SAME CURVE, DIFFERENT REGIMES: A LONGITUDINAL AUDIT OF MEMORY-AUGMENTED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory-augmented agents adapt after deployment by accumulating interaction experience, and their aggregate success often rises before flattening. Does a plateau indicate a stable capability gain? Similar curves can conceal a stable solved-task set, gains offset by losses, changes at the scale of repeated execution, or a task bank with little remaining discriminative range. We propose a longitudinal audit of frozen memory states that jointly examines across-seed coverage, matched reruns of the same state, and the item-level composition of change. On a fixed ALFWorld bank, three backbone operating points and two memory mechanisms reveal gain–loss cancellation, repeat-scale change, and ceiling-limited evidence. On a smaller ScienceWorld panel, two further backbones show different early trajectories and later changes near the repeat scale. A plateau should therefore prompt this audit before it supports a claim of convergence.

1 INTRODUCTION

Memory-augmented agents accumulate trajectories, reflections, rules, or skills after deployment while keeping their language-model weights fixed (Shinn et al., 2023; Wang et al., 2024; Chen et al., 2024). Their effective policy changes with use, and evolving-memory methods are commonly evaluated by success across successive states (Zhang et al., 2026a;b; Wei et al., 2025). A rising curve may then flatten, but aggregate success alone cannot tell whether the agent preserves the same solved tasks or has acquired stable behavior.

Consider two snapshots that both solve 62% of tasks. One preserves almost the same solved set; the other loses twenty solved tasks and gains twenty different ones. A flat curve may also reflect outcomes that flip when the identical state is rerun, or a task bank on which nearly every item is already solved. These explanations call for different next steps: track regressions, collect more repeats, or extend the benchmark.

The ambiguity is one of identification, not an objection to averaging. Multi-turn evaluation has similarly separated aptitude loss from increased unreliability behind one aggregate gap (Laban et al., 2026); here the unit is an agent's own frozen memory state. Figure 1 presents three complementary checks: across-seed coverage bounds what the bank can distinguish, matched reruns calibrate observed variation, and gains and losses reveal the direction of item-level change. We read them together rather than compressing them into another score.

We evaluate this measurement question on a fixed 140-task ALFWorld bank (Shridhar et al., 2021), with three Qwen operating points, three seeds, six MemRL states, matched repeats, and controls. A second memory mechanism, AdaMEM, is audited on the same bank; a smaller 30-item ScienceWorld panel probes two other backbones. Table 1 previews how similar ALFWorld curves lead to different readings. Across panels, the audit tests whether its diagnostics remain useful, not whether one mechanism or benchmark is superior.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

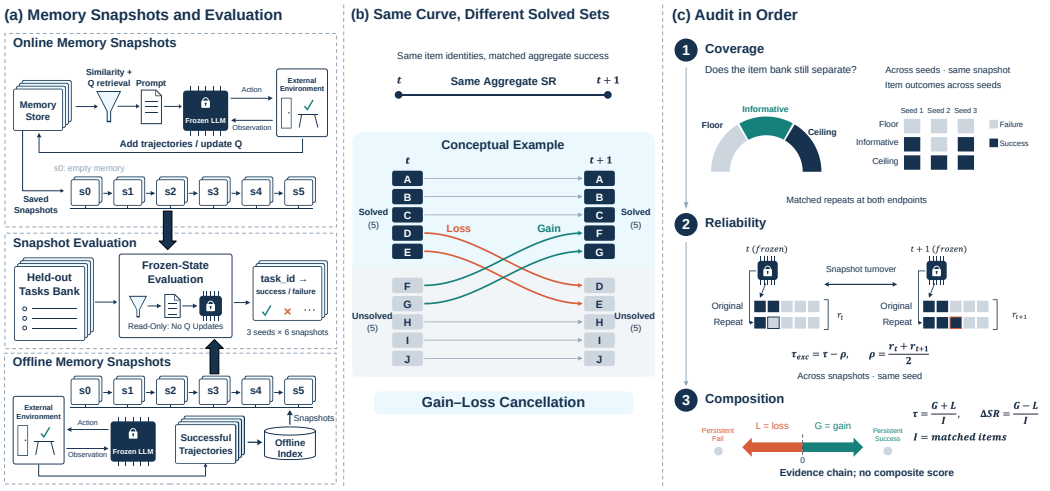


Figure 1: Overview of the longitudinal audit. (a) Online memory updates and offline index prefixes yield frozen states evaluated read-only on the same task bank, with one matched rerun per state. (b) Equal aggregate success can conceal gains offset by losses. (c) Across-seed coverage, same-state disagreement, and gains and losses provide complementary evidence about an apparent plateau.

Table 1: Diagnostic preview of ALFWorld trajectories (three-seed means). 14B/27B denote Qwen2.5-14B/Qwen3.6-27B; late means after s1. Excess is turnover minus the matched-repeat baseline. Ranges are observed transitions, not confidence intervals.

Trajectory	Diagnostic evidence	Reading / next check
MemRL 14B (s0–s5)	19.8–31.2% flips; 13.6–21.5% excess.	Cancellation; report losses.
MemRL 27B (post-s1)	1.2–1.9% flips; 94.3–95.0% at ceiling.	Ceiling; add harder tasks.
AdaMEM 14B (late)	17.1–19.5% flips; 1.4–2.7% excess.	Repeat scale; add reruns.

Our contributions are:

1. **Finding.** Near-flat success can conceal gain–loss cancellation, change at the scale of a frozen-state rerun, or weak evidence on a saturated task bank; these cases warrant different evaluation decisions.
2. **Audit.** We bound what aggregate scores can identify and combine across-seed coverage, matched-repeat disagreement, and item-level gains and losses in an evidence-preserving report.
3. **Evidence and scope.** A dense ALFWorld panel, a second memory mechanism, and a smaller ScienceWorld panel show how the same audit yields different diagnoses. Cross-panel contrasts are descriptive because model families, mechanisms, and protocols are not all matched.

2 Related Work

Runtime memory and longitudinal evaluation. Agents can accumulate reflections, skills, or indexed experience without changing model weights (Shinn et al., 2023; Wang et al., 2024; Chen et al., 2024; Zhang et al., 2026a;b). Sequential benchmarks study their use of experience (Hu et al., 2026; Wei et al., 2025; Wang et al., 2026). Closer longitudinal studies compare read-only memory snapshots for safety (Al-Tawaha et al., 2026) or memory architectures across history lengths with stochastic replicates (Spencer, 2026). SEA-Eval uses token efficiency to question apparent progress from success alone (Jiang et al., 2026). We instead audit the success composition, same-state reliability, and remaining task-bank range of one agent's frozen-state trajectory.

What aggregate success hides. AgentBoard measures within-episode progress beyond final success (Ma et al., 2024); multi-turn evaluation separates aptitude from reliability (Laban et al., 2026). Model-

update studies track negative flips masked by mean accuracy (Träuble et al., 2021; Yan et al., 2021). We retain that item-identity insight across frozen runtime-memory states, and add an identical-state rerun to distinguish snapshot turnover from execution variability.

Reliability and benchmark range. Resampling and repeated agent runs expose evaluation variability (Lior et al., 2025; Mustahsan et al., 2025). Item response theory studies discrimination, while dynamic benchmarks address saturation (Lord, 1980; Kiela et al., 2021; Wang et al., 2025). Our coverage statistic is narrower: floor, informative, and ceiling fractions across observed seeds on a fixed bank. Together with reruns and item-level transitions, it limits what low turnover can establish about broader stability.

3 Longitudinal Audit

3.1 Problem Formulation

For each seed s, memory state t, and item $i \in \{1, \ldots, I\}$, retain the binary outcome $y_{s,t,i}$ rather than only its mean. Conventional evaluation reports

$$\mathrm{SR}_{s,t} = \frac{1}{I} \sum_i y_{s,t,i}. \tag{1}$$

For two snapshots with success rates a and b, define turnover τ as the fraction of items whose outcome changes. Marginal scores constrain but do not identify turnover:

$$|a - b| \leq \tau \leq \min(a + b,\, 2 - a - b). \tag{2}$$

Equal 60% scores, for example, allow anywhere from zero to 80% turnover; a flat curve does not establish item-level stability.

3.2 Sequential Diagnostics

1. Coverage: does the benchmark retain discriminative headroom? With S seeds, let $p_{t,i} = S^{-1} \sum_s y_{s,t,i}$. We call an item floor when $p_{t,i} = 0$, ceiling when $p_{t,i} = 1$, and informative otherwise. *Coverage* is the informative fraction of the fixed bank at a state; we also report floor and ceiling fractions. With three seeds, mean binary entropy is a constant multiple of this fraction and adds no independent diagnostic. Coverage describes separation among the observed seeds, not an intrinsic item property. A large ceiling mass does not rule out temporal changes; it limits how strongly low turnover on this bank supports broader stability.

2. Reliability: does change exceed repeated execution? For every accepted memory state, we reload the exact frozen state and rerun the same item bank without writes. The disagreement between the original and repeat vectors is the repeat flip rate r_t. We define the *repeat baseline* of an adjacent transition as $\rho_{t \to t+1} = (r_t + r_{t+1})/2$, and the resulting *excess turnover* as

$$\tau^{\mathrm{exc}}_{t \to t+1} = \tau_{t \to t+1} - \rho_{t \to t+1}. \tag{3}$$

Near-zero excess means observed snapshot change is at the same scale as rerunning a frozen state; positive excess means that turnover remains beyond that scale. This subtraction is descriptive, not an effect-size estimator. We retain negative values. Each r_t comes from one matched repeat; alternative baselines and their sensitivity appear in the Supplementary Material.

3. Composition: what changed? Between two snapshots, each item either keeps its outcome or changes it; we count gains G (failure to success) and losses L (success to failure). We report

$$\tau = \frac{G + L}{I}, \qquad \Delta\mathrm{SR} = \frac{G - L}{I}, \tag{4}$$

plus failure-set Jaccard overlap. Large G and L with small $\Delta\mathrm{SR}$ indicate *gain–loss cancellation*. Together, coverage, repeat disagreement, and composition distinguish reasons that a plateau alone does not establish convergence; they are not an automatic classifier. Within a trajectory, weights, seed, item identities, and evaluation protocol are fixed while the memory state changes, supporting a bounded within-protocol causal interpretation. Excess turnover remains descriptive rather than a formal causal-effect estimate; cross-mechanism and no-memory comparisons are non-causal because their protocols differ.

4 EXPERIMENTAL SETUP

We ask whether near-flat success preserves the solved-item set, exceeds same-state variability, and occurs on a bank with remaining discriminative range. We then apply the same audit to another memory mechanism and to a smaller second-benchmark panel.

4.1 BENCHMARK

Our primary bank has 140 fixed ALFWorld `valid_seen` household tasks (Shridhar et al., 2021). Item identities are shared across its seeds, states, and mechanisms; held-out evaluation outcomes never update memory.

The ScienceWorld 1.2.3 extension (Wang et al., 2022) uses one fixed easy-setting test variation from each of 30 task families. Strict success requires an official score of 100; we also retain the native 0–100 score. Its item bank differs from ALFWorld, and no backbone is tested on both.

4.2 MODELS AND MEMORY SYSTEMS

Backbones. ALFWorld uses Qwen3.5-9B (Qwen Team, 2026a), Qwen2.5-14B-Instruct (Yang et al., 2024), and Qwen3.6-27B (Qwen Team, 2026b). ScienceWorld uses API-served DeepSeek-V4-Pro (DeepSeek-AI, 2026) and GLM-5.2 (Z.ai, 2026), with internal IDs `deepseek-v4-pro` and `glm-5.2`. We do not update backbone weights. Cross-panel differences confound model family and benchmark.

Memory mechanisms. MemRL (Zhang et al., 2026a) adds episodic trajectories and updates their Q-values. ALFWorld has six indexed states *s0–s5*; ScienceWorld uses *c0/c30/c60/c90*, named for cumulative training attempts per seed. Our ALFWorld AdaMEM (Zhang et al., 2026b) extension freezes six prefixes of an offline trajectory index, also labeled s0–s5. The added s2 and s4 indices use existing trajectories, without new training. MemRL uses Qwen3-Embedding-8B; AdaMEM uses Qwen3-Embedding-4B (Zhang et al., 2025). This AdaMEM instrumentation is not its originally reported evaluation protocol.

Controls. The Qwen2.5-14B study includes a no-memory reference without retrieval or writes and a training-only reflection-memory (TRM) contrast that never incorporates held-out outcomes. Prompt and harness differences make these comparisons descriptive; training exposure and index embedders also differ between MemRL and AdaMEM.

4.3 EVALUATION PROTOCOL

Snapshot construction. Each MemRL transition adds a fixed block of training episodes; ScienceWorld uses 30-attempt blocks. AdaMEM prefixes correspond to 0, 1,065, 2,130, 3,195, 4,260, and 5,325 attempted no-memory training tasks. The ALFWorld s-labels index states within each mechanism, not matched exposure or interventions across mechanisms.

Frozen-state evaluation and repeats. For seeds 42, 123, and 456, each MemRL, TRM, and AdaMEM state is evaluated once and rerun after reloading the identical state. Runs are read-only; item identities are paired within seeds across states and between original and repeat. ALFWorld no-memory references have no matched repeats.

Evaluation scale and integrity. A cell is one full-bank run for a condition, backbone, seed, state, and execution. The main ALFWorld analysis contains 198 cells (27,720 task records): 162 MemRL-study cells including controls and 36 AdaMEM cells. A separate AdaMEM no-memory reference adds 12 cells; the two ScienceWorld panels add 24 cells and 720 episodes each. Each cell has all 140 ALFWorld or 30 ScienceWorld task IDs. Detailed run accounting appears in the Supplementary Material.

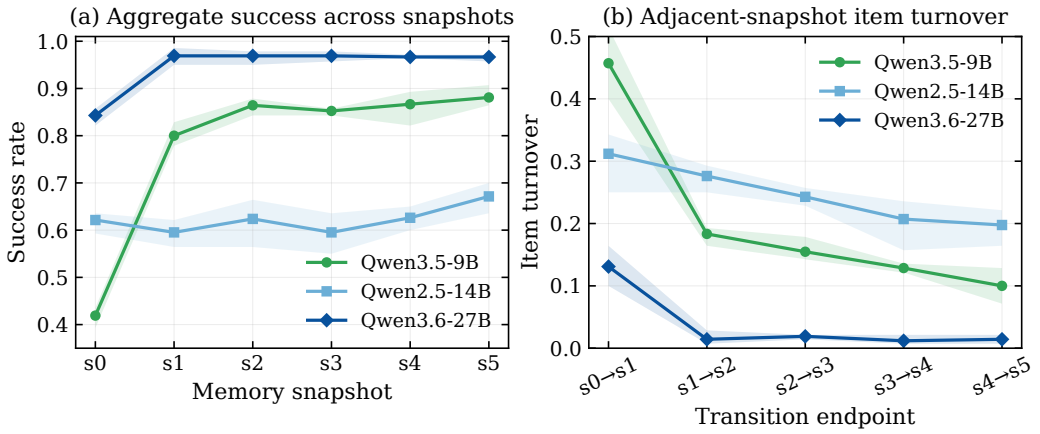


Figure 2: ALFWorld aggregate success and adjacent item turnover (three-seed means). Qwen9B improves sharply at s1; Qwen14B retains turnover despite small net changes; Qwen27B has little turnover after s1. Shading spans observed values across the three seeds.

4.4 Metrics and Aggregation

We report state success and coverage; same-state repeat disagreement; and transition gains, losses, raw and excess turnover, and failure-set overlap (Section 3). State and transition rates are three-seed means; coverage classifies each item across the three original runs. ΔSR uses percentage points, while rates use percentages of each panel's own bank. All reported transitions are adjacent within their panel, including AdaMEM s1→s2 through s4→s5. Shading in Figures 2 and 4(a,c) spans observed values across the three seeds. These ranges are not population-level confidence intervals, and transition ranges are not confidence intervals.

Seed spread and matched-repeat disagreement answer different questions. The former combines differences among training and memory trajectories with variation in their original executions; the latter measures change when an identical frozen state is executed again. We do not treat overlap between shaded bands as a significance test or use seed spread in place of the matched-repeat baseline. More seeds would test how reproducible a trajectory is, whereas more reruns would sharpen the within-state reliability estimate.

5 Results and Analysis

5.1 Aggregate Curves Hide Item Replacement

Figures 2 and 3 show the ALFWorld trajectories; Table 2 gives every adjacent transition. Qwen3.5-9B is initially gain-dominated, rising from 41.9% to 80.0% success at s1, then reaching 88.1% at s5 as turnover declines from 45.7% to 10.0%. Qwen2.5-14B MemRL instead fluctuates within 59.5–67.1% after s0: each adjacent net change is at most 4.5 points, yet 19.8–31.2% of items flip. Gains and losses coexist at every step. The curve hides replacement of the solved-item set, so regression identities should accompany aggregate scores.

Qwen3.6-27B also flattens after s1, but for a different reason: success first rises from 84.3% to 96.9%, then only 1.2–1.9% of items flip while 132–133 of 140 are at ceiling across seeds. Low turnover here is measured on a nearly saturated bank; it does not establish stability beyond that bank.

Failure-set overlap adds a view of which errors persist, but its denominator changes with the union of failed items. For Qwen2.5-14B MemRL, overlap rises from 43.3% at the first transition to 56.0% at the last while gains and losses continue. Some failures therefore recur even as other item outcomes are replaced. For Qwen3.6-27B, only a few failures remain after s1, so a similar overlap percentage would be weak evidence about broader ability. We use overlap to locate persistent errors, not as a stand-alone stability score.

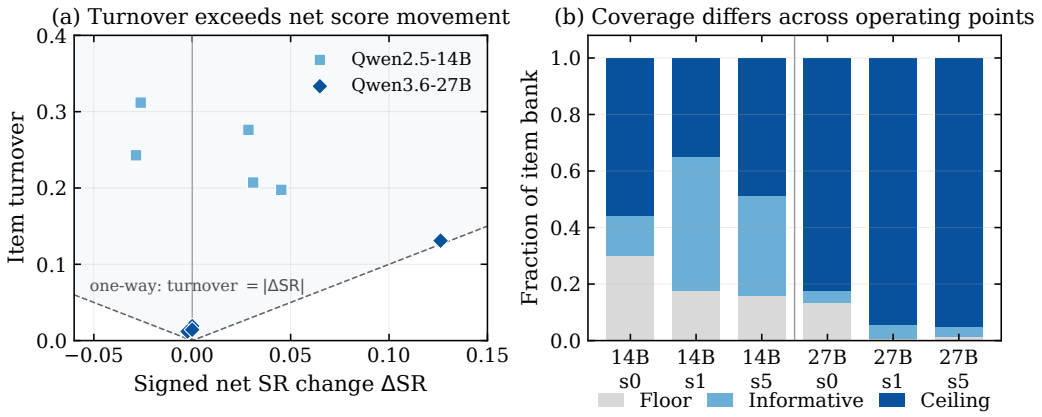


Figure 3: ALFWorld item composition and coverage. Left: turnover above the one-way-change diagonal implies both gains and losses. Right: Qwen27B's ceiling mass reaches 94.3–95.0% after s1, unlike Qwen14B's mixed bank.

Table 2: All adjacent ALFWorld transitions in memory-bearing panels (three-seed means). Entries are percentages except ΔSR in points. G/L: gains/losses; τ: turnover; ρ: endpoint-averaged repeat disagreement; $\tau^{\text{exc}} = \tau - \rho$; J_F: failure-set overlap.

Backbone	Transition	ΔSR	G	L	τ	ρ	τ^{exc}	J_F
MemRL (online memory)								
	s0→s1	+38.1	41.9	3.8	45.7	15.0	30.7	26.3
	s1→s2	+6.4	12.4	6.0	18.3	6.8	11.5	29.1
Qwen3.5-9B	s2→s3	-1.2	7.1	8.3	15.5	6.2	9.3	29.5
	s3→s4	+1.4	7.1	5.7	12.9	7.1	5.7	36.8
	s4→s5	+1.4	5.7	4.3	10.0	7.0	3.0	43.3
	s0→s1	-2.6	14.3	16.9	31.2	9.6	21.5	43.3
	s1→s2	+2.9	15.2	12.4	27.6	7.3	20.4	47.6
Qwen2.5-14B	s2→s3	-2.9	10.7	13.6	24.3	7.4	16.9	52.4
	s3→s4	+3.1	11.9	8.8	20.7	6.3	14.4	58.0
	s4→s5	+4.5	12.1	7.6	19.8	6.2	13.6	56.0
	s0→s1	+12.6	12.9	0.2	13.1	2.1	11.0	18.3
	s1→s2	+0.0	0.7	0.7	1.4	0.7	0.7	65.7
Qwen3.6-27B	s2→s3	+0.0	1.0	1.0	1.9	0.8	1.1	50.8
	s3→s4	-0.2	0.5	0.7	1.2	0.7	0.5	67.8
	s4→s5	+0.0	0.7	0.7	1.4	0.4	1.1	64.4
AdaMEM (offline index)								
	s0→s1	+13.3	20.0	6.7	26.7	18.2	8.5	47.9
	s1→s2	-2.9	7.6	10.5	18.1	15.4	2.7	56.5
Qwen2.5-14B	s2→s3	+1.4	9.5	8.1	17.6	14.9	2.7	58.3
	s3→s4	-1.9	7.6	9.5	17.1	15.7	1.4	59.1
	s4→s5	+1.9	10.7	8.8	19.5	17.0	2.5	55.1

5.2 Turnover Beyond the Repeat Scale

Raw turnover can also reflect stochastic execution. In Qwen2.5-14B MemRL, same-state repeat disagreement falls from 11.4–14.3% at s0 to 3.6–9.3% at s5 across seeds, yet adjacent excess turnover remains 13.6–21.5% (Table 2). Its changing solved set exceeds the observed rerun scale. By contrast, TRM's post-s1 excess is -0.4–0.8%, consistent with repeat-scale change. The same aggregate plateau therefore demands different follow-up: track Qwen14B regressions, but collect more repetitions before attributing TRM's later flips to memory updates.

The controls also separate performance level from evidence of continued change. TRM rises from 60.7% to 70.2% success at s1 and then stays within 71.0–71.4%, with post-s1 excess near zero. The no-memory reference stays within 58.8–61.2% across the same state labels, but lacks matched repeats, so it cannot support an excess-turnover diagnosis. Because its prompt and harness differ, these levels are descriptive controls rather than an estimate of MemRL's effect.

5.3 Ceiling Saturation Limits Stability Claims

Figure 3(b) shows the coverage contrast. Qwen3.5-9B moves from a floor-heavy start toward a large ceiling mass as excess turnover falls from 30.7% to 3.0%; Qwen2.5-14B retains a mixed bank. Qwen3.6-27B instead has only 2.9–5.0% informative items after s1. Its low turnover should prompt harder or refreshed tasks before any broader stability claim. The Supplementary Material shows coverage at every state.

5.4 Similar Turnover, Different Reliability

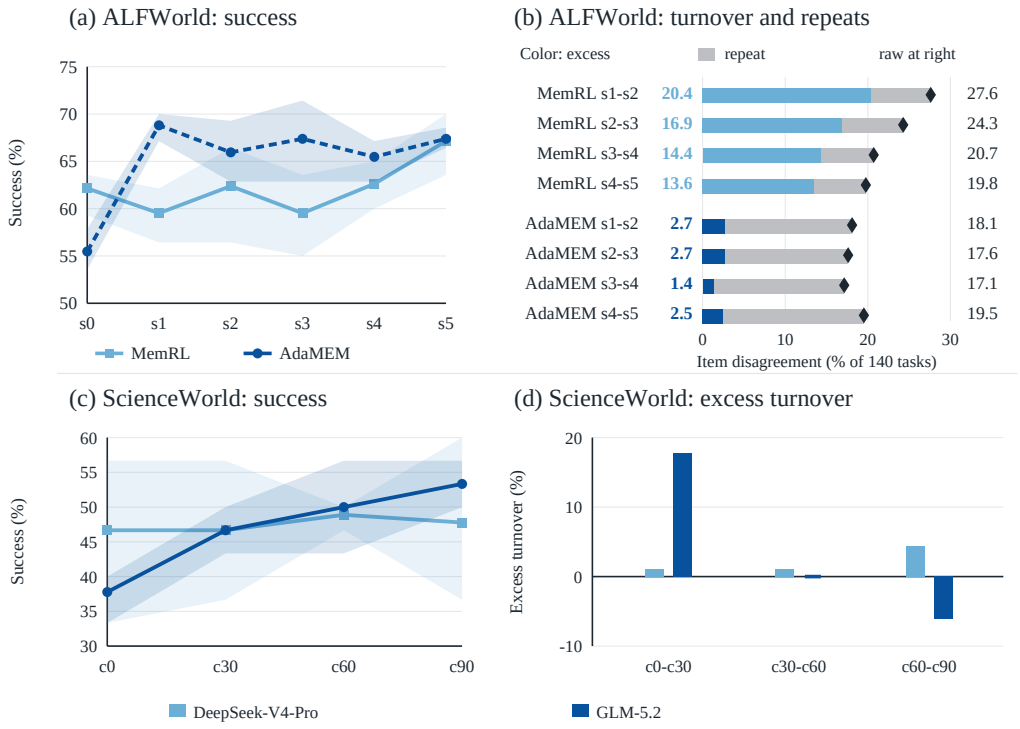


Figure 4: Two extensions of the audit (three-seed means). (a,b) ALFWorld: MemRL and AdaMEM success on the same 140 items, with post-s1 raw turnover separated into excess and matched-repeat components. (c,d) ScienceWorld: success (official score 100) and excess turnover on a separate 30-item bank. Shading in (a,c) spans observed seed values. Colors identify systems within each row; different backbones and protocols preclude cross-panel ranking.

AdaMEM reaches 68.8% mean success at s1 from 55.5% at s0, then stays within 65.5–67.4% through s5. Its first transition has 8.5% excess turnover. Later raw turnover remains 17.1–19.5%, but only 1.4–2.7% exceeds the matched-repeat scale (Figure 4(a,b); Table 2). The informative fraction remains 31.4–33.6%, so this is not the ceiling-limited pattern of Qwen27B.

Across the four post-s1 transitions, Qwen2.5-14B MemRL has 19.8–27.6% raw and 13.6–20.4% excess turnover. AdaMEM's corresponding ranges are 17.1–19.5% and 1.4–2.7% (Figure 4(a,b)). Even where late raw rates are close, reruns change the diagnosis. Different training exposure, harnesses, and embedders preclude a mechanism ranking.

Task examples show what these summaries retain. In MemRL s0→s1, a bowl-under-lamp task becomes solved while a pillow-under-lamp task stops being solved; their opposing contributions cancel in net success. In a nonadjacent AdaMEM s3→s5 comparison, the keychain task changes from failure to success as the original run searches a second drawer rather than revisiting the first; its matched repeats reverse the endpoint outcomes. A CD task changes from success to failure on originals, yet its s3 repeat also fails. These observations make the item-level changes concrete without assigning either outcome to one memory entry. Task IDs and trajectory excerpts appear in the Supplementary Material.

5.5 Different Early Trajectories on ScienceWorld

The same audit remains informative on the smaller ScienceWorld bank, but its two backbones do not share an early trajectory (Figure 4(c,d)). DeepSeek-V4-Pro's strict success is 46.7, 46.7, 48.9, and 47.8% from c0 to c90. Its first mean change is zero while seed changes range from -13.3 to $+23.3$ points; adjacent excess turnover is only 1.1–4.4%. GLM-5.2 instead rises from 37.8% to 46.7% at c30, with gain-dominated turnover, 17.8% excess, and improvement in all three seeds. It reaches 50.0% and 53.3% at c60 and c90, but later mean excess is 0.0 and -6.1% (Table 3). The diagnostic separates GLM's early directional change from both backbones' later repeat-scale changes; later plateaus call for more reruns rather than a convergence verdict. Seed-level results are in the Supplementary Material.

Table 3: All adjacent ScienceWorld transitions (30 fixed items; three-seed means). c0/c30/c60/c90 denote states after 0/30/60/90 training attempts. Strict success requires an official score of 100. Entries are percentages except ΔSR in points; ρ is endpoint-averaged repeat disagreement, $\tau^{\mathrm{exc}} = \tau - \rho$, and J_F is failure-set overlap.

Backbone	Transition	ΔSR	G	L	τ	ρ	τ^{exc}	J_F
DeepSeek-V4-Pro	c0→c30	+0.0	10.0	10.0	20.0	18.9	1.1	68.9
	c30→c60	+2.2	7.8	5.6	13.3	12.2	1.1	77.3
	c60→c90	-1.1	7.8	8.9	16.7	12.2	4.4	72.6
GLM-5.2	c0→c30	+8.9	18.9	10.0	28.9	11.1	17.8	60.1
	c30→c60	+3.3	7.8	4.4	12.2	12.2	0.0	78.6
	c60→c90	+3.3	4.4	1.1	5.6	11.7	-6.1	90.0

Only 16.7–33.3% of ScienceWorld items are informative across seeds, and one item shifts a per-seed rate by 3.3 points. Strict success also misses partial progress: from c30 to c60, GLM's mean strict success rises from 46.7% to 50.0% while native score falls from 61.0 to 59.4. This metric difference is descriptive, not evidence of regression. Because model family and benchmark change together, the extension does not identify a benchmark effect or demonstrate cross-benchmark transfer.

6 Discussion

From diagnosis to evaluation decision. A snapshot report should retain task IDs with seed, state, and execution; show gains, losses, turnover, repeats, and floor/informative/ceiling fractions. These measures point to different next steps. Gain–loss cancellation calls for reporting which tasks regressed; repeat-scale turnover calls for more reruns; a saturated bank calls for harder or refreshed tasks. Evidence of stability is stronger only when the bank still discriminates, repeated execution is reliable, and the solved set no longer changes materially.

The checks must be read together. Composition alone would overstate AdaMEM's late reorganization, because much of its turnover is visible when a frozen state is rerun. Repeat correction alone would overstate Qwen27B's stability, because its ALFWorld bank is nearly saturated. Aggregate success and coverage alone would miss Qwen14B's persistent gains and losses. Keeping these separate measurements preserves the reason for each evaluation decision.

The two extensions test different parts of that reasoning. ALFWorld holds the task IDs and one backbone fixed across memory mechanisms, showing why raw turnover and repeat-adjusted turnover should both be reported. ScienceWorld checks whether the same diagnostic questions remain useful

under a different task and backbone, without isolating which change matters. The transferable result is the audit procedure, not a numerical threshold or a ranking across panels.

Near-zero excess is not evidence that a memory update had no effect: it means that the observed turnover is comparable in magnitude to one matched rerun. Positive excess, in turn, does not by itself establish improvement; gains and losses determine its direction. The audit therefore narrows a claim to what the observed bank and repeat protocol can support.

What the audit establishes. Within a trajectory, changing the memory state is the intended intervention under a fixed evaluation protocol, supporting a bounded causal reading of observed differences. Excess turnover itself is only a comparison with the observed rerun scale, not a causal-effect estimate. Neither a mechanism ranking nor a claim about why one memory entry caused one outcome follows. Complementary evaluations can examine efficiency (Jiang et al., 2026) or safety (Al-Tawaha et al., 2026); the present audit asks what the task-success curve actually supports.

7 LIMITATIONS

The primary bank contains 140 ALFWorld tasks; ScienceWorld has only 30 and low informative coverage. No backbone spans both benchmarks, and ALFWorld mechanisms differ in prompts, harnesses, exposure, and embedders. Three seeds and one repeat per state do not support population-level uncertainty or a formal effect estimate. Binary success omits partial progress; the ScienceWorld exports contain seed-level summaries but not item-level episodes, and its API backbones lack recorded immutable versions. These limits restrict generalization and task-level reproducibility.

8 CONCLUSION

A flat success curve is a starting point for diagnosis, not evidence of convergence. Item-level gains and losses, frozen-state reruns, and benchmark coverage distinguish hidden regressions, repeat-scale change, and limited measurement range across our observed panels. Reporting them with each snapshot makes claims of learning and stability more specific and testable.

AI USE STATEMENT

Generative AI assisted with literature retrieval, manuscript drafting and editing, figure revision, methodological discussion, interpretation checks, and evaluation-code review. The authors designed the study, verified all AI-assisted work, and take responsibility for the results and claims. No research data were AI-generated.

ETHICS STATEMENT

This work uses no human-subject data. It studies benchmark outcomes and training-derived agent state. We report negative and inconclusive findings, keep causal claims bounded to the controlled within-protocol comparison, and do not advocate using the audit to selectively hide regressions or benchmark failures.

REPRODUCIBILITY STATEMENT

The ALFWorld primary numbers are recomputed from item-level JSONL artifacts with explicit run IDs, state hashes where recorded, original–repeat pairings, and integrity checks. The ScienceWorld extension uses accepted aggregate exports whose state and transition summaries reconcile arithmetically; the available exports do not yet permit independent reconstruction of task-level pairings. The analysis retains negative excess-turnover values and separates descriptive from causal claims. The Supplementary Material gives the selection rules and analysis definitions; the anonymized data bundle contains available ALFWorld item outcomes and accepted extension summaries.

REFERENCES

Ahmad Al-Tawaha, Shangding Gu, Peizhi Niu, Ruoxi Jia, and Ming Jin. Remembering more, risking more: Longitudinal safety risks in memory-equipped LLM agents. arXiv preprint arXiv:2605.17830, 2026. URL https://arxiv.org/abs/2605.17830.

Minghao Chen, Yihang Li, Yanting Yang, Shiyu Yu, Binbin Lin, and Xiaofei He. AutoManual: Constructing instruction manuals by LLM agents via interactive environmental learning. In *Advances in Neural Information Processing Systems*, 2024.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. Official model card, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026.

Sihang Jiang, Lipeng Ma, Zhonghua Hong, Keyi Wang, Zhiyu Lu, Tengfei Wang, Shisong Chen, Jinghao Zhang, Tianjun Pan, Weijia Li, Jiaqing Liang, and Yanghua Xiao. SEA-Eval: A benchmark for evaluating self-evolving agents beyond episodic assessment. arXiv preprint arXiv:2604.08988, 2026. URL https://arxiv.org/abs/2604.08988.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of NAACL-HLT*, 2021.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026.

Gili Lior, Eliya Habba, Shahar Levy, Avi Caciularu, and Gabriel Stanovsky. ReliableEval: A recipe for stochastic LLM evaluation via method of moments. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11146–11153, 2025.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, 1980.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2024.

Zairah Mustahsan, Abel Lim, Megna Anand, Saahil Jain, and Bryan McCann. Stochasticity in agentic evaluations: Quantifying inconsistency with intraclass correlation. arXiv preprint arXiv:2512.06710, 2025. URL https://arxiv.org/abs/2512.06710.

Qwen Team. Qwen3.5: Towards native multimodal agents. Official model release, 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model. Official model release, 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

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

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

Quentin Spencer. Ground truth first: A longitudinal evaluation instrument for agent memory, and the tenure crossover in memory-architecture rankings. arXiv preprint arXiv:2607.21962, 2026. URL https://arxiv.org/abs/2607.21962.

Frederik Träuble, Julius von Kügelgen, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Peter V. Gehler. Backward-compatible prediction updates: A probabilistic approach. In *Advances in Neural Information Processing Systems*, 2021.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022. URL https://aclanthology.org/2022.emnlp-main.775/.

Siyuan Wang, Zhuohan Long, Zhihao Fan, Xuanjing Huang, and Zhongyu Wei. Benchmark self-evolving: A multi-agent framework for dynamic LLM evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3310–3328, 2025.

Yuyao Wang, Zhongjian Zhang, Mo Chi, Kaichi Yu, Yuhan Li, Miao Peng, Bing Tong, Chen Zhang, Yan Zhou, and Jia Li. EvoMemBench: Benchmarking agent memory from a self-evolving perspective. arXiv preprint arXiv:2605.18421, 2026. URL https://arxiv.org/abs/2605.18421.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. Evo-Memory: Benchmarking LLM agent test-time learning with self-evolving memory. arXiv preprint arXiv:2511.20857, 2025. URL https://arxiv.org/abs/2511.20857.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

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

Z.ai. GLM-5.2 model card. Official model card, 2026. URL https://huggingface.co/zai-org/GLM-5.2.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Zhuo Li, Yujie Zheng, Weinan Zhang, Ying Wen, Zhiyu Li, Feiyu Xiong, Yutao Qi, Bo Tang, and Muning Wen. MemRL: Self-evolving agents via runtime reinforcement learning on episodic memory. arXiv preprint arXiv:2601.03192, 2026a. URL https://arxiv.org/abs/2601.03192.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176, 2025. URL https://arxiv.org/abs/2506.05176.

Yunxiang Zhang, Yiheng Li, Ali Payani, and Lu Wang. AdaMEM: Test-time adaptive memory for language agents. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.