# WHEN ERRORS BECOME MEMORIES: CAUSAL PATHWAY TRACING IN MULTI-TURN MEMORY-AUGMENTED LLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long-term memory allows large language models (LLMs) to preserve and reuse information across interaction turns, but it can also transform a localized error into a persistent system-level risk. Existing evaluations of memory-augmented LLMs focus primarily on whether information is correctly stored and retrieved, leaving the causal pathways through which errors persist and re-emerge insufficiently characterized. We introduce a structural causal framework for tracing cross-turn error propagation in memory-augmented LLMs. The framework models user questions, model responses, and memory states as a dynamic causal process, separates memory-update and question-feedback pathways, and uses four counterfactual trajectories to estimate their downstream effects and interaction. Error influence is evaluated through memory retention, natural responses, targeted diagnostic probes, and probability-level error preference. Across models, memory categories, injection positions, and memory mechanisms, error influence generally decreases with interaction distance but remains detectable in internal memory states and controlled behavior after disappearing from natural responses. The memory-update pathway produces more persistent effects on memory-state and probe outcomes, while question-feedback effects can be reactivated by later interactions. Pathway-guided restoration further supports this causal decomposition: repairing the corrupted memory state is more effective than repairing the affected question trajectory alone, and jointly restoring both pathways reduces residual propagation to 0.003, corresponding to a 98.5% reduction relative to the unrepaired condition. These findings show that response-only evaluation can miss latent error traces and that causal pathway tracing provides a principled basis for diagnosing and repairing persistent memory errors. Our code is available at https://anonymous.4open.science/r/Causal-Memory-Pathway-Tracing-33B0.

## 1 INTRODUCTION

As large language models (LLMs) are increasingly deployed in long-term conversational settings, personalized assistants, and persistent interactive systems, external long-term memory has become an important mechanism for preserving and reusing information across sessions (Zhang et al., 2025). Memory systems such as MemoryBank (Zhong et al., 2024), Generative Agents (Park et al., 2023), and MemGPT (Packer et al., 2023) enable models to move beyond the limitations of the context window through continuous memory updates, dynamic retrieval, and hierarchical memory management. Meanwhile, benchmarks such as LongMemEval (Wu et al., 2024), MemBench (Tan et al., 2025), and MemoryAgentBench (Hu et al., 2025) evaluate memory systems in terms of information retrieval, knowledge updating, long-range understanding, selective forgetting, and overall memory capability.

Long-term memory, however, also changes the nature of model errors. In conventional single-turn interactions, an incorrect response typically constitutes a localized output failure. In memory-augmented systems, by contrast, the same error may be incorporated into subsequent memory states and repeatedly retrieved and reused, thereby influencing responses across future interactions (Xiong et al., 2025; Zhang et al., 2023; Kwan et al., 2024). In high-stakes settings that rely on persistent information, such as health management, financial planning, and task execution, these enduring effects



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.

may further shape user decisions or agent actions (Dash et al., 2026; Qian, 2026). Prior studies have shown that erroneous or contaminated information may persist once incorporated into long-term memory and continue to affect subsequent responses, plans, and behaviors when retrieved (Xiong et al., 2025; Deng et al., 2026). Consequently, an initially isolated error may propagate over time, evolving into a long-term risk that affects both system states and real-world outcomes.

Existing research has primarily focused on the construction, management, and evaluation of long-term memory, typically assessing whether a model can correctly retain and use historical information through final-answer quality or aggregate task performance (Wu et al., 2024; Tan et al., 2025; Hu et al., 2025). However, long-term memory is not just a static storage component; it is part of a dynamic system where user questions, model responses, and memory states evolve together. At the system level, an error may initially affect future behavior through both internal memory states and external interaction dynamics. Internally, an erroneous response may be written into long-term memory, affecting subsequent generation (Xiong et al., 2025; Zhang et al., 2026; Deng et al., 2026). Externally, it may reshape the user's follow-up questions, causing the conversation to proceed from an incorrect premise (Kwan et al., 2024). These pathways may later interact, prolonging the influence of the original error. Existing approaches provide limited insight into how errors propagate dynamically across responses, memory states, and subsequent questions, and how the pathway through which an error enters the system shapes its downstream influence.

To address this gap, we introduce a structural causal model (SCM)-based framework (Pearl, 2009) for analyzing cross-turn error propagation dynamics. The framework integrates the joint evolution of user questions, model responses, and memory states within a unified causal formulation and employs counterfactual interventions to control whether an error enters through internal memory, external question feedback, or both. We use *error influence* to refer to the measurable downstream effects of an injected error on memory states and subsequent behavior, including memory retention, occurrence in natural responses, re-elicitation under targeted probes, and shifts in generation probabilities. By tracing these manifestations, our framework reveals how errors persist, decay, and re-emerge across long-term interactions, providing a unified perspective for understanding and mitigating systematic error propagation in memory-augmented LLMs. Our main contributions are as follows:

- We develop a dynamic structural causal model and a four-trajectory counterfactual intervention framework that estimates the downstream effects of introducing an error through memory updating and question feedback, along with their interaction.
- Our experiments show that error influence generally decreases with interaction distance. In the primary turn-3 setting, the internal memory pathway has more persistent effects on memory-state and controlled-probe outcomes, while the balance between memory and question pathways for natural responses depends on the injection round. Even after an error disappears from natural responses, it may remain as a latent memory trace or a shift in generation preference.
- Error propagation varies across memory categories and mechanisms, showing how memory organization shapes error persistence. In a Qwen2.5-7B-Instruct MemoryBank restoration experiment, restoring the corrupted memory state suppresses error propagation more effectively than correcting the error-affected question trajectory alone, while joint restoration gets a 98.5% reduction relative to the unrepaired condition.

## 2 Related Work

**Long-Term Memory-Augmented LLMs.** External memory supports long-term dialogue and information reuse. Zhong et al. (2024) introduce MemoryBank, which continuously updates user memories using recency and importance; Park et al. (2023) present Generative Agents, which combine retrieval and reflection for planning; Packer et al. (2023) propose MemGPT for managing active context and external storage; and Wang et al. (2023) develop LongMem to decouple memory encoding, retrieval, and reading. These works improve memory storage and use, but pay less attention to how stored information propagates across interaction rounds.

**Memory Evaluation in LLM Systems.** Wu et al. (2024) introduce LongMemEval for assessing historical information retrieval, cross-session and temporal reasoning, knowledge updating, and ab-

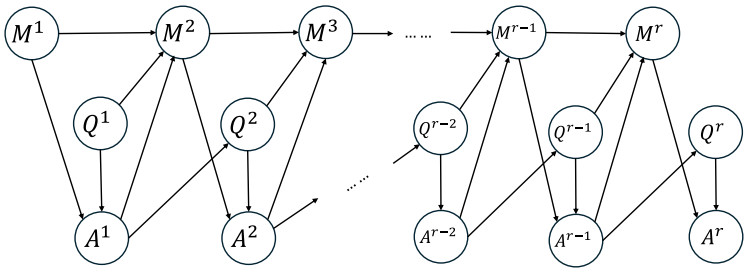



Figure 1: Dynamic structural causal graph of a long-term memory-augmented LLM agent. The memory-update path $A^{r-1} \to M^r \to A^r$ and question-feedback path $A^{r-1} \to Q^r \to A^r$ represent two mechanisms through which an erroneous response can affect subsequent interactions.

stention. Tan et al. (2025) propose MemBench to evaluate memory across types, interaction modes, effectiveness, efficiency, and capacity, while Hu et al. (2025) develop MemoryAgentBench for testing retrieval, test-time learning, long-range understanding, and selective forgetting in incremental interactions. These benchmarks extend memory evaluation to interactive settings but mainly rely on final-answer quality or aggregate performance rather than tracing downstream error propagation.

**Causal Analysis of LLM Behavior.** Meng et al. (2022) use causal tracing to identify computations that contribute to factual predictions and introduce ROME. Activation patching and interchange interventions (Zhang & Nanda, 2023) manipulate model components, while causal mediation analysis decomposes effects transmitted through intermediate representations (Vig et al., 2020). Recent work tracks the changing influence of questions and generated prefixes during autoregressive generation (Xiao et al., 2026). These approaches largely examine individual generations, leaving the causal pathways of cross-turn error propagation among questions, responses, and memory states underexplored.

## 3 Structural Causal Modeling

We model the interaction and memory-update processes of a long-term memory-augmented LLM agent as a structural causal model indexed by interaction round (Pearl, 2009). As shown in Figure 1, $Q^r$, $A^r$, and $M^r$ denote the question, generated response, and memory state available before response generation at round $r$, respectively. We adopt a first-order Markov assumption: earlier interactions may influence round $r$ through intermediate states, while the variables at round $r$ depend only on the preceding and current rounds. This abstraction captures question evolution, response generation, and memory updating.

The graph contains four types of relationships. **(1) Memory persistence:** $M^{r-1} \to M^r$ captures the retention of stored information across rounds (Zhong et al., 2024; Packer et al., 2023). **(2) Memory update:** $Q^{r-1} \to M^r$ and $A^{r-1} \to M^r$ represent how the preceding question and response shape the next memory state, including the possibility that a model response is written into memory (Zhong et al., 2024; Packer et al., 2023; Xiong et al., 2025). **(3) Question continuation:** $A^{r-1} \to Q^r$ represents how a previous response shapes the user's next question, such as a follow-up or request for clarification (Kwan et al., 2024; Laban et al., 2025). **(4) Response generation:** $Q^r \to A^r$ and $M^r \to A^r$ indicate that the current response depends on both the question and the available memory (Zhong et al., 2024).

When $A^{r-1}$ is erroneous, the graph identifies two pathways through which the error can initially affect subsequent interactions. The **memory-update pathway**, $A^{r-1} \to M^r \to A^r$, captures erroneous information being written into the agent's internal memory and subsequently biasing response generation (Xiong et al., 2025; Zhang et al., 2026; Deng et al., 2026). The **question-feedback pathway**, $A^{r-1} \to Q^r \to A^r$, captures the user continuing from an incorrect premise introduced by the previous response, thereby changing the dialogue trajectory (Laban et al., 2025).

These pathways differ in where the error enters the next round: one changes the internal memory state, while the other changes the external question. They may subsequently interact as error-affected

questions elicit further erroneous responses that are written into memory. We therefore distinguish the pathways at the intervention turn and trace how each affects later interactions.

## 4 Causal Intervention and Pathway Effects

Let $t$ denote the error-injection round. Following the two outgoing causal edges from the injected response $A^t$ to the next-round memory state $M^{t+1}$ and user question $Q^{t+1}$ in Figure 1, we intervene on the inputs to the memory updater and the next-question generator. These two successor mechanisms define a $2 \times 2$ joint intervention design.

At round $t$, we replace the naturally generated response with a matched correct response $A_c^t$ or erroneous response $A_e^t$ to the same question. The two responses are constructed to differ in the target proposition while preserving the remaining content and wording as closely as possible; candidate pairs are validated before intervention. Passing them separately to the memory updater produces a clean memory state $M_c^{t+1}$ and an erroneous memory state $M_e^{t+1}$, respectively. Combining the correct and erroneous response variants as inputs to the two mechanisms yields four counterfactual trajectories (Shpitser & Tchetgen, 2014):

$$\mathcal{T}_{cc} = \left(M_c^{t+1}, A_c^t\right), \mathcal{T}_{ec} = \left(M_e^{t+1}, A_c^t\right), \mathcal{T}_{ce} \quad = \left(M_c^{t+1}, A_e^t\right), \mathcal{T}_{ee} = \left(M_e^{t+1}, A_e^t\right). \tag{1}$$

The first component specifies the memory state available at round $t+1$, while the second specifies the response supplied to the next-question generator. Accordingly, the first subscript denotes the intervention on the memory-update mechanism, and the second denotes the intervention on the question-feedback mechanism; $c$ and $e$ indicate correct and erroneous inputs, respectively. Thus, $\mathcal{T}_{cc}$ is the fully clean trajectory, $\mathcal{T}_{ec}$ introduces the error through memory updating only, $\mathcal{T}_{ce}$ introduces it through question feedback only, and $\mathcal{T}_{ee}$ introduces it through both mechanisms.

After the intervention at round $t$, the four trajectories evolve independently according to their respective memory states, dialogue histories, and previous responses. Their differences therefore estimate the downstream contrasts induced by the assigned error-entry interventions, while subsequent memory updating, question evolution, and response generation unfold naturally within each trajectory.

### 4.1 Error-Propagation Outcomes

Let $k$ denote the interaction distance from the error-injection round, where $k = 1$ corresponds to the first post-intervention round. We measure the downstream influence of the injected error at four complementary levels: memory retention, natural responses, targeted diagnostic probing, and generation preference. Within a given trajectory, let $M(k)$ denote the memory state available immediately before response generation at interaction distance $k$. The memory-state outcome is defined as

$$Y_M(k) = \mathbb{I}\left[\text{the target error is detected in } M(k)\right], \tag{2}$$

where $Y_M(k) = 1$ indicates that the target error remains in the current memory state. The natural-response outcome is defined as

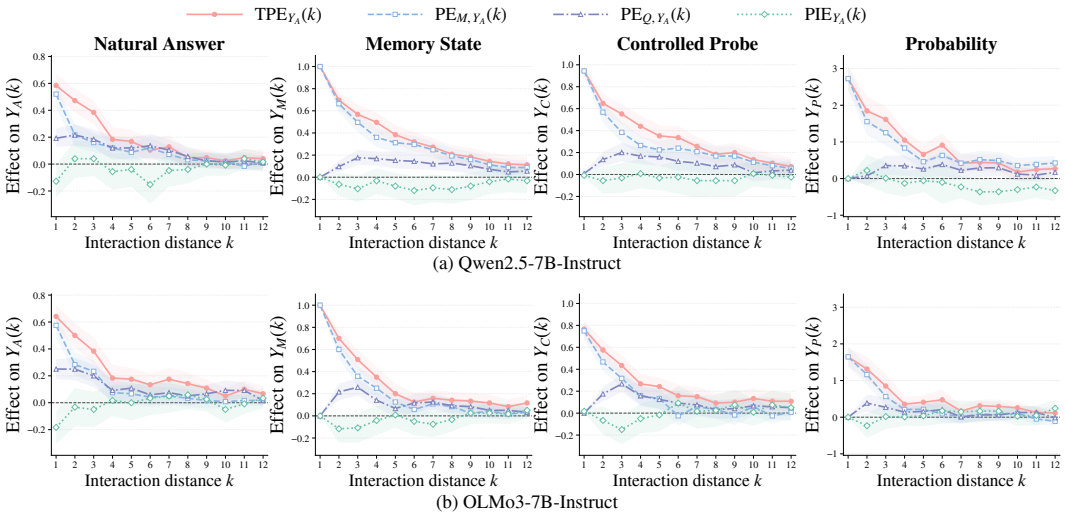
$$Y_A(k) = \mathbb{I}\left[\text{the natural response contains the target error}\right]. \tag{3}$$

Here, $Y_A(k) = 1$ when the response explicitly states, necessarily assumes, or directly relies on the target error.

Because natural follow-up questions may not directly concern the target fact, we also use a targeted diagnostic question $q_k$ that asks about it directly. At each interaction distance $k$, the same probe $q_k$ is used across all four trajectories to ensure comparability. The controlled-probe outcome is defined as

$$Y_C(k) = \mathbb{I}\left[\text{the response to } q_k \text{ contains the target error}\right], \tag{4}$$

where $Y_C(k) = 1$ indicates that the target error can be directly elicited from the current trajectory state. The outcomes $Y_M(k)$, $Y_A(k)$, and $Y_C(k)$ are determined by a proposition-level evaluator. A text is labeled erroneous only if it explicitly states, necessarily assumes, or directly relies on the complete target error; merely quoting, negating, or correcting the erroneous proposition does not count as propagation. Thus, $Y_A$ measures end-to-end behavioral propagation, including pathway-induced changes in the natural question, whereas $Y_C$ measures state-conditioned elicitation under a common diagnostic question.




(a) Qwen2.5-7B-Instruct

(b) OLMo3-7B-Instruct

Figure 2: Pathway effects of an injected error across interaction distances for two models. The first and second rows show results for Qwen2.5-7B-Instruct and OLMo3-7B-Instruct, respectively. From left to right, the columns correspond to $Y_A(k)$, $Y_M(k)$, $Y_C(k)$, and $Y_P(k)$. The curves show the total propagation effect TPE, memory-update pathway effect $\mathrm{PE}_M$, question-feedback pathway effect $\mathrm{PE}_Q$, and pathway interaction effect PIE; shaded regions indicate 95% bootstrap confidence intervals.

To capture latent shifts that may not appear in the final decoded response, we construct a matched erroneous target answer $a_{\mathrm{err}}$ and correct target answer $a_{\mathrm{cor}}$ for $q_k$, and define the probability-level error preference as

$$Y_P(k) = \frac{1}{\ln 2}\left[\frac{\log P(a_{\mathrm{err}} \mid q_k, M(k))}{|a_{\mathrm{err}}|} - \frac{\log P(a_{\mathrm{cor}} \mid q_k, M(k))}{|a_{\mathrm{cor}}|}\right], \tag{5}$$

where $|a|$ denotes the number of tokens in answer $a$, and the factor $1/\ln 2$ converts the difference in natural log-likelihoods into bits. A positive $Y_P(k)$ indicates a preference for the erroneous target, whereas a negative value indicates a preference for the correct target. The four outcomes capture, respectively, error retention in memory, explicit error occurrence in natural responses, error re-elicitation through targeted probing, and latent shifts in generation preference.

### 4.2 Pathway Effects

Motivated by causal contrasts for pathway and joint interventions (Avin et al., 2005; Shpitser & Tchetgen, 2014), we define trajectory-level effects according to the mechanism through which the error initially enters the system. For any outcome $Y \in \{Y_M, Y_A, Y_C, Y_P\}$, let $Y_{xz}(k)$ denote its value under trajectory $\mathcal{T}_{xz}$ at interaction distance $k$, where $x, z \in \{c, e\}$, and let $\mathbb{E}[\cdot]$ denote the average over experimental instances.

**Definition 1** (**Total Propagation Effect**). *The total propagation effect at interaction distance $k$ is defined as*

$$\mathrm{TPE}_Y(k) = \mathbb{E}[Y_{ee}(k)] - \mathbb{E}[Y_{cc}(k)]. \tag{6}$$

It measures the overall downstream effect of introducing the error through both entry mechanisms relative to the fully clean trajectory.

**Definition 2** (**Memory-Update Pathway Effect**). *The memory-update pathway effect at interaction distance $k$ is defined as*

$$\mathrm{PE}_{M,Y}(k) = \mathbb{E}[Y_{ec}(k)] - \mathbb{E}[Y_{cc}(k)]. \tag{7}$$

It measures the downstream effect associated with introducing the error only through the memory updater.

**Definition 3** (**Question-Feedback Pathway Effect**). *The question-feedback pathway effect at interaction distance $k$ is defined as*

$$\mathrm{PE}_{Q,Y}(k) = \mathbb{E}[Y_{ce}(k)] - \mathbb{E}[Y_{cc}(k)]. \tag{8}$$

It measures the downstream effect associated with introducing the error only through the next-question generator.

**Definition 4** (**Pathway Interaction Effect**). *The pathway interaction effect is defined as*

$$\begin{aligned}\mathrm{PIE}_Y(k) &= \mathrm{TPE}_Y(k) - \mathrm{PE}_{M,Y}(k) - \mathrm{PE}_{Q,Y}(k) \\ &= \mathbb{E}[Y_{ee}(k)] - \mathbb{E}[Y_{ec}(k)] - \mathbb{E}[Y_{ce}(k)] + \mathbb{E}[Y_{cc}(k)].\end{aligned} \tag{9}$$

These estimands quantify downstream effects associated with the pathway through which the error is introduced at the injection turn, while each trajectory subsequently evolves according to its own states and dialogue history. The interaction effect measures non-additivity between the two pathway interventions, with positive and negative values indicating super-additive and sub-additive effects, respectively. The estimands are defined by these direct mechanism interventions rather than inferred from observational correlations. For example, $\mathrm{TPE}_{Y_M}(k)$, $\mathrm{PE}_{M,Y_A}(k)$, and $\mathrm{PE}_{Q,Y_C}(k)$ denote the total effect on memory retention, the memory-update pathway effect on natural responses, and the question-feedback pathway effect on controlled elicitation, respectively. All trajectories share the same pre-intervention history, and the same diagnostic question is used across all four trajectories for controlled-probe and probability-level evaluation.

## 5 Experimental Evaluation

**Dataset Construction.** Existing multi-turn dialogue datasets typically use fixed interaction trajectories and do not support the pathway-specific interventions required by our causal design. We therefore construct 144 controlled synthetic multi-turn interactions across six memory categories, with 24 instances each. Follow-up questions use fixed temperature, top-$p$, and decoding settings. The primary experiment injects one target error at turn 3 and traces the following 12 turns; a position-sensitivity experiment repeats the same protocol with injection at $t \in \{3, 6, 9, 12\}$. Thus, the available post-injection distances are $k =$ 1–12, 1–9, 1–6, and 1–3, respectively. Unless otherwise noted, main-text results use the primary $t = 3$ setting. The complete supplementary record is provided in Appendix A, with construction, validity filtering, and run manifests summarized in Appendix A.1.

**Memory Mechanisms and Models.** The main setting follows a MemoryBank-style mechanism (Zhong et al., 2024) where the memory state is updated after each interaction round. We additionally evaluate an episodic-retrieval variant inspired by prior retrieval-based memory systems (Park et al., 2023; Packer et al., 2023). This variant stores each interaction as an independent memory entry and retrieves the three most relevant entries using a deterministic, length-normalized lexical-overlap score, with recency used to break ties. We use Qwen2.5-7B-Instruct (Qwen et al., 2025) and OLMo3-7B-Instruct (Olmo et al., 2026) as answer models, while Qwen2.5-32B-Instruct (Qwen et al., 2025) is used for question simulation, counterfactual construction, and proposition-level evaluation.

**Experimental Settings.** Natural responses, memory updates, controlled-probe questions and responses, and proposition-level evaluator outputs use greedy decoding. Subsequent questions are sampled with temperature 0.7 and top-$p = 0.9$; failed answer-pair validations are retried with temperature 0.7 and top-$p = 0.95$, for at most five attempts. Maximum generation lengths are 96, 128, 256, 256, and 192 tokens for questions, answers, memory updates, answer pairs, and evaluator outputs, respectively.

To test protocol sensitivity, we repeated the turn-3 experiment with an independent seed (20261001) and compared Qwen2.5-7B-Instruct answer limits of 128 and 256 new tokens. We also moved the intervention to turns $t \in \{3, 6, 9, 12\}$ and compared the positions on their common valid instances. These analyses preserve the same four-trajectory construction and outcome definitions as the primary experiment.

**Valid Instances and Statistical Reporting.** We retain instances where the correct response yields a clean memory state and the erroneous response writes the target error, yielding 125 Qwen and 120 OLMo instances in the primary cohort. This condition fixes the first post-injection memory-state effect; persistence is assessed from later rounds. Each valid instance is executed under each of the four counterfactual trajectories, and reported effects are averaged over instances. Shaded regions denote 95% percentile bootstrap confidence intervals computed from 5,000 instance-level resamples. All experiments use 8 NVIDIA H200 GPUs.

The validity filter is applied separately to each sensitivity run. In the independent-seed length check, the valid counts were 120/144 for Qwen at the 128-token limit, 117/144 for Qwen at the 256-token limit, and 120/144 for OLMo at the 128-token limit. Position comparisons use only instances valid at all compared positions; position-specific cohorts are reported in Appendix A.2, with the seed and length checks in Appendix A.4. Record-level integrity checks are reported in Appendix A.8.

### 5.1 Error Propagation and Pathway Effects

We assess propagation through four outcomes defined in Section 4.1: natural responses $Y_A$, memory states $Y_M$, controlled probes $Y_C$, and probability-level error preference $Y_P$. For cross-model robustness, we use Qwen2.5-7B-Instruct and OLMo3-7B-Instruct with identical data, memory, interventions, and evaluation.

Figure 2 shows strong short-range effects that gradually decay across all outcomes and both models. Natural-response effects decay fastest: from $0.584$ and $0.642$ at $k = 1$ to $0.040$ and $0.067$ at $k = 12$ for Qwen2.5-7B-Instruct and OLMo3-7B-Instruct, respectively. Memory-state and controlled-probe effects persist longer. At $k = 1$, the controlled-probe effect reaches $0.944$ and $0.767$. The memory-state total propagation effect is $1.0$ at this distance by construction of the validity filter; over $k = 2$–$12$, its mean total effect remains $0.319$ and $0.239$, respectively.

The finer-grained $Y_P$ reveals substantial shifts toward the erroneous target at $k = 1$: $2.726$ for Qwen2.5-7B-Instruct and $1.646$ for OLMo3-7B-Instruct. Errors can therefore vanish from natural responses yet remain in memory, surface under probing, or influence generation probabilities.

In the primary turn-3 setting, the memory-update pathway contributes more to $Y_M$, $Y_C$, and $Y_P$, whereas question feedback becomes more influential when later questions reactivate error-related information. The mostly negative or near-zero interaction effects indicate limited additional influence once either pathway has induced the error. These patterns hold across models, although controlled-probe and probability-level traces last longer in Qwen2.5-7B-Instruct and attenuate faster in OLMo3-7B-Instruct.

### 5.2 Sensitivity to Injection Position and Generation Length

Propagation magnitude and pathway balance vary across injection positions. On common four-position cohorts (84 Qwen and 68 OLMo instances valid at all four injection positions), Qwen's $Y_A$ TPE averaged over $k = 1$–$3$ is $0.468$, $0.246$, $0.456$, and $0.615$ for $t = 3, 6, 9, 12$, while OLMo's values are $0.520$, $0.338$, $0.574$, and $0.608$. Across positions, $Y_C$ TPE ranges from $0.687$ to $0.817$ for Qwen and from $0.618$ to $0.824$ for OLMo. For $Y_C$, the memory-update pathway effect exceeds the question-feedback effect at every tested position, whereas the relative contributions to $Y_A$ shift toward question feedback at later positions. Full position-specific curves, confidence intervals, and pathway contrasts are reported in Appendix A.2.

The independent-seed and answer-length checks leave the primary natural-response effects essentially unchanged. Increasing Qwen's answer limit from 128 to 256 tokens reduces the answer-generation touch-or-exceed rate from $41.3\%$ to $23.4\%$; the paired $k = 1$–$12$ natural-response $Y_A$ TPE difference is $-0.006$ $[-0.061, 0.047]$. The new seed changes the original $k = 1$–$12$ natural-response $Y_A$ TPE by $-0.004$ $[-0.063, 0.057]$ for Qwen and $-0.025$ $[-0.081, 0.031]$ for OLMo. Full length summaries and seed comparisons are provided in Appendix A.4. Since generation finish reasons are unavailable, these rates are interpreted as length diagnostics rather than direct truncation counts.

**Evaluator robustness.** We re-evaluate all $35{,}280$ proposition-level outcomes with GLM-5.2 (GLM-5-Team et al., 2026) and a stratified sample of 900 outcomes with GPT-5.5 (OpenAI,

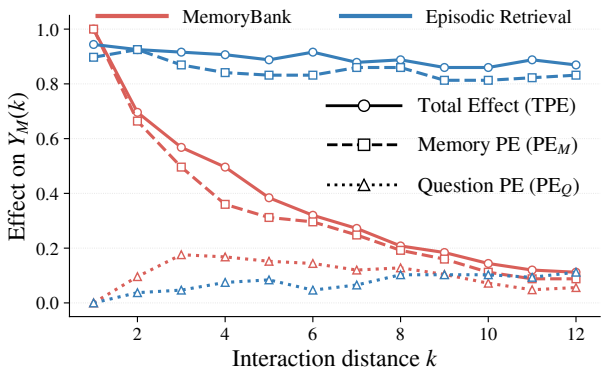



Figure 3: Effects of the injected error on the memory-state outcome $Y_M(k)$ under different memory mechanisms. Red and blue curves denote MemoryBank and episodic retrieval, respectively. The outcome uses the full pre-response memory; for episodic retrieval, this includes all stored entries, not only retrieved entries. Line styles indicate TPE, $\mathrm{PE}_M$, and $\mathrm{PE}_Q$.

2026). Agreement statistics and Cohen's $\kappa$ values are reported in Appendix A.5; the re-judged results preserve the direction of all 72 TPE estimates and the temporal and pathway-level conclusions.

### 5.3 Propagation Patterns across Memory Types

We further compare error propagation across six memory categories: *Constraint*, *Decision*, *Plan*, *Preference*, *Profile*, and *Routine*. To reduce variation caused by natural follow-up questions, we use a fixed diagnostic question that directly probes the target fact.

Category-specific results are summarized in the main text, while Appendix A.3 reports the corresponding cohort counts and expanded-cohort effects. In the primary cohort, preference errors have the highest controlled-probe total effect in Qwen2.5-7B-Instruct, while plan errors have the highest point estimate in OLMo3-7B-Instruct. In Qwen, preference errors receive more contribution from memory updating, whereas plan and routine errors involve more question feedback; in OLMo, both pathways contribute to plan errors. Constraint, decision, and profile effects generally decay faster or fluctuate more. Thus, category shapes persistence and pathway composition.

For the expanded turn-3 category analysis, 84 independently constructed candidates yielded 72 additional Qwen and 65 OLMo instances, increasing the cohorts to 197 and 185, respectively. The combined cohorts retain positive propagation and decay; preference remains largest in Qwen, while preference and plan are nearly tied in OLMo. Detailed cohort counts, category intervals, and pairwise comparisons are reported in Appendix A.3.

### 5.4 Propagation Patterns across Memory Mechanisms

To examine whether the identified propagation patterns generalize across memory mechanisms, we compare two long-term memory architectures: continuously updated MemoryBank (Zhong et al., 2024) and episodic retrieval with independent entries and Top-3 retrieval (Park et al., 2023; Packer et al., 2023). MemoryBank consolidates historical interactions into a dynamic memory state, whereas episodic retrieval preserves interactions as independent entries and retrieves those most relevant to the current query.

Figure 3 shows that TPE and $\mathrm{PE}_M$ decline as MemoryBank updates its state, whereas erroneous episodic entries persist in full storage. On the episodic memory-only trajectory, the target error appears in 86.9% of full-storage observations but 18.4% of controlled-probe answers. In both mechanisms, $\mathrm{PE}_M$ dominates the memory-state effect and $\mathrm{PE}_Q$ is smaller. This separates memory retention from downstream use; the storage and retrieval definitions are given in Appendix A.6.

| Strategy | Residual MRE ↓ | Reduction ↑ |
|---|---|---|
| *No Repair* | 0.178 | – |
| *Question Repair* | 0.129 | 27.7% |
| *Memory Repair* | 0.053 | 70.0% |
| *Joint Repair* | **0.003** | **98.5%** |

Table 1: Pathway-guided state-restoration results on controlled probes. Lower mean residual effect (MRE) indicates more effective restoration.

### 5.5 Can Pathway Effects Guide Error Restoration?

To examine whether pathway effects can guide effective interventions, we conduct a pathway-guided state-restoration experiment. The error is injected at turn 3, and a one-time restoration is applied before the turn at interaction distance $k = 3$. We then evaluate the remaining propagation over subsequent interactions. We conduct this experiment using Qwen2.5-7B-Instruct with the MemoryBank-style mechanism. We consider four strategies: *No Repair*, *Question Repair*, *Memory Repair*, and *Joint Repair*. *Question Repair* restores the dialogue history and previous response while retaining the error-affected memory. *Memory Repair* restores the memory state while retaining the error-affected dialogue context. *Joint Repair* restores the complete clean state, including memory, dialogue history, previous response, and current question, and serves as a reference condition.

Using fixed controlled probes, we define the residual effect of strategy $s$ at post-restoration distance $j$ as $R_s(j) = |Y_s(j) - Y_{\text{clean}}(j)|$, where $Y_s(j)$ and $Y_{\text{clean}}(j)$ denote the controlled-probe outcomes under strategy $s$ and the clean trajectory, respectively. We summarize the remaining propagation using the mean residual effect: $\text{MRE}(s) = \frac{1}{J}\sum_{j=0}^{J-1} R_s(j)$, where $J$ is the number of post-restoration distances. Lower MRE indicates more effective restoration. We also report the relative reduction compared with *No Repair*:

$$\text{Reduction}(s) = \frac{\text{MRE}(\textit{No Repair}) - \text{MRE}(s)}{\text{MRE}(\textit{No Repair})} \times 100\%.$$

As shown in Table 1, *Memory Repair* lowers residual MRE more than *Question Repair* (paired difference $-0.075$, 95% CI $[-0.125, -0.030]$), consistent with the stronger memory-update pathway effects. Across instances, larger question-feedback effects are associated with greater gains from *Question Repair* on controlled probes ($\rho = 0.335$, 95% CI $[0.158, 0.493]$). *Joint Repair* restores the complete clean state, and its 0.003 MRE arises from inconsistent evaluator labels for otherwise identical outputs. State-reset semantics and paired summaries are given in Appendix A.7.

## 6 Conclusion

We present a dynamic structural causal framework for characterizing cross-turn error propagation in long-term memory-augmented LLMs and decomposing its total effect into memory-update, question-feedback, and interaction components. The four-trajectory intervention separates internal memory contamination from externally propagated question feedback, effects that aggregate final-answer metrics otherwise conflate. Experiments across models, memory categories, and memory mechanisms show that errors can persist across multiple turns, with the memory-update pathway typically contributing more strongly to memory-state and controlled-probe outcomes in the primary turn-3 setting. Injection-position and generation-length analyses show that the qualitative temporal pattern is preserved across protocol variations, while effect magnitude and pathway balance vary with the injection round and sampling realization. The expanded turn-3 cohort reproduces these effects with 197 valid Qwen instances and 185 valid OLMo instances. Pathway-guided restoration further shows that repairing the memory-update pathway is more effective than repairing the error-affected question-feedback pathway alone, while restoration to the complete clean state provides a reference for residual propagation.

### AI USE STATEMENT

In this work, we used generative AI tools only for language polishing, including improving grammar, clarity, and fluency of the manuscript. We have not used generative AI tools for generating research ideas, developing the methodology, designing experiments, analyzing results, or drawing scientific conclusions; other disclosure categories are not applicable to this work. We reviewed all AI-assisted edits and verified that they did not alter the technical content or scientific claims. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

## REFERENCES

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. 2005.

Pritam Dash, Tongyu Ge, Aditi Jain, Tanmay Shah, and Zhiwei Shang. From untrusted input to trusted memory: A systematic study of memory poisoning attacks in llm agents, 2026. URL https://arxiv.org/abs/2606.04329.

Xinle Deng, Ruobin Zhong, Hujin Peng, Xiaoben Lu, Yanzhe Wu, Guang Li, Buqiang Xu, Yunzhi Yao, Jizhan Fang, Haoliang Cao, et al. Memtrace: Tracing and attributing errors in large language model memory systems. *arXiv preprint arXiv:2605.28732*, 2026.

GLM-5-Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*, 2025.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, 2024.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in neural information processing systems*, 2022.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

OpenAI. GPT-5.5 model. https://developers.openai.com/api/docs/models/gpt-5.5, 2026.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: towards llms as operating systems. 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Jiachen Qian. Visual inception: Compromising long-term planning in agentic recommenders via multimodal memory poisoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20846–20862, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.954. URL https://aclanthology.org/2026.acl-long.954/.

Qwen, 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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Ilya Shpitser and Eric Tchetgen Tchetgen. Causal inference with a graphical hierarchy of interventions. *arXiv preprint arXiv:1411.2127*, 2014.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Shuyao Xiao, Shengling Wang, and Ke Chao. Bridging internal consistency and external alignment: A causal and dynamic interpretability framework for llm generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20378–20392, 2026.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts llm agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*, 2025.

Dylan Zhang, Yanshan Lin, Zhengkun Wu, Yihang Sun, Bingxuan Li, Dianqi Li, and Hao Peng. Useful memories become faulty when continuously updated by llms. *arXiv preprint arXiv:2605.12978*, 2026.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47, 2025.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

# A Supplementary Experimental Details

This appendix gives the validity counts, position-specific estimates, expanded-cohort analyses, generation-length checks, evaluator comparisons, and mechanism-level storage statistics used in the main text. All confidence intervals are instance-level bootstrap intervals; the bootstrap unit is the instance or `group_id`, never an individual turn.

## A.1 Construction, Validity, and Reproducibility

The primary dataset contains 144 controlled synthetic interactions, with 24 candidates in each of six memory categories. Each interaction has 15 rounds. The primary $t = 3$ analysis retains 125 Qwen instances and 120 OLMo instances after requiring a clean memory state for the correct response and successful writing of the target error for the erroneous response. Invalid injections are retained in the manifest with their failure reason and are not replaced by relaxed criteria. The primary and position-sensitivity analyses use 5,000 bootstrap resamples; the expansion and seed-length analyses use 10,000 resamples. Candidate and valid-instance counts for every injection position are reported in Table 2.

| Model | $t$ | Candidates | Valid | Invalid | Valid rate |
|---|---|---|---|---|---|
| Qwen | 3 | 144 | 125 | 19 | 86.8% |
| Qwen | 6 | 144 | 107 | 37 | 74.3% |
| Qwen | 9 | 144 | 124 | 20 | 86.1% |
| Qwen | 12 | 144 | 118 | 26 | 81.9% |
| OLMo | 3 | 144 | 120 | 24 | 83.3% |
| OLMo | 6 | 144 | 115 | 29 | 79.9% |
| OLMo | 9 | 144 | 116 | 28 | 80.6% |
| OLMo | 12 | 144 | 103 | 41 | 71.5% |

Table 2: Candidate and valid-instance counts for the four injection positions. The main position comparison uses only the 84 Qwen and 68 OLMo instances valid at all four positions.

The position runs use seed 20260925 for the newly generated positions; the original $t = 3$ records are reused. The legacy $t = 3$ records do not expose an explicit random seed, so the position comparison changes the intervention position under the same construction and decoding protocol but does not claim identical random streams. The independent-seed check in Appendix A.4 quantifies this remaining sampling variation.

## A.2 Injection-Position Sensitivity

Figures 4 and 5 show the complete position-specific curves. The main comparison averages $k = 1, 2, 3$ within each instance before bootstrapping. The $t = 3, 6, 9$ runs additionally share a common $k = 1, \ldots, 6$ window; the original $t = 3$ run retains the full $k = 1, \ldots, 12$ curve. The paired position summaries and pathway contrasts are collected in Table 3.

| Model | $t$ | $Y_A$ TPE | $Y_C$ TPE | $Y_A$ $\mathrm{PE}_M - \mathrm{PE}_Q$ | $Y_C$ $\mathrm{PE}_M - \mathrm{PE}_Q$ |
|---|---|---|---|---|---|
| Qwen | 3 | .468 [.381,.552] | .687 [.607,.758] | .111 [.016,.198] | .516 [.433,.595] |
| Qwen | 6 | .246 [.151,.341] | .607 [.520,.690] | .020 [-.083,.123] | .456 [.377,.540] |
| Qwen | 9 | .456 [.385,.528] | .746 [.671,.813] | -.202 [-.290,-.111] | .429 [.341,.512] |
| Qwen | 12 | .615 [.544,.683] | .817 [.758,.873] | -.123 [-.226,-.020] | .536 [.448,.623] |
| OLMo | 3 | .520 [.431,.608] | .618 [.539,.691] | .157 [.049,.265] | .382 [.299,.471] |
| OLMo | 6 | .338 [.211,.456] | .623 [.539,.706] | .196 [.093,.304] | .402 [.299,.495] |
| OLMo | 9 | .574 [.495,.652] | .765 [.696,.829] | -.103 [-.221,.015] | .353 [.260,.446] |
| OLMo | 12 | .608 [.529,.686] | .824 [.760,.887] | -.250 [-.353,-.142] | .221 [.137,.309] |

Table 3: Position-specific effects averaged over $k = 1$–3 on the four-position common cohorts. Positive pathway differences indicate a larger memory-update contribution.

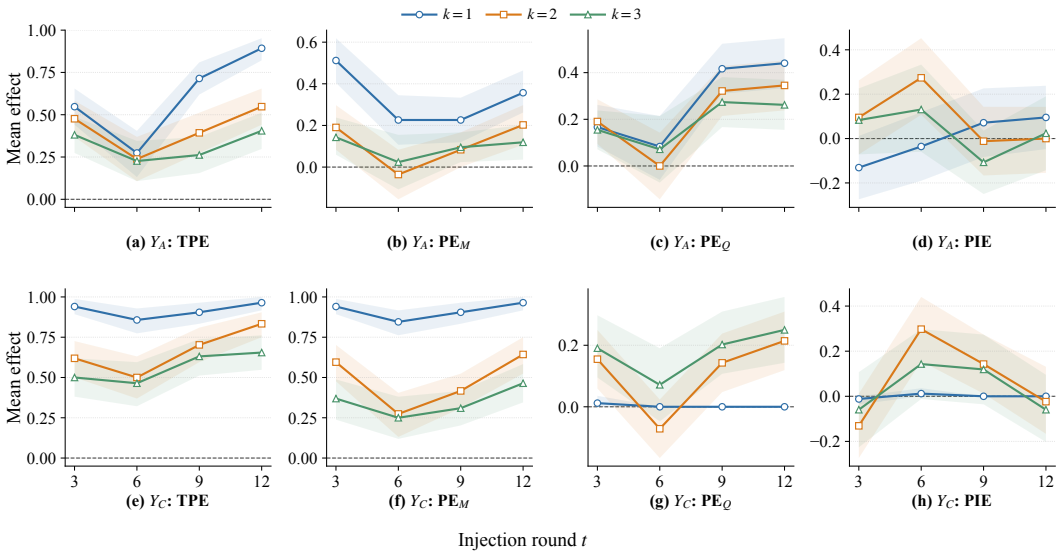



Figure 4: Position sensitivity for Qwen2.5-7B-Instruct on the common valid instances. Curves report the four outcomes and path effects for $t \in \{3, 6, 9, 12\}$; here, $k$ is the interaction distance, i.e., the number of subsequent turns after injection. Shaded regions are 95% instance-level bootstrap intervals.

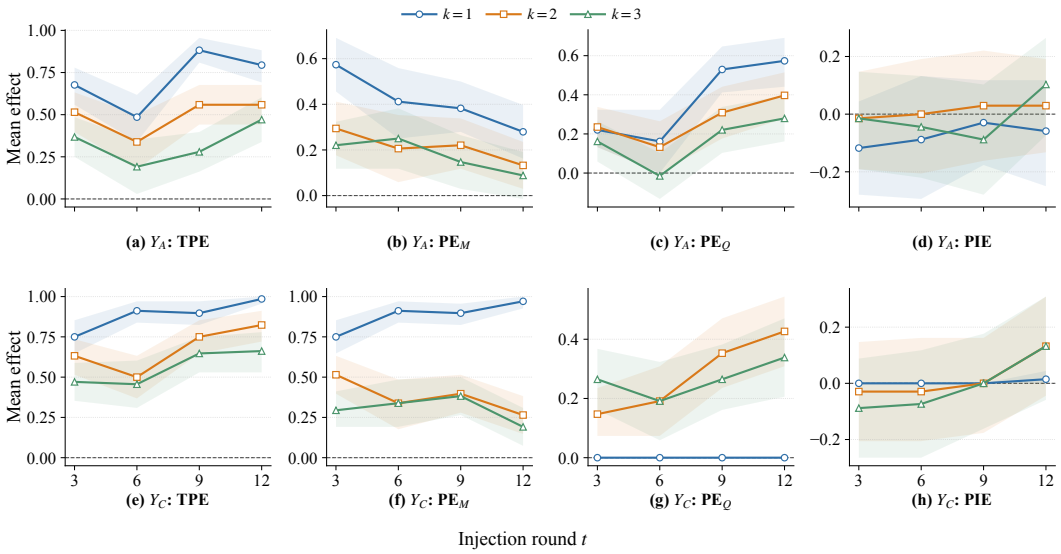



Figure 5: Position sensitivity for OLMo3-7B-Instruct on the common valid instances. Curves report the four outcomes and path effects for $t \in \{3, 6, 9, 12\}$; here, $k$ is the interaction distance, i.e., the number of subsequent turns after injection. Shaded regions are 95% instance-level bootstrap intervals.

The main paired contrasts are Qwen $t = 6 - t = 3$ $Y_A$ TPE $-.222$ $[-.357, -.091]$, Qwen $t = 12 - t = 3$ $Y_A$ TPE $+.147$ [.048,.246], and OLMo $t = 9 - t = 3$ $Y_C$ TPE $+.147$ [.044,.245]. The shift in the natural-response pathway balance from $t = 3$ to $t = 9$ is $-.313$ $[-.437, -.190]$ for Qwen and $-.260$ $[-.407, -.118]$ for OLMo. These are pre-specified position contrasts; the complete pairwise table is included with the released analysis files.

### A.3 Expanded Turn-3 Cohort

We added 84 independently constructed $t = 3$ candidates. The new candidates are kept in a separate cohort and combined with the original instances only for the analyses reported here. Category-wise validity counts are shown in Table 4, and the combined-cohort effects are summarized in Table 5.

| Model | Cohort | Constraint | Decision | Plan | Preference | Profile | Routine |
|---|---|---|---|---|---|---|---|
| Qwen | Old | 12/24 | 24/24 | 23/24 | 21/24 | 21/24 | 24/24 |
| Qwen | New | 22/24 | 12/12 | 12/12 | 6/12 | 8/12 | 12/12 |
| Qwen | Combined | 34/48 | 36/36 | 35/36 | 27/36 | 29/36 | 36/36 |
| OLMo | Old | 20/24 | 21/24 | 24/24 | 16/24 | 19/24 | 20/24 |
| OLMo | New | 20/24 | 10/12 | 12/12 | 7/12 | 5/12 | 11/12 |
| OLMo | Combined | 40/48 | 31/36 | 36/36 | 23/36 | 24/36 | 31/36 |

Table 4: Valid instances by memory category. Entries are valid/candidate counts; the expansion is intentionally reported separately from the original cohort.

| Model | Window | $Y_A$ TPE | $Y_C$ TPE |
|---|---|---|---|
| Qwen | $k = 1$–$3$ | .450 [.394,.506] | .680 [.631,.728] |
| Qwen | $k = 1$–$12$ | .195 [.159,.232] | .328 [.286,.370] |
| OLMo | $k = 1$–$3$ | .488 [.432,.546] | .613 [.564,.661] |
| OLMo | $k = 1$–$12$ | .218 [.178,.258] | .272 [.236,.310] |

Table 5: Combined-cohort propagation effects after the 84-instance expansion.

For the combined $k = 1$–$3$ controlled-probe ranking, Preference has the largest Qwen point estimate and a bootstrap first-rank probability of .982. OLMo's Preference and Plan point estimates are .667 and .713, respectively, with first-rank probabilities .219 and .554; the complete pairwise intervals do not support a stable category ranking across all outcomes.

### A.4 Seed and Generation-Length Sensitivity

The independent-seed runs use seed 20261001, $t = 3$, 15 rounds, four trajectories, and $k = 1, \ldots, 12$. The Qwen max256 output was checked for duplicate checkpoint blocks; 27 exact duplicate blocks were removed before analysis, and the resulting groups passed full structural validation. Run-level validity is summarized in Table 6, and observed output lengths are reported in Table 7.

| Run | Groups | Valid | Invalid | Valid rate | Records |
|---|---|---|---|---|---|
| Qwen max128 | 144 | 120 | 24 | 83.3% | 1872 |
| Qwen max256 | 144 | 117 | 27 | 81.3% | 1836 |
| OLMo max128 | 144 | 120 | 24 | 83.3% | 1872 |

Table 6: Validity of the independent-seed and answer-length runs.

| Run | Field | Mean tokens | P95 | Maximum | Limit | Touch/exceed |
|---|---|---|---|---|---|---|
| Qwen max128 | Answer | 97.5 | 128 | 130 | 128 | 41.25% |
| Qwen max256 | Answer | 139.8 | 256 | 257 | 256 | 23.37% |
| OLMo max128 | Answer | 111.2 | 128 | 130 | 128 | 64.89% |
| Qwen max128 | Memory update | 98.2 | 236 | 256 | 256 | 3.31% |
| Qwen max256 | Memory update | 111.3 | 256 | 256 | 256 | 8.34% |
| OLMo max128 | Memory update | 44.9 | 104 | 104 | 256 | 0.27% |
| Qwen max128 | Question | 22.6 | 34 | 90 | 96 | 0.00% |
| Qwen max256 | Question | 22.8 | 34 | 97 | 96 | 0.014% |
| OLMo max128 | Question | 22.8 | 34 | 67 | 96 | 0.00% |

Table 7: Output-length diagnostics. Touch/exceed rates indicate that the saved text reached or exceeded the configured limit; logs did not include finish reasons.

On 111 common Qwen groups, max256 minus max128 changes natural-response $Y_A$ TPE by $+.066$ $[-.009, .144]$ over $k = 1$–$3$ and $-.006$ $[-.061, .047]$ over $k = 1$–$12$. Relative to the original $t = 3$ records, the new seed changes the $k = 1$–$12$ $Y_A$ TPE by $-.004$ $[-.063, .057]$ for Qwen and $-.025$ $[-.081, .031]$ for OLMo. The corresponding common-group counts are 114 and 111. The complete paired contrasts for $\mathrm{PE}_M$, $\mathrm{PE}_Q$, and PIE are included in the released CSV files.

### A.5 Evaluator Robustness

The primary proposition-level labels were compared with GLM-5.2 on all $35{,}280$ outcomes and with GPT-5.5 on a stratified sample of $900$ outcomes. Agreement is reported together with Cohen's $\kappa$ in Table 8.

| Comparison | Agreement | $\kappa$ |
|---|---|---|
| Qwen–GLM | 91.9% | .709 |
| Qwen–GPT | 89.3% | .591 |
| GLM–GPT | 96.1% | .805 |

Table 8: Agreement among proposition-level evaluators. The GPT comparison uses a stratified sample; the primary Qwen–GLM comparison covers all proposition-level outcomes.

Re-judging preserves the direction of all 72 reported $\mathrm{TPE}$ estimates, with temporal-curve correlations $r = .975$–$.996$. The external judges therefore change absolute label rates without changing the temporal or pathway-level conclusions.

### A.6 Episodic Storage and Retrieval

The episodic analysis contains 107 valid groups, 1,284 post-intervention records, and 5,136 trajectory-distance observations. Each round retrieves the top three entries with the deterministic length-normalized lexical-overlap score described in the main text. The storage metric labels the full episodic memory before the current round; the retrieval metric labels only the top-three entries used for the controlled probe. The resulting storage, retrieval, and expression rates are reported in Table 9.

| Metric | $\mathcal{T}_{ec}$ | $\mathcal{T}_{ee}$ |
|---|---|---|
| Stored error | .869 [.814,.918] | .914 [.875,.950] |
| Retrieved error | .901 [.855,.942] | .939 [.906,.967] |
| Natural-answer error | .061 [.037,.090] | .171 [.132,.214] |
| Controlled-probe error | .184 [.139,.234] | .217 [.162,.275] |

Table 9: Episodic storage, retrieval, and expression rates averaged over $k = 1$–$12$. The high full-storage rate and lower controlled-probe rate separate retention from current-query use.

### A.7 PATHWAY-GUIDED RESTORATION

Question Repair restores the dialogue history and previous response while retaining the error-affected memory. Memory Repair restores the memory state while retaining the error-affected dialogue context. Joint Repair restores memory, dialogue history, previous response, and current question. The paired Memory Repair minus Question Repair residual-MRE difference is $-.075$ with 95% CI $[-.125, -.030]$. Across instances, the association between question-feedback effect and Question Repair gain is $\rho = .335$, 95% CI $[.158, .493]$. The Joint Repair MRE of .003 is caused by inconsistent proposition-level labels for otherwise identical outputs.

### A.8 ANALYSIS INTEGRITY CHECKS

The position analysis contains 28,164 instance-effect records with no duplicate primary keys. Valid trajectories contain the theoretical post-injection distances: 12 for $t = 3$, 9 for $t = 6$, 6 for $t = 9$, and 3 for $t = 12$. The identity $\mathrm{TPE} = \mathrm{PE}_M + \mathrm{PE}_Q + \mathrm{PIE}$ holds with maximum absolute numerical error $5.4 \times 10^{-15}$. All analyses preserve the original valid/invalid decisions and write per-instance outputs, manifests, and bootstrap summaries separately from the primary results.