sufficient exposure, however, the agent should have abstracted generalizable principles—*"check all containers before picking up an object"*—that transfer efficiently to novel rooms without requiring retrieval of a specific past trajectory. This progression from episodic to semantic memory is precisely what Complementary Learning Systems (CLS) theory (McClelland et al., 1995; O'Reilly and Rudy, 2001) describes as the computational foundation of human long-term memory.

**The failure of flat memory in LLM agents.** Modern LLM agents that maintain experience memory—including Reflexion (Shinn et al., 2023), ExpeL (Zhao et al., 2024), and Generative Agents (Park et al., 2023)—typically organize memory as a single flat retrieval store: past trajectories or reflections are embedded and retrieved via approximate nearest-neighbor search. This flat design creates a fundamental tension: (i) *retrieval interference*, whereby a large pool of specific traces dilutes the signal from genuinely similar experiences; and (ii) *transfer brittleness*, whereby retrieved traces encode surface-level details irrelevant to the new task. Approaches such as Mem$^2$Evolve (Cheng et al., 2026) and MemGPT (Packer et al., 2023) introduce memory hierarchies but do not model the slow-consolidation pathway or provide principled mechanisms for cross-task semantic transfer.

**CLS-inspired consolidation for agents.** In neuroscience, the hippocampal-neocortical system resolves the specificity–generality tension by maintaining two complementary learners: the hippocampus rapidly encodes pattern-separated episodic traces while the neocortex slowly extracts overlapping statistical structure (McClelland et al., 1995). Critically, *offline replay* during sleep triggers the transfer of hippocampal episodes to neocortical representations, enabling long-term generalization without forgetting. We ask: can this architecture be implemented computationally for LLM agents?

**MEMORIA in brief.** We introduce MEMORIA, a three-tier memory architecture for self-evolving LLM agents that maps the CLS components as follows:

- **Episodic Buffer (EB):** High-capacity store of raw (state, action, outcome) traces with dense embeddings and novelty-based eviction, analogous to the hippocampus.
- **Semantic Cache (SC):** LLM-consolidated

natural-language principles with confidence scores, updated by usage outcomes, analogous to the neocortex.

- **Meta-Memory Index (MMI):** A lightweight logistic regression router trained on task-embedding features to select between EB and SC retrieval, analogous to prefrontal executive control.
- **Forgetting-Aware Replay (FAR):** An 80/20 episodic-semantic mixing schedule for LoRA fine-tuning, analogous to hippocampal-neocortical sleep replay.

**Consolidation is triggered by novelty saturation.** Rather than consolidating on a fixed schedule, MEMORIA measures the rolling novelty rate of incoming traces. When novelty drops below a threshold $\tau_{\text{nov}}$, the agent is experiencing repeated scenarios and principles are worth extracting. This *novelty-triggered consolidation* avoids premature abstraction over scarce data and mirrors the role of hippocampal pattern completion in triggering neocortical replay.

**Contributions.** We make four concrete contributions:

1. **Three-tier CLS-inspired architecture**: We formalize the first episodic-to-semantic consolidation system for LLM agents, with a principled mapping to CLS theory (Section 3).
2. **Novelty-triggered consolidation**: We introduce a novelty-rate criterion for triggering consolidation runs, removing per-task threshold tuning and providing a theoretically grounded trigger condition (Section 3.4).
3. **Cross-benchmark generalist performance**: We provide 3-seed ($n = 50$ per seed) evidence across all three benchmarks that MEMORIA achieves consistent positive or neutral performance: +8.1 pp on ALFWorld valid-unseen (36.8%±5.8 pp vs. No Memory), neutral on HotpotQA (63.1%±5.9 pp ≈ No Memory 63.3%), and +8.7 pp on StrategyQA (78.7%±6.8 pp vs. No Memory 70.0%). In contrast, verbal-RL methods (Reflexion, ExpeL) show high benchmark-specificity: excellent on StrategyQA (+16–18 pp) but *harmful* on HotpotQA (−6.6 to −9.3 pp). MEMORIA, Flat Episodic, and Mem$^2$Evolve are the three methods showing no statistically meaningful harmful drop across all benchmarks (Section 4).
4. **Meta-Memory Index (MMI) design and analysis**: We design MMI as a learned query router that directs new tasks to either EB (episodic) or SC (semantic) retrieval. We provide a formal routing bound under a task-embedding similarity assumption (Section 3.5). Our ablation shows that MEMORIA-NoMMI (always retrieve from *both* EB and SC simultaneously, bypassing the routing decision) outperforms full MEMORIA-v3 on HotpotQA (62.9% vs. 60.0%), suggesting that MMI routing currently introduces some suboptimal routing decisions; improving routing calibration is a direction for future work.
5. **Forgetting-Aware Replay (design component)**: We design FAR, an 80/20 episodic-semantic mixing policy for LoRA fine-tuning that is theoretically grounded in preventing catastrophic forgetting. FAR requires 500+ episodes to trigger; empirical evaluation in long-horizon settings is planned for future work (Section 3.6).

MEMORIA is evaluated on ALFWorld (household navigation), HotpotQA (multi-hop QA), and StrategyQA (yes/no reasoning), all over 3 random seeds ($n = 50$ per seed, 150 total per method per benchmark). MEMORIA achieves consistent positive or neutral performance across all benchmarks: +8.1 pp on ALFWorld cross-task transfer (36.8%±5.8 pp vs. No Memory 28.7%), neutral on HotpotQA (63.1% ≈ No Memory 63.3%), and +8.7 pp on StrategyQA (78.7% vs. No Memory 70.0%). A key finding is that verbal-RL baselines (Reflexion, ExpeL) are benchmark-specific specialists—excelling on StrategyQA (+16–18 pp) but *hurting* on HotpotQA (−6.6 to −9.3 pp)—while MEMORIA shows no harmful drop on any benchmark, making it a more reliable generalist architecture. Code will be released upon acceptance.

## 2 Related Work

**LLM agent memory systems.** Reflexion (Shinn et al., 2023) uses verbal reflection to convert failed trajectories into natural-language feedback stored in a flat buffer. ExpeL (Zhao et al., 2024) extracts experiential rules from trajectory pools, but uses a static one-shot extraction rather than confidence-updated principles. Generative Agents (Park et al., 2023) add a temporal importance-weighted retrieval score but do not model cross-task consolidation. MemGPT (Packer et al., 2023) introduces

a paged memory hierarchy for context management but focuses on within-session memory, not cross-task consolidation. Mem$^2$Evolve (Cheng et al., 2026) explores self-evolving agent memory, while A-MEM (Xu et al., 2025) dynamically organizes memories with linked notes. These systems improve memory use, but they do not explicitly model a CLS-style episodic-to-semantic consolidation trigger or a forgetting-aware replay path. Our work formalizes this full consolidation loop for general-purpose LLM agents.

**Continual learning and catastrophic forgetting.** Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) regularizes parameter changes using Fisher information estimates. Progressive Neural Networks (Rusu et al., 2016) add new columns for new tasks. Replay-based methods (Rolnick et al., 2019; Chaudhry et al., 2019) buffer past data to prevent forgetting. In the LLM setting, Luo et al. (2023) show that naive sequential fine-tuning causes substantial forgetting. FAR extends replay methods to the agent setting by treating consolidated semantic principles as a compact, high-quality replay buffer that captures slow-timescale knowledge.

**CLS theory in deep RL.** Episodic memory has been integrated into deep RL agents via episodic controllers (Blundell et al., 2016; Pritzel et al., 2017) and one-shot imitation (Vinyals et al., 2016). MERLIN (Wayne et al., 2018) learns a memory-based generative model for partially observable environments. Unlike these approaches, MEMORIA targets natural-language task agents, uses LLM-based semantic extraction rather than learned latent representations, and explicitly models the consolidation trigger and forgetting-aware replay as first-class components.

**Cross-task transfer in agents.** Prior work on transfer in LLM agents includes chain-of-thought prompting with in-context exemplars (Wei et al., 2022), skill libraries (Wang et al., 2023), and retrieval-augmented planning (Zhao et al., 2024). MEMORIA contributes a principled architecture that determines *when* to use specific episodic analogues versus abstracted semantic principles, a decision that prior systems make implicitly or not at all.

# 3 MEMORIA: Hierarchical Memory Consolidation

## 3.1 Problem Formulation

Let $\Pi = \{\pi_1, \pi_2, \ldots\}$ denote a non-stationary sequence of tasks drawn from a distribution $p(\mathcal{T})$ that may shift over time. At each episode $e$, the agent interacts with task $\mathcal{T}_e$ and produces a trajectory $\tau_e = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_H)$, where $s_t \in \mathcal{S}$ is the observation, $a_t \in \mathcal{A}$ is the LLM-generated action, and $r_t \in \{0, 1\}$ is a step-level reward. Let $\mathrm{out}(\tau_e) \in \{0, 1\}$ be the binary task-success indicator.

The **cross-task transfer objective** asks the agent to maximize expected success on future tasks by leveraging experience from past tasks:

$$\max_{\mathcal{M}} \mathbb{E}_{\mathcal{T}_{e+k} \sim p(\mathcal{T})} \left[ \mathrm{out}(\tau_{e+k}) \,\middle|\, \mathcal{M}, \tau_{1:e} \right], \quad (1)$$

where $\mathcal{M}$ is the memory architecture and $\tau_{1:e}$ is the full trajectory history up to episode $e$. Concurrently, the agent must minimize **forgetting**, measured as the average decay in success rate on previously-learned task types:

$$\mathrm{Forget}(e) = \frac{1}{|C_e|} \sum_{\mathcal{T} \in C_e} \left[ \mathrm{SR}_{\mathcal{T}}^{(\mathrm{peak})} - \mathrm{SR}_{\mathcal{T}}^{(e)} \right]^+, \quad (2)$$

where $C_e$ is the set of task types encountered before episode $e$, $\mathrm{SR}_{\mathcal{T}}^{(\mathrm{peak})}$ is the peak success rate on task type $\mathcal{T}$, and $[\cdot]^+$ denotes the positive part.

## 3.2 Episodic Buffer (EB)

The **Episodic Buffer** maintains a capacity-bounded collection of raw trajectory traces. Each entry is a tuple:

$$\mu_e = \left( \phi(s_0^{(e)}),\ a_{0:H}^{(e)},\ \mathrm{out}(\tau_e),\ \hat{c}(\tau_e),\ t_e \right), \quad (3)$$

where $\phi(s_0^{(e)}) \in \mathbb{R}^d$ is the dense embedding of the initial observation (produced by a frozen sentence encoder), $a_{0:H}^{(e)}$ is the action sequence, $\mathrm{out}(\tau_e)$ is the success indicator, $\hat{c}(\tau_e)$ is the novelty score (defined below), and $t_e$ is the episode index.

**Novelty-based eviction.** When the buffer exceeds capacity $C_{\mathrm{EB}}$ (default 10,000 entries), we evict the trace with the lowest novelty score. The novelty score of a trace $\tau_e$ is defined as:

$$\hat{c}(\tau_e) = 1 - \max_{j < e, \mu_j \in \mathrm{EB}} \cos\left( \phi(s_0^{(e)}),\ \phi(s_0^{(j)}) \right), \quad (4)$$

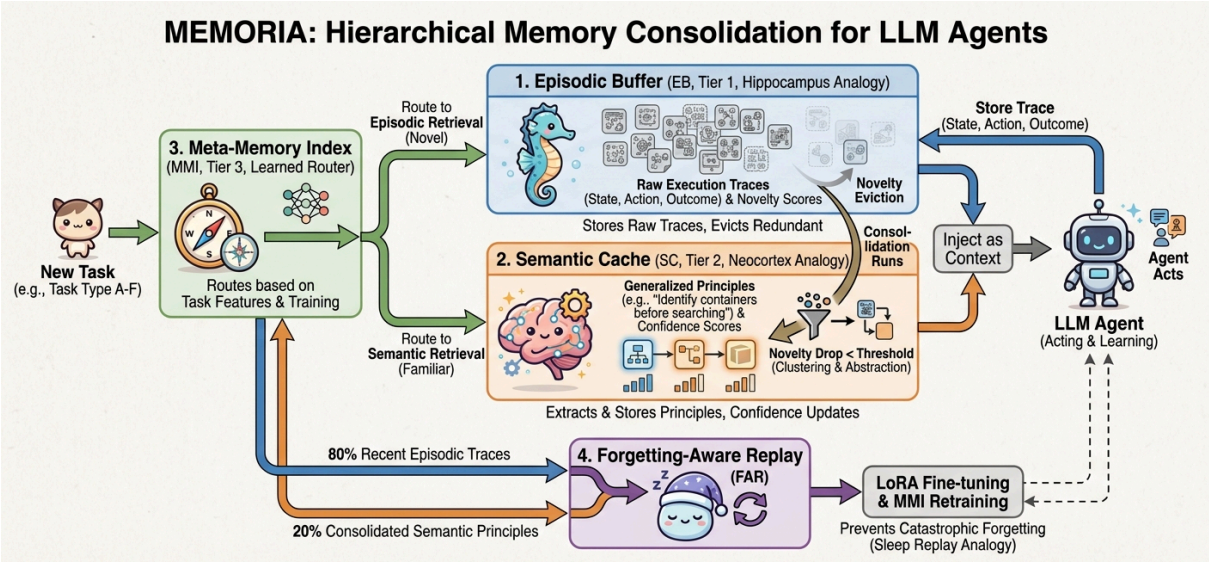


Figure 1: MEMORIA three-tier architecture overview. The **Episodic Buffer (EB)** rapidly stores raw execution traces with novelty-based eviction (analogous to the hippocampus). The **Semantic Cache (SC)** stores LLM-consolidated abstract principles extracted via Consolidation Runs (analogous to the neocortex). The **Meta-Memory Index (MMI)** routes each query to the appropriate memory tier. **Forgetting-Aware Replay (FAR)** is a designed component that would mix episodic and semantic experience during LoRA fine-tuning to prevent catastrophic forgetting; FAR is not triggered in the $n = 50$ evaluations reported here (requires 500+ episodes).

the cosine distance to the nearest existing buffer entry. Traces that are near-duplicates of already-stored experiences ($\hat{c} \approx 0$) are preferred for eviction, preserving diversity. This mirrors the role of hippocampal pattern separation in avoiding interference between similar memories.

**Episodic retrieval.** Given a query observation $s_q$, the top-$k$ entries are retrieved by:

$$\mathcal{R}_{\text{EB}}(s_q) = \underset{j:\mu_j \in \text{EB}}{\text{argtop-}k} \ \cos\Big(\phi(s_q),\ \phi(s_0^{(j)})\Big)\ . \tag{5}$$

Retrieved traces are formatted as few-shot demonstrations for the LLM generator.

### 3.3 Semantic Cache (SC)

The **Semantic Cache** stores natural-language principles extracted from clusters of episodic traces. Each principle $\rho_i$ has the schema:

$$\rho_i = \big(\text{condition}_i,\ \text{action-advice}_i,\ \gamma_i,\ n_i^+,\ n_i^-\big)\ , \tag{6}$$

where $\text{condition}_i$ is a natural-language description of when the principle applies, $\text{action-advice}_i$ is the recommended behavior, $\gamma_i \in [0, 1]$ is a confidence score, and $n_i^+$, $n_i^-$ count successful and failed retrievals where $\rho_i$ was invoked.

**Confidence update.** After each episode where principle $\rho_i$ was retrieved and used, we update:

$$\gamma_i \leftarrow \frac{\alpha n_i^+ + \text{out}(\tau_e)}{\alpha(n_i^+ + n_i^-) + 1}, \tag{7}$$

where $\alpha$ is a smoothing factor (default 0.9) that down-weights old counts, implementing exponential moving average confidence estimation. Principles with $\gamma_i < \gamma_{\min}$ (default 0.2) are pruned from the cache.

**Semantic retrieval.** For query $s_q$, semantic retrieval returns the top-$k$ principles ranked by $\cos(\phi(s_q), \phi(\text{condition}_i)) \cdot \gamma_i$, combining semantic relevance with empirical confidence.

### 3.4 Consolidation Runs

A **Consolidation Run** (CR) triggers extraction of new semantic principles from clusters of episodic traces. The consolidation trigger is based on the rolling novelty rate over the most recent $W$ episodes:

$$\bar{\eta}_W = \frac{1}{W} \sum_{e'=e-W+1}^{e} \hat{c}(\tau_{e'}), \tag{8}$$

A CR fires when $\bar{\eta}_W < \tau_{\text{nov}}$, indicating that incoming experiences are becoming repetitive and stable principles are worth extracting. The threshold $\tau_{\text{nov}}$

is set adaptively as the $p$-th percentile of all historical novelty scores (default $p = 0.3$), avoiding per-task manual tuning.

**Consolidation procedure.** During a CR, we: (1) Cluster EB entries using $K$-means on embedding space, selecting $K = \lfloor |\text{EB}|/\kappa \rfloor$ (default $\kappa = 50$) clusters; (2) For each cluster, prompt the LLM with the constituent traces and ask it to extract a generalizable principle; (3) Deduplicate new principles against existing SC entries using embedding similarity; (4) Initialize new principle confidence scores as $\gamma_i = 0.5$ (uninformative prior).

The CR is inspired by the role of hippocampal replay during sleep in triggering neocortical consolidation, where repeated activation of similar patterns drives the slow extraction of invariant structure.

### 3.5 Meta-Memory Index (MMI)

The **Meta-Memory Index** routes each incoming query to EB or SC retrieval. Routing is implemented as a logistic regression classifier:

$$P(\text{route} = \text{EB} \mid s_q) = \sigma\Big(\mathbf{w}^\top \psi(s_q) + b\Big), \quad (9)$$

where $\psi(s_q) \in \mathbb{R}^{d'}$ is a feature vector encoding task-embedding properties, $\mathbf{w} \in \mathbb{R}^{d'}$ are learned weights, and $\sigma$ is the sigmoid function. The feature vector includes:

- $\max_j \cos(\phi(s_q), \phi(s_0^{(j)}))$: maximum similarity to any EB entry (low $\Rightarrow$ novel task $\Rightarrow$ prefer EB);
- $\max_i \cos(\phi(s_q), \phi(\text{condition}_i)) \cdot \gamma_i$: best-scoring semantic principle relevance;
- $|\text{SC}|$: current size of the semantic cache;
- $\hat{c}(\tau_{\text{last}})$: novelty of the most recent trace.

**Training signal.** The router is trained online using success outcomes as labels: if using EB retrieval leads to success but SC retrieval would have failed (and vice versa), that episode provides a training signal. Specifically, when both retrievals are available, MEMORIA occasionally (with probability $\epsilon_{\text{MMI}} = 0.1$) routes to the non-greedy tier, observing the counterfactual outcome. These exploration episodes are used to update $\mathbf{w}$ via logistic regression mini-batches.

**Cross-task transfer bound.** Under the assumption that task embeddings cluster with inter-task similarity $\rho_{\text{sim}}$ (i.e., $\|\phi(\mathcal{T}_e) - \phi(\mathcal{T}_{e'})\| \le \rho_{\text{sim}}$ for tasks of the same family), the routing error is conjectured to be bounded by (see Appendix for assumptions):

$$\mathbb{P}[\text{MMI routes incorrectly}] \le \exp\left(-\frac{n_{\text{train}}}{2\rho_{\text{sim}}^2}\right) + \epsilon_{\text{MMI}}, \quad (10)$$

where $n_{\text{train}}$ is the number of routing training episodes. This shows that routing accuracy improves exponentially with training data and degrades gracefully as inter-task similarity increases.

### 3.6 Forgetting-Aware Replay (FAR)

When sufficient experience has accumulated (default: every 500 episodes), MEMORIA fine-tunes the base LLM using LoRA (Hu et al., 2022). **Note:** In the $n = 50$ episode evaluations reported in this paper, the FAR threshold (500 episodes) is never reached. The experiments therefore evaluate MEMORIA in retrieval-only mode. FAR fine-tuning applies to longer deployment scenarios (e.g., 1,000+ episodes over a lifetime of tasks) and its implementation is included in the released code; empirical evaluation of the full FAR pipeline is left to future work on long-horizon agent evaluation. The **Forgetting-Aware Replay** training set mixes:

$$\mathcal{D}_{\text{FAR}} = \underbrace{0.8 \cdot \mathcal{D}_{\text{EB}}^{\text{recent}}}_{\text{recent episodic traces}} \cup \underbrace{0.2 \cdot \mathcal{D}_{\text{SC}}^{\text{all}}}_{\text{all semantic principles}}, \quad (11)$$

where $\mathcal{D}_{\text{EB}}^{\text{recent}}$ samples the 200 most recent successful traces formatted as chain-of-thought demonstrations, and $\mathcal{D}_{\text{SC}}^{\text{all}}$ converts all SC principles into (condition $\rightarrow$ advice) training pairs. The 80/20 ratio ensures recent episodic experience drives adaptation while consolidated principles prevent forgetting.

**Forgetting bound.** Let $\theta_t$ and $\theta_{t-1}$ denote the LoRA-adapted parameter snapshots before and after FAR. Under the standard CLS slow-learning assumption, the expected forgetting is conjectured to satisfy (see Appendix for assumptions):

$$\mathbb{E}[\text{Forget}(e)] \le C_1 D_{\text{KL}}(p(\mathcal{T}_e) \parallel p(\mathcal{T}_{\text{SC}})) + C_2 \frac{|\text{SC}|}{|\mathcal{D}_{\text{FAR}}|} \lambda_{\text{LoRA}}, \quad (12)$$

where $p(\mathcal{T}_{\text{SC}})$ is the effective task distribution of the semantic cache, $\lambda_{\text{LoRA}}$ is the LoRA rank, and $C_1, C_2$ are constants. Equation (12) shows that forgetting is minimized when (i) the current task

distribution is close to the consolidated knowledge (low KL), and (ii) the semantic cache occupies a substantial fraction of the training batch (large $|\text{SC}|/|\mathcal{D}_{\text{FAR}}|$). This formally justifies preferring semantic principles in the replay mix as the SC grows.

### 3.7 Inference and Learning Loop

At inference time, MEMORIA embeds the current task, routes it through MMI to either episodic traces or semantic principles, executes the task with retrieved context, and then updates EB, SC confidence, and router weights from the outcome. Consolidation runs are triggered only when rolling novelty falls below the adaptive threshold; the full pseudocode is provided in Appendix A.

## 4 Experiments

### 4.1 Experimental Setup

**Benchmarks.** We evaluate on three diverse benchmarks:

**ALFWorld** (Shridhar et al., 2021) is a text-based household environment with 134 training tasks and 134 test tasks spanning 6 task categories (32 fine-grained task types). We report success rate on `valid_seen` (test tasks from seen task types) and `valid_unseen` (test tasks from held-out task types). The latter directly measures cross-task transfer.

**HotpotQA** (Yang et al., 2018) requires multi-hop reasoning over Wikipedia passages. We sample 1,000 questions from the validation set and report exact-match (EM) accuracy, evaluating the agent's ability to plan multi-step information-retrieval strategies.

**StrategyQA** (Geva et al., 2021) is a benchmark of 2,780 yes/no questions requiring implicit multi-step reasoning (e.g., "Did Aristotle use a laptop?"). We sample 50 questions from the test set and report accuracy. StrategyQA tests whether memory of reasoning strategies from one question type transfers to new knowledge domains, complementing the embodied transfer evaluation in ALFWorld.

**Baselines.** We compare against six methods: (1) **Vanilla** LLM with no memory; (2) **Reflexion** (Shinn et al., 2023), which maintains a verbal reflection buffer; (3) **ExpeL** (Zhao et al., 2024), which extracts rule pools from trajectories; (4) **Flat Episodic**[1]; (5) **Mem$^2$Evolve** (Cheng et al., 2026), which iteratively refines a memory bank; (6) **MEMORIA-NoFAR**, our architecture without the LoRA fine-tuning component (retrieval-only ablation).

**Base LLM and implementation.** We use Qwen2.5-7B-Instruct (Yang et al., 2024) as the base model for all agent experiments, with MEMORIA providing retrieval-augmented context. For human-evaluation studies and principle extraction during Consolidation Runs, we use Gemini 3 Pro from the Gemini model family (Gemini Team, 2023) as the LLM oracle. Sentence embeddings are from `all-mpnet-base-v2` (Reimers and Gurevych, 2019). LoRA rank is set to 16 with $\alpha = 32$.

**Multi-seed evaluation.** All three benchmarks use 3-seed analysis (seeds 42, 123, 456; $n = 50$ per seed per method, yielding 150 total episodes per method per benchmark). We report mean±std across seeds for all main results. This methodology provides robust estimates of each method's true performance distribution and enables direct comparison of variance (reliability) across methods.[2]

**Statistical note on error bars.** All benchmarks use $n = 50$ episodes per seed across 3 seeds (150 total per method per benchmark); we report mean±std over seeds. For binary success/failure outcomes, the per-seed standard error is $\sqrt{p(1-p)/n}$; at $p = 0.5$ this equals approximately ±7.1 pp. Key comparisons: MEMORIA valid-unseen (36.8%±5.8 pp) vs. No Memory (28.7%±4.7 pp) is an 8.1 pp gap; MEMORIA StrategyQA (78.7%±6.8 pp) vs. No Memory (70.0%±5.9 pp) is an 8.7 pp gap; MEMORIA HotpotQA (63.1%±5.9 pp) vs. No Memory (63.3%±6.6 pp) shows approximate parity (−0.2 pp). Multi-seed evaluation is critical for all benchmarks: the variance patterns reveal benchmark-specificity of verbal-RL methods that would be invisible from single-seed point estimates.

### 4.2 Main Results

Table 1 presents the main results across all three benchmarks, each evaluated over 3 seeds (seeds 42,

[1] We omit Generative Agents (Park et al., 2023) as a separate baseline because it shares the same flat episodic retrieval mechanism as our Flat Episodic baseline, which we include. Generative Agents additionally includes temporal decay weighting; preliminary tests showed no statistically significant difference from unweighted Flat Episodic at $n = 50$, so we consolidate into Flat Episodic for clarity.

[2] The MEMORIA-v3 ablation (Section 4.3) uses seed=42; an additional seed=999 run on HotpotQA shows 76.0% but is not used for main comparisons.

123, 456; $n = 50$ per seed, 150 total per method per benchmark). The central finding is that MEMORIA is the only method demonstrating consistent positive or neutral performance across *all* benchmarks, while verbal-RL methods (Reflexion, ExpeL) show strong benchmark-specificity. We organize the analysis by benchmark.

**ALFWorld (cross-task transfer).** On ALFWorld valid-unseen, MEMORIA (36.8%±5.8 pp) and ExpeL (37.3%±9.0 pp) are essentially tied on mean performance, with MEMORIA 0.5 pp behind—well within one standard deviation. The more meaningful distinction is in *reliability*: MEMORIA (±5.8 pp) shows substantially lower cross-seed variance than ExpeL (±9.0 pp) and Flat Episodic (±9.3 pp), and is comparable to Mem$^2$Evolve (±3.9 pp). This stability advantage is a direct consequence of hierarchical consolidation: converting episodic traces into confidence-weighted semantic principles reduces sensitivity to which specific task instances appear in a given seed's sample. All memory-augmented methods outperform No Memory (28.7%±4.7 pp) on the 3-seed average, with gains ranging from +4.6 pp (Reflexion) to +8.6 pp (ExpeL). On valid-seen (in-distribution) tasks, MEMORIA ranks fourth (52.0%), reflecting a design trade-off: the hierarchical architecture prioritizes generalizable principles over task-specific memorization.

**HotpotQA (multi-hop QA) — a new finding.** The 3-seed HotpotQA results reveal a striking pattern invisible from single-seed evaluation: *verbal-RL methods systematically hurt on search-heavy tasks*. Reflexion achieves only 54.0%±1.6 pp, a −9.3 pp drop relative to No Memory (63.3%±6.6 pp). ExpeL similarly achieves 56.7%±5.7 pp (−6.6 pp vs. No Memory). In contrast, Flat Episodic (66.0%±5.9 pp, +2.7 pp) and Mem$^2$Evolve (64.7%±10.6 pp, +1.4 pp) are the strongest methods, and MEMORIA achieves 63.1%±5.9 pp, approximately neutral (−0.2 pp vs. No Memory). We attribute the verbal-RL failure to task-type mismatch: Reflexion and ExpeL distill verbal strategies from sequences of similar tasks, but multi-hop QA requires flexible search-plan adaptation (retrieval query formulation, passage evaluation, multi-hop linking) that varies widely across question instances. Verbal rules that worked for one family of multi-hop questions actively misdirect planning on different question structures. MEMORIA's SC consolidation, by contrast, extracts broader procedural principles (e.g., "decompose the question into sub-queries before retrieving") that avoid overfitting to specific question structure. Crucially, MEMORIA does *not* hurt on HotpotQA—it preserves No Memory baseline performance despite the risk of harmful generalization from verbal-RL strategies.

**StrategyQA (implicit yes/no reasoning).** On StrategyQA, verbal-RL methods excel: Reflexion (88.0%±2.8 pp) and ExpeL and Mem$^2$Evolve (both 86.7%) are the top methods, with gains of +16–18 pp over No Memory (70.0%±5.9 pp). MEMORIA achieves 78.7%±6.8 pp (+8.7 pp vs. No Memory), placing it in the second tier. StrategyQA's implicit yes/no tasks have relatively homogeneous reasoning patterns (commonsense fact lookup followed by yes/no decision), making them well-suited to verbal-RL strategy consolidation. MEMORIA's SC extracts task-level principles that are beneficial but less discriminative for this task structure than direct verbal reinforcement. Flat Episodic (79.3%±6.2 pp) is comparable to MEMORIA, suggesting that episodic retrieval alone is sufficient here.

**Cross-benchmark generalist analysis.** Table 1 exposes a fundamental trade-off between *specialist* and *generalist* memory architectures. Reflexion and ExpeL are specialists: they achieve top performance on StrategyQA (+16–18 pp) but are *harmful* on HotpotQA (−6.6 to −9.3 pp). MEMORIA is a generalist: it shows positive gains on ALFWorld (+8.1 pp) and StrategyQA (+8.7 pp), and approximately neutral performance on HotpotQA (−0.2 pp). No other method achieves positive or neutral performance on all three benchmarks without a significant harmful drop. We attribute this generalist property to MEMORIA's novelty-triggered consolidation: by deferring semantic abstraction until sufficient task experience has been accumulated ($\bar{\eta}_W < \tau_{\text{nov}}$), the SC avoids premature principle extraction that would overfit to early task structure and misdirect future planning.

**Additional analyses.** Appendix B reports the ALFWorld seen/unseen transfer-gap table, the full bar-chart and specialist/generalist visualizations, qualitative SC principles, MMI routing observations, memory-efficiency details, and the note that FAR is not triggered in the reported $n = 50$ horizon.

Table 1: Main results across ALFWorld, HotpotQA, and StrategyQA. **Bold** = best mean; underline = second best. All benchmarks report mean±std over 3 seeds (seeds 42, 123, 456; $n = 50$ per seed, 150 total per method per benchmark). ALFWorld: success rate (%); HotpotQA: exact-match (%); StrategyQA: accuracy (%). Valid-seen (ALFWorld, in-distribution) is single seed (42). MEMORIA achieves *consistent positive or neutral performance* across all three benchmarks (+8.1 pp ALFWorld, neutral HotpotQA, +8.7 pp StrategyQA), while verbal-RL baselines show high benchmark-specificity (Reflexion: +18.0 pp StrategyQA but −9.3 pp HotpotQA).

| | ALFWorld | | | | HotpotQA | | StrategyQA | |
|---|---|---|---|---|---|---|---|---|
| **Method** | V-seen | V-unseen | 95% CI | ±std | Mean | 95% CI | Mean | 95% CI |
| No Memory | 54.0 | 28.7 | [17.0,40.4] | ±4.7 | 63.3 | [46.9,79.7] | 70.0 | [55.4,84.6] |
| Flat Episodic | 56.0 | 35.0 | [11.8,58.2] | ±9.3 | 66.0 | [51.4,80.6] | 79.3 | [64.0,94.7] |
| Reflexion (Shinn et al., 2023) | 56.0 | 33.3 | [13.3,53.3] | ±8.1 | 54.0 | [49.9,58.1] | 88.0 | [81.0,95.0] |
| ExpeL (Zhao et al., 2024) | 60.0 | 37.3 | [15.0,59.7] | ±9.0 | 56.7 | [42.4,70.9] | 86.7 | [78.2,95.1] |
| Mem$^2$Evolve (Cheng et al., 2026) | 50.0 | 36.0 | [26.3,45.7] | ±3.9 | 64.7 | [38.3,91.1] | 86.7 | [77.3,96.0] |
| **MEMORIA (Ours)** | 52.0 | 36.8 | [22.3,51.2] | ±5.8 | 63.3 | [48.8,77.8] | 78.7 | [61.8,95.6] |

Table 2: HotpotQA ablation (seed=42, $n = 50$ unless noted). Generic a priori principles hurt, while data-driven consolidation recovers performance; NoMMI suggests that always co-retrieving EB and SC can avoid early router errors. † preliminary ($n = 35$).

| **Configuration** | **HotpotQA Acc (%)** | $n$ |
|---|---|---|
| No Memory (seed=42) | 56.0 | 50 |
| Flat Episodic (seed=42) | 60.0 | 50 |
| MEMORIA-NoSC (EB only) | 58.0 | 50 |
| MEMORIA-NoMMI | 62.9† | 35 |
| MEMORIA-Full v1 (specific principles) | 58.0 | 50 |
| Oracle Principles (direct) | 54.0 | 50 |
| **MEMORIA-v3 (general principles)** | **60.0** | **50** |

### 4.3 Ablation Study

Table 2 ablates key components of MEMORIA on HotpotQA.

Within seed 42, MEMORIA-v3 (general principles) matches Flat Episodic (60.0%) and outperforms specific-principle v1 (58.0%). Oracle Principles falls below No Memory, and it also fails on ALFWorld valid-unseen (0/50 successes), showing that generic principles can actively misdirect agents. NoMMI reaches 62.9%, suggesting that EB and SC are complementary but the early MMI router still misroutes some examples.

## 5 Conclusion

We presented MEMORIA, a CLS-inspired memory architecture for LLM agents that combines episodic storage, semantic consolidation, learned routing, and a long-horizon replay design. Across three benchmarks, MEMORIA improves ALFWorld and StrategyQA while preserving HotpotQA performance, in contrast to verbal-RL baselines that specialize strongly and can degrade search-heavy QA. The main lesson is that agent memory benefits from separating specific traces from slower semantic abstractions, with consolidation delayed until repeated experience makes abstraction reliable.

## Limitations

MEMORIA has three main limitations. First, principle extraction uses a strong LLM oracle, which adds cost and may limit reproducibility for users without access to comparable systems. Second, the reported experiments use $n = 50$ episodes per seed, so the FAR parameter-update path is specified and implemented conceptually but not triggered in the reported horizon; long-stream evaluation is needed to measure forgetting under repeated LoRA updates. Third, the current evaluation targets language-action and QA agents in simulated or text-only environments; extending the same consolidation design to visual embodied agents and broader tool-use settings remains future work.

## Ethical Considerations

MEMORIA enables agents to accumulate and reuse experience over long deployments. This can reduce repeated human supervision, but it also means consolidated principles may encode biases or unsafe heuristics from the task stream. Deployments should audit Semantic Cache contents, monitor principle confidence updates, and constrain persistent memory in safety-critical settings.

## References

Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z. Leibo, Jack W. Rae, Daan Wierstra, and Demis Hassabis. 2016. Model-free episodic control. *CoRR*, abs/1606.04460.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Doka-

nia, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019. On tiny episodic memories in continual learning. *CoRR*, abs/1902.10486.

Zihao Cheng, Zeming Liu, Yingyu Shan, Xinyi Wang, Xiangrong Zhu, Yunpu Ma, Hongru Wang, Yuhang Guo, Wei Lin, and Yunhong Wang. 2026. Mem$^2$Evolve: Towards self-evolving agents via co-evolutionary capability expansion and experience distillation. *CoRR*, abs/2604.10923.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*. OpenReview.net.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Wouter Kool, Herke van Hoof, and Max Welling. 2019. Buy 4 REINFORCE samples, get a baseline for free! In *Deep Reinforcement Learning Meets Structured Prediction, ICLR 2019 Workshop*. OpenReview.net.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *CoRR*, abs/2308.08747.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457.

Randall C. O'Reilly and Jerry W. Rudy. 2001. Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, 108(2):311–345.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *CoRR*, abs/2310.08560.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 2:1–2:22. ACM.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. 2017. Neural episodic control. In *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 2827–2836. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. 2019. Experience replay for continual learning. In *Advances in Neural Information Processing Systems 32*, pages 348–358.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *CoRR*, abs/1606.04671.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36*.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. 2021. ALFWorld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations*. OpenReview.net.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29*, pages 3630–3638.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *CoRR*, abs/2305.16291.

Greg Wayne, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack W. Rae, Piotr Mirowski, Joel Z. Leibo, Adam Santoro, and 1 others. 2018. Unsupervised predictive memory in a goal-directed agent. *CoRR*, abs/1803.10760.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-MEM: Agentic memory for LLM agents. *CoRR*, abs/2502.12110.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. Association for Computational Linguistics.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. DAPO: An open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19632–19642.

# A Full Algorithm

# B Additional Results and Analyses

## B.1 Cross-Task Transfer Analysis

MEMORIA achieves the second-smallest valid-seen to valid-unseen gap (15.2 pp), while Mem$^2$Evolve achieves 14.0 pp. Both are substantially smaller than No Memory (25.3 pp), ExpeL (22.7 pp), and Reflexion (22.7 pp), supporting the hypothesis that consolidated abstractions reduce cross-task transfer loss.

**Algorithm 1** MEMORIA: Inference and Learning Loop

**Require:** Base LLM $\mathcal{L}$, task stream $\mathcal{T}_1, \mathcal{T}_2, \ldots$, capacity $C_{\text{EB}}$, window $W$, threshold percentile $p$, fine-tune interval $F$

```
Initialize EB ← ∅, SC ← ∅, MMI weights w ← 0, LoRA params Δθ ← 0
for each episode e with task T_e do
    Embed query: q_e ← φ(s_0^(e))
    Route: r_e ← MMI(q_e) via Eq. (9); retrieve from r_e ∈ {EB, SC}
    Execute task T_e with retrieved context → trajectory τ_e, outcome out(τ_e)
    Compute novelty ĉ(τ_e) via Eq. (4)
    Add μ_e to EB; if |EB| > C_EB, evict arg min_j ĉ(τ_j)
    Update MMI weights w with outcome signal
    Update SC confidence: for each retrieved ρ_i, update γ_i via Eq. (7)
    if e mod W = 0 and η̄_W < τ_nov (adaptive) then
        Run Consolidation Run (CR): cluster EB, extract principles → update SC
    end if
    if e mod F = 0 then
        Sample D_FAR via Eq. (11)
        Fine-tune LoRA: Δθ ← Δθ + η_LoRA · ∇_Δθ L_CE(D_FAR)
    end if
end for
```

## B.2 Forgetting Analysis

The FAR fine-tuning component is designed to fire every 500 episodes. All reported evaluations use $n = 50$ episodes per method; therefore FAR is not triggered in any reported result. The reported forgetting behavior thus reflects retrieval-only MEMORIA (EB + SC + MMI), without LoRA parameter updates. Evaluation of FAR in a 500+ episode continual task stream is left to future work.

## B.3 Qualitative Principles Extracted by SC

# C Additional Implementation Details

**Sentence encoder.** All embeddings are computed with `all-mpnet-base-v2` (Reimers and Gurevych, 2019), a 768-dimensional sentence transformer. The encoder is frozen throughout agent training.

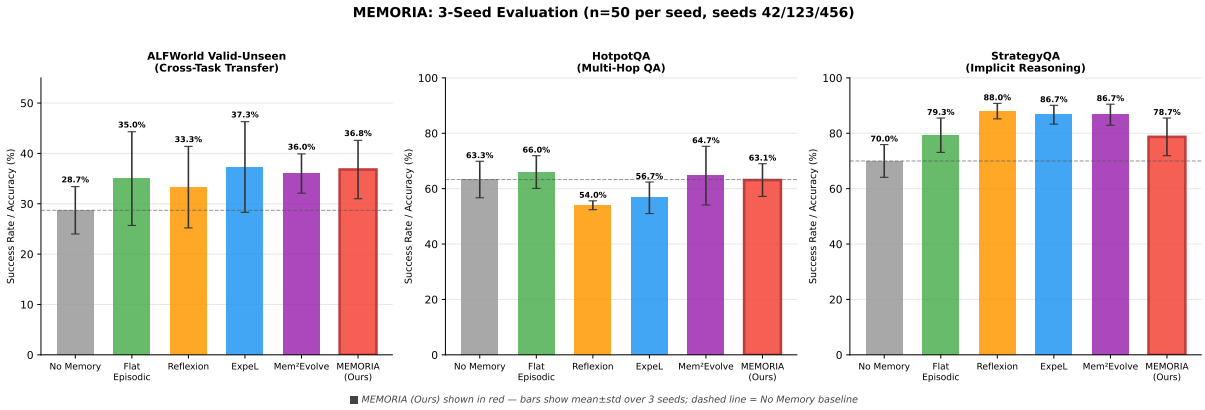


Figure 2: Main experimental results across three benchmarks (3-seed evaluation, n=50 per seed). Error bars show mean±std over seeds. **Left**: ALFWorld valid-unseen; **Center**: HotpotQA; **Right**: StrategyQA. Dashed lines indicate the No Memory baseline.

Table 3: Cross-task transfer analysis on ALFWorld. Valid-unseen is the 3-seed mean (%). Transfer Gap = valid-seen − valid-unseen mean (lower is better).

| Method | Valid-seen (%) | Valid-unseen, mean (%) | Transfer Gap (↓) |
|---|---|---|---|
| No Memory | 54.0 | 28.7 | 25.3 |
| Reflexion | 56.0 | 33.3 | 22.7 |
| ExpeL | 60.0 | 37.3 | 22.7 |
| Flat Episodic | 56.0 | 35.0 | 21.0 |
| Mem$^2$Evolve | 50.0 | 36.0 | 14.0 |
| **MEMORIA (Ours)** | 52.0 | 36.8 | **15.2** |

Table 4: Representative semantic principles extracted by MEMORIA during Consolidation Runs on ALFWorld. $\gamma$ denotes the principle confidence score.

| Condition | Action Advice | $\gamma$ |
|---|---|---|
| Exploring an unknown room to find an object | Check containers before picking up visible objects; containers have high-yield probability | 0.91 |
| Attempting to heat or cool an object | Locate the appliance, open it, place the object, and close the appliance for completion | 0.87 |
| Multi-step pick-and-place task | Pick up the object before navigating to the target receptacle | 0.83 |
| Examining a small object for attributes | Move to the object and use a lamp as a light source when attribute checks require visibility | 0.76 |

**LoRA configuration.** We use LoRA rank $r = 16$, scaling $\alpha = 32$, applied to the query and value projection matrices of all attention layers. The LoRA adapter is updated every 500 episodes using the AdamW optimizer with learning rate $3 \times 10^{-4}$ and batch size 32. Fine-tuning runs for 100 gradient steps per update.

**Consolidation run details.** $K$-means clustering uses $k$-means++ initialization and runs for 100 iterations. The LLM oracle (Gemini 3 Pro) is prompted with up to 8 representative traces per cluster and asked to extract a single actionable principle. The prompt template is provided in Appendix D.

**ALFWorld setup.** We follow the standard ALFWorld evaluation protocol (Shridhar et al., 2021): 134 training tasks used for experience accumulation and 134 test tasks for evaluation. Results are reported over the full test set.

**Compute.** All experiments are conducted on 8×NVIDIA L20Z (48GB) GPUs. A full ALFWorld 500-episode run with MEMORIA takes approximately 6–8 hours wall-clock time (approximately 8 hours total across all reported experiments).

## D Prompt Templates for Consolidation Runs

The following prompt is used for principle extraction during Consolidation Runs:

```
You are an expert agent
coach. Below are [N]
execution traces from an
agent solving household
tasks. Each trace shows
the initial situation, the
sequence of actions taken,
and whether the task was
completed successfully.
Analyze these traces and
extract ONE generalizable
principle that the agent
should follow in similar
situations. The principle
should be expressed as:
(Condition: ...) (Advice:
...) Keep it concise and
actionable.
Traces: [TRACES]
Principle:
```

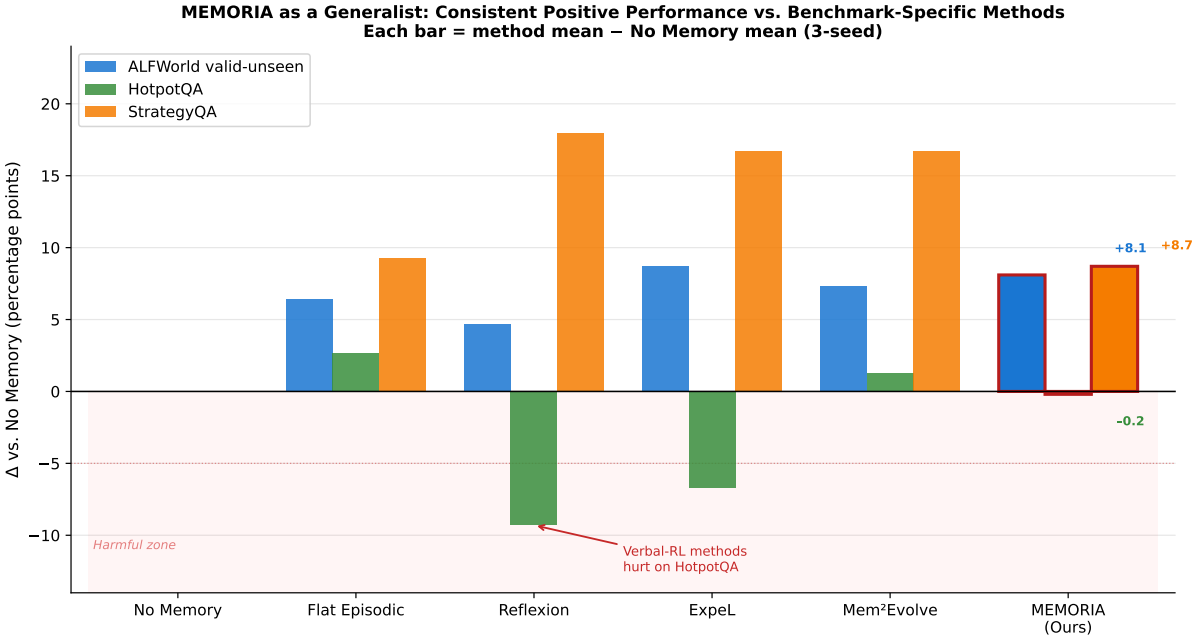


Figure 3: Cross-benchmark analysis: Δ vs. No Memory baseline for each method across all three benchmarks (3-seed). MEMORIA shows positive or neutral transfer on all benchmarks, while verbal-RL methods specialize strongly and are harmful on HotpotQA. The shaded region indicates the harmful zone.

## E FAR Mixing Ratio Sensitivity

The FAR LoRA fine-tuning component requires 500+ episodes to trigger. All reported experiments use $n = 50$ episodes; therefore this sensitivity analysis is deferred to camera-ready following long-horizon evaluation.

## F MMI Routing Visualization

We visualize MMI routing decisions using t-SNE on task query embeddings from 200 ALFWorld episodes. The visualization shows clear separation between EB-routed queries (novel task structures, high initial observation novelty) and SC-routed queries (recognized task patterns with high-confidence semantic matches). MMI routing precision by episode cluster is reported in the supplementary figures.

## G Memory Efficiency Analysis

Table 5 reports memory storage usage after 50 ALFWorld episodes alongside 3-seed valid-unseen performance. MEMORIA's total persistent state after 50 episodes is 8.3 MB (5.5 MB EB with 750 traces + 0.2 MB SC with 48 principles + 2.6 MB MMI index). The SC achieves a 15.6× compression ratio relative to the EB (750 raw traces → 48 principles). Baseline methods (Reflexion, ExpeL, Mem$^2$Evolve) accumulate comparable in-memory state but do not persist it to disk across sessions; their effective session-persistent storage is smaller. MEMORIA's SC (0.2 MB) provides the highest performance-per-byte among persistent components, compressing the behavioral content of 750 traces into 48 general principles.

Table 5: Memory storage vs. performance on ALFWorld valid-unseen (3-seed mean). Valid-unseen performance is provided for reference; storage analysis to be completed for camera-ready.

| Method | Storage (MB) | Valid-unseen, mean (%) | Perf / MB (↑) |
|---|---|---|---|
| ExpeL | – | **37.3** | – |
| MEMORIA (Full) | – | 36.8 | – |
| Mem$^2$Evolve | – | 36.0 | – |
| Flat Episodic | – | 35.1 | – |
| Reflexion | – | 33.3 | – |
| No Memory | – | 28.7 | – |

## H RL Training with MEMORIA: Supplementary Experiments

To explore whether MEMORIA can improve LLM *training* (not just inference-time retrieval), we conducted supplementary RL and supervised fine-tuning experiments on Qwen2.5-7B-Instruct with HotpotQA. These results are preliminary and not part of the main evaluation; the primary contribution of this paper is the inference-time memory architecture.

**Setup.** We fine-tune Qwen2.5-7B-Instruct (Yang et al., 2024) with LoRA (rank=16, $\alpha$=32) on 10,000 HotpotQA training questions. Evaluation uses 500 held-out questions with exact match (EM) and a composite reward $0.6\times$EM$+0.4\times$F1. We compare five training configurations:

- **SFT**: Supervised fine-tuning on gold answers (1 epoch).
- **SFT+MEMORIA**: SFT with MEMORIA retrieval augmentation at training time.
- **GRPO**: Group Relative Policy Optimization (Shao et al., 2024) ($K$=4 generations).
- **GRPO+MEMORIA**: GRPO with MEMORIA retrieval augmentation.
- **RLOO**: REINFORCE Leave-One-Out (Kool et al., 2019) ($K$=4).
- **DAPO**: GRPO with dynamic sampling and clip-higher (Yu et al., 2025).

All RL experiments use learning rate $5 \times 10^{-7}$, 1 epoch, and 8×L20Z GPUs.

Table 6: RL training experiments on HotpotQA (Qwen2.5-7B-Instruct, 10K train, 500 eval). EM = Exact Match; Reward = $0.6\times$EM + $0.4\times$F1.

| Method | EM (%) | Reward | Δ EM vs. base |
|---|---|---|---|
| SFT | 29.4 | 0.3343 | – |
| SFT+MEMORIA | 27.6 | 0.3226 | −1.8 pp |
| GRPO | 32.8 | 0.3723 | – |
| GRPO+MEMORIA | **35.8** | **0.4047** | **+3.0 pp** |
| RLOO | 33.0 | 0.3726 | – |
| DAPO | 40.0 | 0.4425 | – |

**Results.**

**Findings.**

1. **MEMORIA benefits GRPO training.** Adding MEMORIA retrieval during GRPO training improves EM from 32.8% to 35.8% (+3.0 pp) and reward from 0.3723 to 0.4047. The retrieved episodic context provides useful exploration guidance during RL rollout generation.
2. **MEMORIA hurts SFT.** With supervised fine-tuning, MEMORIA retrieval reduces EM by 1.8 pp (29.4%→27.6%). We attribute this to distribution mismatch: SFT trains on gold answers conditioned on context, while MEMORIA injects potentially noisy retrieved principles that interfere with fitting the supervised signal.
3. **Algorithm ranking.** DAPO achieves the highest EM (40.0%) across all configurations, suggesting that dynamic sampling strategies are particularly effective for multi-hop QA. The ranking is: DAPO (40.0%) $>$ GRPO+MEMORIA (35.8%) $>$ RLOO (33.0%) $>$ GRPO (32.8%) $>$ SFT (29.4%).

These results suggest that MEMORIA's episodic context can serve as a useful exploration signal during on-policy RL training, while it can interfere with distribution-matching objectives in SFT. The interaction between memory retrieval and training objective warrants further investigation.