MEMLADDER: CONTROLLED AUDITING OF AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent memory is commonly evaluated by comparing endpoint success with and without retrieved experience, but that comparison is not an attribution test for the retrieved text. It jointly changes the prompt channel, control text, action policy, exposure duration, and realized continuation. We introduce MemLadder, a controlled audit that separates behavioral rerouting, single-decision downstream value, and remaining-episode exposure-policy return. Its five mutually exclusive conditions range from bare context to retrieved source-grounded memory, including structural, length- and format-matched generic, and mismatched-real controls. From the same restored ALFWorld state, MemLadder executes each induced action, removes exposure text, and estimates return with five repeated memory-free executions under the frozen no-memory serving configuration. In a contemporaneous, human-verified audit, retrieved memory versus an actionable domain-general control produced actor-state-averaged substantive rerouting rates of 25.6% and 28.7%. On a fresh episode-disjoint 180-state panel, however, the one-step value contrast was -0.67 percentage points for both Qwen actors (episode 95% CI $[-1.78, +0.45]$). Under the binary $K = 5$ instrument, 94–98% of fresh-panel states were at an empirical text-arm floor or ceiling, so this small average contrast is a low-resolution estimate rather than evidence that memory is generally valueless. An exhaustive environment-only audit of all 5,231 admissible state–action choices finds a controller-relative outcome boundary in 50/180 states (27.8%), yet retrieved memory and the actionable control fall on opposite sides in only 2/180 and 5/180 states for 35B and 9B, respectively; mean preservation is 90.2% on controller-solvable states. Thus both weak-continuation failure and strong-continuation recovery can compress observed one-step separation. Behavioral rerouting, one-step downstream value, and remaining-episode exposure-policy return therefore require separate controls and should be reported as distinct estimands.

1 INTRODUCTION

Language-model agents increasingly retrieve past trajectories, reflections, and procedural rules to guide later decisions (Park et al., 2023; Packer et al., 2023; Zhong et al., 2023; Zhang et al., 2024). In interactive environments, this experience reuse ranges from verbal self-reflection (Shinn et al., 2023; Zhao et al., 2024) to raw trajectory exemplars (Zheng et al., 2024), distilled strategies (Ouyang et al., 2026), and state-conditioned guidance (Zhang et al., 2026). Their evaluation is commonly endpoint-oriented: run an agent with memory, run it without memory, and compare terminal success. That contrast is informative about the deployed system, but it is not by itself an attribution test for historical content.

MemLadder is a controlled evaluation and attribution protocol for agent memory, not a memory-repair algorithm or a learned exposure policy. It uses a five-arm control ladder, same-state action interventions, and repeated memory-free suffixes to keep policy rerouting, one-step decision value, and remaining-episode exposure-policy return as separate measurement targets.

Memory exposure changes several things at once. It opens a prompt channel, introduces control text, inserts source-grounded text, selects a historical item, changes its location and length in context, and may persist across multiple decisions. Any induced first action then hands control to a long-horizon continuation under a serving policy whose repeated executions may vary. Prompt wording,

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.

demonstration format, irrelevant content, ordering, and location can each change model behavior (Webson & Pavlick, 2022; Min et al., 2022; Shi et al., 2023; Sclar et al., 2024; Liu et al., 2024a; Zhao et al., 2021; Lu et al., 2022). Consequently, a memory/no-memory terminal difference combines content, channel, generic-control, match, dose, and continuation effects.

This matters for both science and system design. A successful episode may award credit to every retrieved item even if the agent would have succeeded without them. Conversely, one unlucky continuation may condemn a useful induced action. Our diagnostic replays show that identical first actions and post-action states can lead to different terminal outcomes. We therefore distinguish

$$\text{behavioral rerouting} \neq \text{single-decision value} \neq \text{deployment utility},$$
$$Y_{\text{terminal}} \not\equiv Q^{\pi_0}(h, a).$$

The middle quantity is the expected value of an induced first action under a frozen no-memory continuation policy π_0, not the value of a fully memory-enabled policy.

We instantiate MemLadder in ALFWorld (Shridhar et al., 2021), an interactive household benchmark built from TextWorld (Cote et al., 2018). From one restorable state, five mutually exclusive arms supply bare context, a structural null, length-matched generic control text, a real but mismatched historical memory, or the naturally retrieved memory. We execute each arm's first action, remove all exposure text, and share suffix evaluations only when the complete continuation identity matches. We estimate downstream value from 5 uncached no-memory suffixes. A separate persistent panel estimates remaining-episode exposure-policy contrasts. Finally, as a post-primary validation, a finite-state planted suite tests whether the analysis pipeline recovers analytically known effects. It was completed after the main ALFWorld heldout analysis and does not retroactively change endpoint status.

The results separate influence from value. In the strongest contemporaneously interleaved post-primary repeat panel, retrieved and actionable-generic text produced mean cross-repeat action/state disagreement of 25.55%/28.83% and substantive rerouting of 25.55%/28.65% for 35B/9B, versus 15.33%/20.62% for actionable-generic text against structural null. An earlier independently annotated repeat audit showed the same direction: 24.54% versus 15.24% (paired difference +9.31 pp after averaging four repeat-pair comparisons within each actor-state). We then used a separate 180-state episode-disjoint value panel, Fresh180. In that panel, H4 minus HP was -0.67 pp for both actors, with 94–98% of states at an empirical text-arm floor or ceiling. Increasing replication to $K = 25$ did not materially expand nonzero support, and a fixed-expert continuation did not uncover a stable H4 advantage. A distinct expert/random prefix panel provides a multi-action detection check rather than evidence that the one-step audit has the same sensitivity.

The paper makes three contributions:

1. MemLadder distinguishes behavioral rerouting, one-step value under Q^{π_0}, and remaining-episode exposure-policy return as distinct agent-memory estimands.
2. We provide a five-arm same-state audit with repeated memory-free suffix evaluation. For the original held-out audit, calibrated studentized inference is accompanied by an independent applicability audit that identifies settings where exact continuation sharing makes intervals conservative or places them outside the validated simulation scope.
3. Across independently annotated and prospective panels, source-grounded memory text produces substantial rerouting beyond generic controls, while the contribution of retrieval match is reported separately. Endpoint saturation under π_0 and high recovery under a stronger controller accompany compressed observable one-step return differences. Exhaustive admissible-action execution quantifies the local action-space support for this separation; supporting diagnostics do not reveal a stable retrieved-memory advantage within the evaluated ALFWorld policy and resolution regime.

2 Related Work

Agent memory and experience reuse. Interactive language agents build on text-game and web environments such as Jericho, ScienceWorld, WebShop, WebArena, and AgentBench (Hausknecht et al., 2020; Wang et al., 2022; Yao et al., 2022; Zhou et al., 2024; Liu et al., 2024b), and often use

ReAct-style interleaving of reasoning and action (Yao et al., 2023). Long-term memory systems span retrieval-augmented text stores (Lewis et al., 2020), hierarchical context management (Packer et al., 2023), reflective learning (Shinn et al., 2023; Zhao et al., 2024), social-agent memory (Park et al., 2023), and executable skill libraries (Wang et al., 2023a). Synapse retrieves raw trajectories, ReasoningBank distills reusable strategies, and AdaMEM synthesizes short-term state-conditioned guidance (Zheng et al., 2024; Ouyang et al., 2026; Zhang et al., 2026). Agent Workflow Memory induces reusable routines from past trajectories, while AutoGuide extracts and selects concise context-aware guidelines (Wang et al., 2025; Fu et al., 2024). Together these systems motivate auditing raw experience, distilled workflows, and conditional guidance rather than treating memory as one uniform intervention. We audit representative components of these mechanisms under a common frozen harness; these adaptations are not official endpoint reproductions.

Memory attribution and auditing. CAMEO studies paired factorial effects of memory entries and banks in ALFWorld and motivates separating passenger credit from marginal contribution (Rocchi, 2026). Its deployment estimand includes prompt length, attention displacement, and altered stopping behavior as mediators. Our complementary target is the value of a memory-induced action after explicit channel and text controls with repeated suffixes. MEMAUDIT isolates budgeted memory writing with an exact package oracle (Bhargava & Barrento, 2026). These works reinforce that "memory quality" is not a single end-to-end quantity.

Causal replay, controls, and step value. Causal Agent Replay intervenes on steps, runs futures forward, and includes synthetic planted calibration for long-horizon attribution (Shah, 2026); executed-replay audits also question single-run credit labels (Zhang, 2026). We likewise use actual environment execution, but target memory exposure and repeated value from the induced post-action state. Monte Carlo rollouts and process supervision were already used to estimate or supervise intermediate-step value in reasoning and reinforcement learning (Kazemnejad et al., 2025; Wang et al., 2024); our contribution is their use inside a controlled memory-attribution audit, not the first introduction of rollout-based step evaluation. The control ladder is motivated by evidence that models respond to format, label space, irrelevant context, and prompt position independently of intended semantics (Webson & Pavlick, 2022; Min et al., 2022; Shi et al., 2023; Sclar et al., 2024; Liu et al., 2024a). Retrieval studies further show that irrelevant or random context can alter downstream performance (Cuconasu et al., 2024), motivating explicit non-memory and mismatched-text controls.

MemLadder does not introduce Monte Carlo step-value estimation or counterfactual replay. It combines non-memory and matched-text controls, same-state execution of the induced action, three separated estimands, and diagnostics of whether the one-step contrast has observable support.

Repeated evaluation and inference. Repeated stochastic decoding has long been used to average reasoning paths (Wang et al., 2023b); here replicas estimate a policy-specific environment return rather than vote on a textual answer. We report clustered bootstrap uncertainty (Efron & Tibshirani, 1993; Cameron et al., 2011), paired McNemar tests (McNemar, 1947), Wilson intervals for rates (Wilson, 1927). Human-label reliability is summarized jointly with agreement, Cohen's κ, and Gwet's AC1 (Cohen, 1960; Gwet, 2008).

3 MemLadder: Auditing Memory Across Exposure Regimes

3.1 Three estimands

Let h denote a restorable history and environment state, $A_z \sim \pi_z(\cdot \mid h)$ the possibly stochastic canonical first action induced by arm z, and S'_z its successor state. Here "possibly stochastic" refers to the realized serving policy: the configured decoder is greedy, but repeated uncached requests need not be bitwise identical under the serving stack. The five arms are mutually exclusive interventions, not cumulative prompt additions. *Behavioral rerouting* is

$$R(z, z') = \Pr[A_z \neq A_{z'} \ \vee \ S'_z \neq S'_{z'}]\,. \tag{1}$$

Equation 1 asks whether exposure changed the sampled action or successor state after canonicalization. Distinct canonical actions may still count as rerouting when their successor states coincide; the OR estimand therefore does not remove all behaviorally equivalent substitutions. We separately report

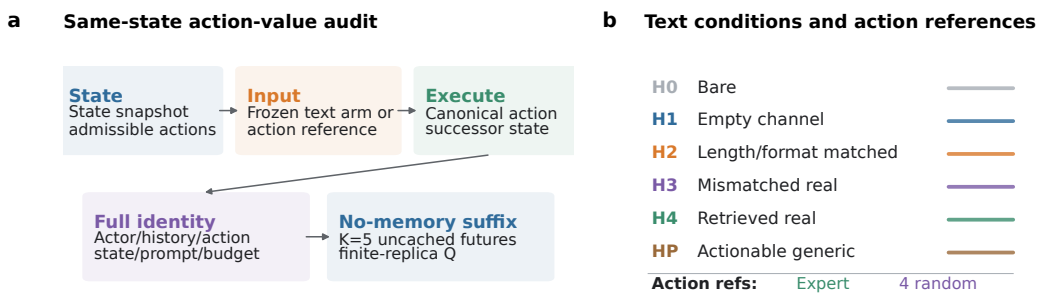

Figure 1: MemLadder's same-state single-decision audit and five mutually exclusive text conditions. H0 is the bare reference and H4 is the retrieved-memory treatment; H1–H3 are matched controls and HP is a stronger post-primary generic control. H2 is length- and format-matched generic text, while HP is an actionable domain-general control. Expert and sampled-random actions are value references, not additional text conditions. All arms branch from the same restored state. The single-decision audit executes one induced action, removes exposure text, and estimates return with repeated no-memory suffixes. A post-primary contemporaneous panel adds the actionable generic control HP and uniformly sampled admissible first actions. Remaining-episode exposure-policy return is measured by a separate experiment and is not depicted by the suffix branch.

action-only, state-only, and joint differences. Human substantive labels operationalize task-level rerouting from observable actions and context, not latent intent. In the analyzed ALFWorld panels, action changes always coincide with successor-state changes; we retain the OR definition because this coincidence is dataset-specific rather than part of the estimand.

For *single-decision downstream value*, all exposure text disappears after the first action. Under the frozen no-memory continuation policy π_0,

$$Q^{\pi_0}(h,a) = \mathbb{E}[Y \mid h, \mathrm{do}(a), \pi_0], \quad V_z(h) = \mathbb{E}_{A\sim\pi_z(\cdot|h)}[Q^{\pi_0}(h,A)]. \tag{2}$$

The frozen single-draw audit estimates

$$\mathrm{SVE} = \mathbb{E}[Q^{\pi_0}(h, a_{H4}) - Q^{\pi_0}(h, a_{H2})]. \tag{3}$$

Equation 3 is the retrieved-memory versus generic-control value contrast, H4–H2. It does not by itself isolate pure historical semantics and is policy- and environment-specific, not an intrinsic memory score. Finally, *remaining-episode exposure-policy return* retains exposure through the rest of the episode. It includes dose, policy adaptation, channel load, and their interactions, and must be interpreted for that exposure regime.

3.2 Five mutually exclusive input conditions

H0 is bare no memory; H1 opens the same wrapper with a structural null; H2 contains length- and format-matched generic text; H3 contains real, source-grounded but deterministically mismatched memory; and H4 contains the naturally retrieved source-grounded memory. H4 is "historical-specific" only as an operational label for retrieved source-grounded text, not as an estimate of pure provenance. The generic control is active, not presumed inert, and mismatch approximates non-relevance rather than guaranteeing it. These matched arms define the operational contrasts

$$\begin{aligned} CE &= Q(H1) - Q(H0), \quad & GSE &= Q(H2) - Q(H1), \\ SGE &= Q(H3) - Q(H2), \quad & RME &= Q(H4) - Q(H3), \\ SVE &= Q(H4) - Q(H2), \quad & NME &= Q(H4) - Q(H0). \end{aligned} \tag{4}$$

The adjacent contrasts in Equation 4 telescope as

$$NME = CE + GSE + SGE + RME, \tag{5}$$

an accounting identity rather than a causal-decomposition theorem. GSE denotes the operational H2–H1 generic-control contrast; it does not identify an independent procedural- scaffold effect. In

particular, H3–H2 contrasts mismatched source-grounded text with generic text; it does not isolate provenance as an intrinsic property. Contrast names describe the frozen intervention, not universal properties of memory systems. The generic control was matched on length and format but was not validated as an equally actionable, executable procedure. Consequently, the study does not rule out that a stronger generic procedural control could induce behavioral or value effects closer to those of retrieved memory. We therefore also audit an actionable domain-general procedural control mapped to the same candidate templates and token-matched under both Qwen tokenizers. The original five-arm H2 contrast is retained for protocol provenance, whereas the later contemporaneous H4–HP contrast is our stronger actionable-control sensitivity. Its earlier H2′ use was a cross-runtime behavioral sensitivity; the contemporaneous validation denotes the frozen same text as HP and evaluates its first-action distribution and downstream value alongside H0 and H4. Neither use replaces the original five-arm endpoint or the separate weak-control factorial study. The stronger-control, post-primary estimand is

$$\tau_{4P}^{\pi_0} = \mathbb{E}[Q^{\pi_0}(h, a_{H4}) - Q^{\pi_0}(h, a_{HP})]\,. \tag{6}$$

We call Equation 6 the retrieved-vs-actionable-control contrast. It is distinct from the original SVE (H4–H2) and RME (H4–H3), and we do not rename it as either quantity.

3.3 Repeated-suffix Q^{π_0}

Each arm generates one action from the same snapshot. We canonicalize and execute it, then construct a full continuation identity containing the actor checkpoint/runtime/decoding fingerprint, target history, exact and canonical first action, post-state, rebuilt no-memory continuation messages, tool and admissible-action state, remaining step budget, and harness state. Suffixes are shared only when this complete identity is equal. An implementation audit found zero raw-action, group-payload, or memory-removal mismatches across 1,697 stored groups (Appendix A). For every unique continuation identity we run 5 uncached no-memory continuations with the same actor model:

$$\widehat{Q}_5^{\pi_0}(h, a) = \frac{1}{5}\sum_{k=1}^{5} Y_k. \tag{7}$$

Although the configured decoding temperature is zero, π_0 is defined operationally as the distribution of executions produced by the frozen no-memory serving stack. Repeated uncached suffixes therefore average observed execution variability under that configuration; we do not attribute its source to a particular backend mechanism. Appendix A reports the stored binary and normalized-trajectory variability by panel and actor. Binary replica-level variation is low under the greedy-serving π_0 configuration: only 2.1–7.1% of identities contain both success and failure among $K = 5$ suffixes, even though 37–73% contain multiple normalized trajectories. Thus repeated suffixes average an operational execution distribution, but the binary endpoint provides limited additional resolution for most identities. This is a finite Monte Carlo reference estimate on a coarse outcome grid, not ground-truth action value and not the value of a memory-enabled continuation policy (Sutton & Barto, 2018). Binary success is the prespecified outcome. Branch-normalized residual goal progress is a prospectively frozen post-primary diagnostic reconstructed from immutable suffixes.

3.4 Calibrated inference

For paired state effects d_i, the state-weighted estimator is $\widehat{\theta} = N^{-1}\sum_i d_i$. Its episode standard error is the sample standard deviation of d_i divided by $\sqrt{N}$. For G retrieved-candidate clusters, the CR1 sensitivity uses centered cluster scores $U_g = \sum_{i\in g}(d_i - \widehat{\theta})$ and

$$\widehat{\mathrm{SE}}_{\mathrm{CR1}} = \frac{1}{N}\sqrt{\frac{G}{G-1}\sum_{g=1}^{G} U_g^2}. \tag{8}$$

Intervals use $|\widehat{\theta} - \theta_0|/\widehat{\mathrm{SE}}$ and the maximum empirical absolute-studentized quantile over the frozen calibration grid. Equal-candidate inference instead averages candidate means and targets a different population.

Frozen calibration grids and independent seeds gave 95.10–96.92% paired-value coverage, at most 4.86% null rejection, and 5.14% maximum boundary false-equivalence (257/5,000), within the

frozen 7% tolerance but not strictly below 5%. An identity-aware applicability audit passed every budget-family gate but found conservative sparse layouts elsewhere. The guarantee is family- and layout-specific; Appendix J gives full grids, Monte Carlo uncertainty, and limits.

Fresh180 is a separate post-primary audit using paired-state Student-t and unit-weighted 24-cluster candidate-CR1 intervals; neither inherits the earlier matched calibration-grid guarantee. Appendix L separately stress-tests sparse exact-identity sharing. The state-weighted mean is the headline estimand; candidate-CR1 and equal-candidate analyses are dependence and weighting sensitivities, and all later diagnostics retain their post-primary status.

4 Experimental Setup

Environment, states, and actor models. ALFWorld contains six household task types and exposes admissible textual actions (Shridhar et al., 2021). Source generation, prospective train states, and official `valid_seen`/`valid_unseen` states were isolated. A 288-state train-only panel preceded one-time opening of 140 seen and 134 unseen official states, one per episode. The actor model is the language model that selects actions. Both Qwen3.5-35B-A3B and Qwen3.5-9B evaluate all 274 states; a prespecified Gemma-4-12B-it panel supplies post-primary cross-family directional evidence. The primary actors were served through an OpenAI-compatible vLLM stack. A post-run audit of the primary serving host and public configuration fields is reported in Appendix A; private infrastructure locators are neither scientific inputs nor release artifacts.

Stimuli and controls. A no-memory source actor produced natural successful trajectories from an outcome-blind, template-grouped pool. A fixed writer generated episodic and atomic procedural candidates. A treatment-blind content audit checked source support, hallucinated rules, abstraction, atomicity, identifier leakage, and actionability. The frozen library contains 94 source-grounded, content-eligible stimuli; it is an audit library, not a production bank. Retrieval, mismatch selection, candidate-specific generic control, actor prompt, action canonicalization, and suffix policy were frozen before official evaluation.

Annotation and inference. For the original single-action panel, every official-heldout adjacent-arm action/post-state shift was side-blinded and independently labeled by two humans, with a third annotator adjudicating disagreements. The repeat and cross-runtime actionable-control audits use a separate versioned mapping: two humans independently labeled 210 new exact-visible tasks (100% primary agreement, $\kappa = 1.000$), a third human adjudicated 71 secondary-label disagreements, and certified exact matches restored the remaining records. This yields 2,236 statistical comparisons with their original 0.25 weights; deduplication reduces annotation effort, not statistical weight or denominator. Value inference uses the predefined studentized procedure for its specified target and dependence structure; behavioral repeat-rate intervals are separate post-primary cluster-bootstrap sensitivities. A later CPU-only applicability audit modeled exact sharing of the same suffix-outcome vector wherever a complete continuation identity recurs. Its frozen gate passed for the budget family but not for paired-value, random-comparator, or prefix families because some configurations were more conservative than the prespecified coverage range. We therefore retain the original rowwise intervals without extending a shared-identity coverage guarantee (Appendix J).

Additional panels. An earlier repeat audit measures same-arm stability and concurrent H0/H1/H2/H4 action distributions. A later H2′ run is a cross-runtime sensitivity because it lacks a same-arm service-window bridge. Separate frozen panels examine text by exposure dose, an earlier real-versus-null mechanism contrast, and common-harness Synapse-style, ReasoningBank-style, and AdaMEM-low components; all are post-primary and not native endpoint reproductions.

Post-primary validation adds 180 train episodes disjoint from prior targets and all 94 stimulus-source episodes, plus expert/random and continuation-policy sensitivities; it does not replace official heldout.

The contemporaneous validation removes the earlier runtime mismatch by interleaving H0–H4 and HP within actor-state blocks in one declared service window. Each arm receives two uncached first-action draws ($6,576$ requests); repeat 0 is fixed for value before outcomes are observed. H0, H4, and HP are evaluated on all 274 states. A deterministic SHA-stratified subset of 180 states adds H1–H3 and four uniformly sampled, without-replacement admissible first actions per state. All

branches use $K = 5$ no-memory suffixes. Complete continuation identity sharing reduces 20,820 logical slots to 10,945 executed or deterministic-terminal replica records without changing arm-level weights.

Component	Frozen setting
Environment	ALFWorld 0.5.0; TextWorld 1.7.0; six task types
Actor models	Qwen3.5-35B-A3B and Qwen3.5-9B
Decoding	temperature 0; thinking disabled; 128-token action response
Context / horizon	12 history turns; 30 total environment steps
Action interface	exact choice from current admissible-command list
Official heldout	140 seen + 134 unseen episodes; one state per episode
Exposure arms	H0–H4, mutually exclusive; one decision only
Memory library	94 source-grounded, content-eligible stimuli
Suffix value	$K = 5$ uncached memory-free continuations per unique identity
Primary value inference	predefined studentized procedure; identity-sharing scope audited separately
Clusters	35 retrieved candidates; 274 episodes (two actor models per episode)
Earlier repeat audit	H0/H1/H2/H4; two uncached actions per arm; 4,384 requests
Contemporaneous validation	H0–H4/HP, $R = 2$; 274-state core, 180-state random panel
New suffix execution	10,945 physical records; 20,820 logical arm–replica slots
Factorial extension	96 train states; real/control $\times$ dose; 3,840 trajectories

Table 1: Core frozen configuration. The 548 official actor-state records are 274 episodes evaluated by two actor models, not 548 independent episodes. The final value procedure distinguishes the stated target estimand and dependence structure; historical one-way percentile intervals are reported separately in Appendix E.

5 RESULTS

5.1 RQ1: CONCURRENT STRONG-TEXT BEHAVIORAL DIFFERENCES

Contemporaneous stronger-control behavior. The post-primary interleaved panel repeats every H0–H4/HP arm twice. For 35B, the cross-repeat action/post-state difference C is 25.55% for H4–HP and 15.33% for HP–H1; for 9B the corresponding rates are 28.83% and 20.62%. The descriptive distribution distance is

$$D_{z,z'} = C_{z,z'} - \tfrac{1}{2}(W_z + W_{z'}), \tag{9}$$

where W_z is within-arm repeat disagreement. It is 22.99% and 27.19% for H4–HP. Within-arm disagreement is 0.73–2.55% across these arms. D is a population-level repeat-distribution contrast, not a denoised behavioral-influence rate and not a human semantic label. In the subsequently frozen human audit, the substantive H4–HP rates were 25.55% and 28.65%, versus 15.33% and 20.62% for HP–H1. Two annotators agreed on all primary labels; a third adjudicated 118 secondary-taxonomy conflicts. These labels establish substantive rerouting, not downstream utility.

Retrieval specificity on the common 180-state subset. Repeat-0 Eq. 1 rates for H3–H2 versus H4–H2 are 21.1/22.8% for 35B and 24.4/22.2% for 9B; H4 and H3 themselves differ on 21.1% and 24.4%. Fixed-expert H3/H4 agreement is 23.3/21.1% for 35B and 20.6/23.9% for 9B, reversing direction across actors. Thus source-grounded text, retrieved or mismatched, reroutes beyond generic controls, while a retrieval-specific directional benefit remains unresolved. Complete descriptive post-primary rates are in Appendix G.

Earlier supporting panels. An earlier 274-state repeat audit showed low within-arm instability and an H4–H2 substantive-rate excess of +9.31 pp over H2–H1. A stronger cross-runtime control retained the direction but lacks a same-arm epoch bridge, so it remains sensitivity evidence. Full rates, provenance, and the original 25.7% versus 14.8% panel are in Appendix D. In the original single-action panel H4 and H3 differ on 25.9% of units, showing behavioral separation but not that H4 selects a higher-value action.

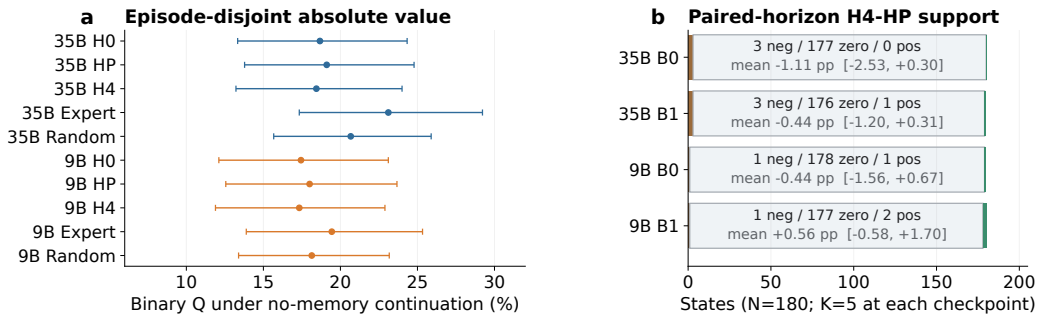

Figure 2: Episode-disjoint post-primary value on one 180-state panel. Panel a reports H0, HP, H4, expert, and four-action-random means under identical state weights. Panel b reports H4–HP sign support and state-level Student-t 95% intervals at B_0 and $B_1 = B_0 + 20$ from fixed first actions and uncached continuations ($N = 180$, $K = 5$). Text-arm floor/ceiling counts fall from 170 to 167 for 35B and from 176 to 172 for 9B. Random is a sampled first-action reference followed by the same no-memory policy, not a random continuation policy.

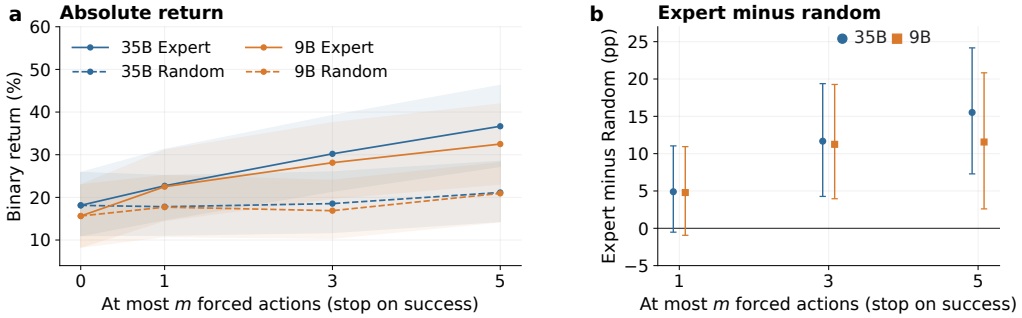

Figure 3: Intervention-length diagnostic on one jointly frozen 96-state panel. Panel a reports absolute return after expert or matched sampled-random prefixes of at most $m \in \{0, 1, 3, 5\}$ actions, stopping early on success. Panel b reports the paired expert-minus-random contrast. Error bars and shading are pointwise state-bootstrap 95% intervals at the discrete values of m; connecting lines are visual guides and provide no continuous-m confidence guarantee. The curve detects a multi-action intervention; it does not establish one-action precision or an intrinsic action-gap ceiling.

5.2 RQ2: Same-panel and episode-disjoint one-step return

On the episode-disjoint panel in Figure 2, the retrieved-vs-actionable-control contrast $\tau_{4P}^{\pi_0}$ in Equation 6 is evaluated alongside the absolute H0/HP/H4/Expert/Random returns, which are 18.67/19.11/18.44/23.11/20.67% for 35B and 17.44/18.00/17.33/19.44/18.14% for 9B. The state-weighted H4–HP contrast is -0.67 pp for both actors (episode 95% CI $[-1.78, +0.45]$ pp); candidate-CR1 sensitivities are $[-1.92, +0.59]$ pp and $[-1.63, +0.29]$ pp. Among the 54 repeat-0 rerouted units per actor, the post-treatment conditional contrast is -2.22 pp (95% intervals $[-6.67, 0.00]$ and $[-6.56, 0.00]$ pp). These conditional intervals are not tests against the overall ± 3 pp equivalence margin. The near-zero H4–HP estimate constrains the observable state-weighted contrast under this frozen one-step instrument; it is not a tight upper bound on latent memory utility.

The binary instrument has limited resolution on this panel. Across H0/HP/H4, 141/180 and 146/180 states are at an empirical all-zero floor for 35B and 9B, while 29/180 and 31/180 are at an empirical all-one ceiling. Thus 170/180 (94.4%) and 177/180 (98.3%) states, respectively, cannot distinguish the three text arms at the observed $K = 5$ resolution. Per-state H4–HP differences are 0 positive, 2 negative, and 178 zero for each actor. A zero difference can therefore reflect an identical continuation

identity, finite-replica coarsening, or an endpoint tie; it is not a universal value-neutrality claim. Retrospective 80%-power resolution diagnostics based on observed paired-state dispersion range from 1.59 pp for Fresh180 H4–HP to 8.61 pp for the $m = 1$ expert–random prefix contrast; these descriptive MDEs do not declare smaller effects absent (Appendix L).

H4 and HP differ in complete continuation identity on 54/180 Fresh180 states, yet only 2/180 have a nonzero empirical $K = 5$ contrast. At $K = 25$, nonzero contrasts remain in 3/17 identity-changed 35B states and 2/20 9B states (Table 44). Under fixed-expert continuation, the full action set contains an outcome boundary in 50/180 states, but H4 and HP straddle it in only 2/180 and 5/180 (Table 41). Thus observable separation also requires a consequential boundary and adequate resolution. Weak-continuation floor compression sends many π_0 branches to failure; strong-controller recovery compression repairs many first-action differences; and low binary execution variance means only 2.1–7.1% of stored identities mix success and failure across five greedy-serving replicas. The third is a replica-level resolution limitation under the frozen serving configuration, not an inherent property of ALFWorld. These are distinct compression modes, not a claim that only 2–5 states have a true value effect.

In Figure 2b, H4–HP changes from $-1.11/-0.44$ pp at B_0 to $-0.44/+0.56$ pp at $B_0 + 20$ for 35B/9B; all intervals cross zero. A fixed expert solves 64/80/104 states at $B_0/B_0 + 20$/150-step-compatible allowance. These are controller-achievable, not optimal, and the new uncached B_0 need not reproduce earlier endpoint masses (Appendix L).

Post-primary Expert–H0 is $+4.44/+2.00$ pp. Of 5,231 admissible actions, 90.22% preserve fixed-expert recovery on 64 baseline-solvable states (Appendix L.5); H0/HP/H4 cross this boundary in only 4/180 and 6/180 states, versus 29/180 for expert/random references (Table 41). These are controller-relative, not Q^{π_0} or optimal-solvability, diagnostics.

Earlier official-state H4–HP ($-0.15/+1.90$ pp; 9B candidate-sensitive) and five-arm H4–H2 (+0.44 pp/+2.34 pp) are not pooled with Fresh180 (Appendices E and L).

5.3 RQ3: Intervention length, budget, and exposure sensitivity

Expert minus random is $+4.90/+11.67/+15.52$ pp at $m = 1/3/5$ for 35B and $+4.79/+11.25/+11.56$ pp for 9B. Only $m = 3, 5$ intervals exclude zero, so this detects longer interventions, not equal one-action sensitivity; random returns rise descriptively without identifying a mechanism.

Neither the earlier weak-control factorial panel nor the strong-control HP/H4 factorial panel provides multiplicity-adjusted interaction evidence (Appendices H and L).

Strong-control one-step/persistent H4–HP is $-1.88/+0.21$ pp (35B) and $-1.04/+2.29$ pp (9B), with interactions crossing zero (Holm $p = 0.576$). Aligned Gemma gives $+1.46/-3.75$ pp and a -5.21 pp interaction, but evidence differs by dependence axis (Holm $p = 0.303$). Profiles are not stable across actors; remaining-episode exposure return is not one-step Q^{π_0}.

6 Implications, Limitations, and Conclusion

Audits should declare their estimand and report identity sharing, saturation, candidate concentration, and controller-relative recoverability. Source-grounded memory text induces substantial ALFWorld rerouting, yet Fresh180 H4–HP is -0.67 pp for both Qwen actors under saturated binary $K = 5$. Graded, longer-horizon, High-K, and fixed-expert checks find no stable H4 advantage, while an up-to-five-action expert intervention is detectable. This supports distinct compression regimes, not a claim of zero utility or a model-size rule.

For deployment policy π, the performance-difference identity writes $J(\pi) - J(\pi_0)$ as discounted $A^{\pi_0}(h_t, a_t)$ over visits by π. Same-state H4–HP is a local advantage difference because $V^{\pi_0}(h)$ cancels. Persistent return additionally requires each exposure policy's occupancies and later actions, which our frozen-state audit does not estimate.

MemLadder first screens rerouting, stability, identity sharing, endpoint mass, and controller-relative recoverability, then reserves K-replica suffixes for supported contrasts. Its compression patterns are limited to ALFWorld and the tested actor policies; generalization to other environments and policies is open.

REPRODUCIBILITY STATEMENT

The authors have institutional approval to release, upon acceptance, the complete experimental code and settings together with the sanitized, project-generated data needed for statistical replay. The release will include the evaluation harness and HTTP adapter; prompts and control-construction code; retrieval and memory-extraction code; environment, model, decoding, and budget configurations; target-state and episode manifests; intervention stimuli; state-level outcomes, continuation identities, analysis weights, and cluster assignments; anonymized item-level human labels and adjudications; statistical, calibration, figure, and table code; exact commands, dependencies, and smoke tests. Sanitized project-generated trajectory records will also be released where upstream terms permit. These materials will support recomputation of the reported estimates and intervals and rerunning the protocol with independently obtained compatible models and environment assets. Because π_0 is defined operationally, reruns under a different serving software/hardware stack may induce a different execution distribution; the release supports protocol and statistical reproduction rather than bitwise reproduction of individual suffix trajectories. Personal annotator identities, credentials, infrastructure addresses and ports, workflow identifiers, and private paths will be removed. Third-party ALFWorld assets and model weights will not be redistributed; the release will instead provide pinned identifiers, notices, hashes where applicable, and official acquisition instructions. The supplementary appendices document the protocols, annotation procedures, inference diagnostics, validation panels, and evidence boundaries.

ETHICS STATEMENT

The study uses a simulated household environment and does not involve personal data or human subjects as experimental participants. Human annotators labeled agent actions; human annotation protocols, adjudication procedures, and reliability statistics are documented in the appendices. The main practical risk is overclaiming causal or general conclusions from a policy-specific audit, which we address through explicit scope and limitations.

AI USE STATEMENT

Generative-AI tools assisted literature discovery, hypothesis refinement, protocol drafting, repository inspection, implementation and testing, data formatting, analysis code, interpretation, figure generation, translation, and manuscript editing. The authors reviewed the protocols, code, outputs, numbers, citations, and text and take responsibility for the final work. Original records were retained, and post-primary analyses and subsequent revisions to interpretation were documented separately.

REFERENCES

Nishant Bhargava and Rodrigo Sobral Barrento. MEMAUDIT: An exact package-oracle evaluation protocol for budgeted long-term LLM memory writing, 2026. URL https://arxiv.org/abs/2605.02199.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29(2):238–249, 2011. doi: 10.1198/jbes.2010.07136.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Marc-Alexandre Cote, Akos Kadar, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. TextWorld: A learning environment for text-based games. In *Computer Games Workshop at ICML/IJCAI*, 2018. URL https://arxiv.org/abs/1806.11532.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. doi: 10.1145/3626772.3657834. URL https://arxiv.org/abs/2401.14887.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. AutoGuide: Automated generation and selection of context-aware guidelines for large language model agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3811. URL https://papers.nips.cc/paper_files/paper/2024/hash/d8efbb5dd415974eb095c3f06bff1f48-Abstract-Conference.html.

Kilem L. Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: 10.1348/000711006X126600.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Cote, and Xingdi Yuan. Interactive fiction games: A colossal adventure. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://arxiv.org/abs/1909.05398.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590. PMLR, 2025. URL https://proceedings.mlr.press/v267/kazemnejad25a.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.11401.

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, 2024a. doi: 10.1162/tacl_a_00638.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating llms as agents. In *International Conference on Learning Representations*, 2024b. URL https://arxiv.org/abs/2308.03688.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Annual Meeting of the Association for Computational Linguistics*, 2022. URL https://arxiv.org/abs/2104.08786.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947. doi: 10.1007/BF02295996.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Empirical Methods in Natural Language Processing*, 2022. URL https://arxiv.org/abs/2202.12837.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory, 2026. URL https://arxiv.org/abs/2509.25140.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023.

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. In *ACM Symposium on User Interface Software and Technology*, 2023. URL https://arxiv.org/abs/2304.03442.

Alessio Rocchi. Does memory credit travel? paired factorial audits of LLM-agent memory. SSRN working paper, 2026. URL https://ssrn.com/abstract=7160321.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures, 2026. URL https://arxiv.org/abs/2606.08275.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Scharli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023. URL https://arxiv.org/abs/2302.00093.

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. URL https://arxiv.org/abs/2303.11366.

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. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

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, 2023a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Cote, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Empirical Methods in Natural Language Processing*, 2022. URL https://arxiv.org/abs/2203.07540.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2203.11171.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63897–63911. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25bx.html.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *North American Chapter of the Association for Computational Linguistics*, 2022. URL https://arxiv.org/abs/2109.01247.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2207.01206.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Haiyue Zhang. Credit without ground truth: Auditing step-level credit assignment in LLM agents against executed replay, 2026. URL https://arxiv.org/abs/2608.19760. Version 2, 25 August 2026.

Yunxiang Zhang, Yiheng Li, Ali Payani, and Lu Wang. AdaMEM: Test-time adaptive memory for language agents. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.05684.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, 2024.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (17), 2024. doi: 10.1609/aaai.v38i17.29936.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2102.09690.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.07863.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory, 2023.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.13854.