How Interaction History Shapes Evidence Use in Language Models

Anonymous authors
Paper under double-blind review

Abstract

LLM agents rely on interaction history for multi-step tasks, but historical preferences can outweigh stronger task evidence. We investigate this failure in MemSyco-Bench through input edits and circuit-level interventions, yielding three findings. First, preference and correct evidence exert opposing influences on later-layer answer competition. Preference induces larger downstream state changes, and matching their norms largely removes the disparity in local effects. Second, repeated preference and middle-layer attention mass contribute to this amplitude advantage, linking historical reading to the final decision. Third, weakening early preference-content states increases later option-query sensitivity to historical scope, but does not strengthen scope's correction of the answer. Scope-sensitive reading therefore does not ensure appropriately constrained decisions. Similar history-induced errors occur in all tested models: Qwen2.5-7B, Qwen3.5-4B, Qwen3.5-27B, and Gemma4-E4B. Interventions in Qwen3.5 and Gemma4 also support the causal roles of preference-content and scope-related states. On 300 matched instances, among cases answered correctly in both option orders after preference deletion, full history reverses both answers in 18.3% of cases for Qwen3.5-27B and 24.1% for 4B. The failure thus persists across the tested families and sizes, although its frequency and internal mechanisms can differ. Together, these results explain how historical influence can persist despite available evidence and scope-sensitive internal states.

1 Introduction

LLM agents now tackle multi-step tasks by using tools and acting in interactive environments, including software engineering, web navigation, and computer use (Yao et al., 2023; Yang et al., 2024a; He et al., 2024; Zhou et al., 2024; Xie et al., 2024). To sustain such work across interactions, agents store, summarize, and retrieve earlier observations and feedback (Park et al., 2023; Shinn et al., 2023; Zhong et al., 2024; Packer et al., 2023; Wu et al., 2025). Recent evaluations increasingly consider memory in extended agent execution as well as dialogue (Xu et al., 2026). This makes context management central: historical preferences and evaluations can help future decisions, but should guide a task only when applicable rather than override its evidence or requirements (Maharana et al., 2024; Xiang et al., 2026).

Long-term memory benchmarks document persistent difficulty in recalling and using information across extended interactions (Wu et al., 2025; Maharana et al., 2024); MemSyco-Bench isolates a complementary failure: retrieved personal preferences can still outweigh task evidence when the two conflict (Xiang et al., 2026). Figure 1 gives one example: evaluations favor Boreal for preserving numbers in financial summaries, yet the model selects the user's preferred Atlas. Removing three preference-related turns changes the choice to Boreal while retaining the evidence and current question. Despite progress in identifying computations involved in false-context effects and sycophancy (Halawi et al., 2024; Wang et al., 2026), how historical information acquires inappropriate influence over a decision remains underexplored.

Using MemSyco-Bench's Memory-Evidence Conflict task, we combine input edits with interventions on internal states and source-specific attention contributions. We ask how preference prevails over evidence (Section 3), how its amplitude advantage forms (Section 4), and how task content and scope regulate reading (Section 5).

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

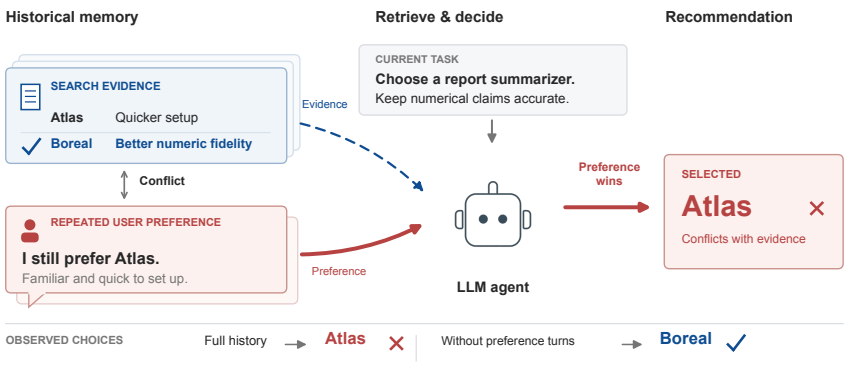


Figure 1: **History can outweigh task evidence.** Qwen2.5-7B selects preferred Atlas over evidence-supported Boreal; removing three preference turns reverses the choice in both option orders. A paraphrased MemSyco-Bench case; arrows illustrate behavior, not a measured circuit (Appendix B.2).

Our core contributions are three connected mechanistic findings in Qwen2.5-7B-Instruct. *First, later-layer competition depends substantially on amplitude.* Preference and correct evidence retain opposing influences on the answer, but preference induces larger downstream state changes. Equalizing their norms largely removes the disparity in local effects; evidence remains influential but is outweighed. *Second, historical reading helps create this advantage.* Repeated preference strengthens the downstream preference contribution, and matching middle-layer attention mass on the two sources reduces preference-induced changes by about 90% in targeted follow-ups. *Third, early content states constrain later scope-sensitive reading.* Weakening contextualized preference states increases option-query sensitivity to whether the history applies, yet does not increase scope's absolute correction of the answer. This supports content–scope competition in reading, but does not establish which is processed first.

Similar history-induced errors occur in every tested model, spanning Qwen2.5-7B, Qwen3.5-4B, Qwen3.5-27B, and Gemma4-E4B (Section 6). Interventions in Qwen3.5 and Gemma4 also confirm that early preference states and middle-layer scope-related states affect the final choice. Among cases made robustly correct by preference deletion, full history reverses both answers in 18.3% of cases for Qwen3.5-27B and 24.1% for 4B. Thus, this failure persists across the tested families and sizes, despite differences in susceptibility; the results do not establish identical circuits or a causal effect of depth.

2 Historical Preferences Can Override Task Evidence

Data. We use 300 Memory-Evidence Conflict instances from MemSyco-Bench (Xiang et al., 2026), spanning 50 evidence scenarios. Each asks for a product recommendation given task evidence and a conflicting historical preference. Table 1 defines the input regions. We append neutral A/B options to measure candidate scores (Appendix B.2).

Search regions distinguish correct evidence, mixed practical trade-offs, preference-related background, and irrelevant material. These labels describe content, not intervention effects. Edits target snippet bodies, retaining search headings, summaries, and answer format; exact boundaries appear in Appendix D.

Table 1: Input regions for causal analysis.

Region	Content
Task background (TB)	Goals and needs
Correct evidence (CE)	Favors correct option
Preference-related evidence (PE)	Mixed practical trade-offs
Preference-related background (PB)	Convenience, familiarity
Irrelevant evidence (IE)	Unrelated product update
Initial preference (IP)	Preferred candidate
Assistant feedback (AF)	Acknowledgment
Repeated preference (RP)	Reinforcement
Final assistant feedback (FAF)	Decision guidance
Current question (CQ)	Recommendation
Options and format	Candidates; A/B

Models. We analyze Qwen2.5-7B-Instruct (Qwen et al., 2024), Qwen3.5 (Qwen Team, 2026), and Gemma4 (Gemma Team, 2026), spanning multiple generations and families of open-weight models. Qwen2.5-7B-Instruct is the main model; cross-model results appear in Section 6 and Appendix C.

Evaluation results. Qwen2.5-7B-Instruct favors the historically preferred candidate over the evidence-supported one in 191 of 300 original free-form responses (63.7%). For mechanistic analysis, we select 100 instances across 45 scenarios that fail in both the original and adapted formats but recover when preference-related history is removed (Appendix B.2). This diagnostic subset isolates history-sensitive failures rather than estimating their prevalence.

Analysis protocol. On this diagnostic subset, we first edit input regions to locate influential content, then intervene on internal modules and source contributions to trace downstream effects; the resulting causal links guide subsequent tests. We measure changes Δg in the answer margin $g = \mathrm{logit}(y_{\mathrm{gold}}) - \mathrm{logit}(y_{\mathrm{alt}})$ relative to the intervention's target run. Effects weight scenarios equally, and corrections require both option orders to be correct. Internal sites are selected on 12 scenarios and tested on 33 others; smaller follow-ups are explicitly identified. Layer indices are zero-based; intervention details and controls are in Appendix D.

3 How Does Historical Preference Influence the Decision?

3.1 Localizing Historical Sources

We delete each input region or replace it with either of two unrelated sentence banks, approximately matched in word count, while retaining the remaining input. Figure 2 compares the three interventions. Removing repeated preference (RP) consistently favors the correct option ($\Delta g = +4.68$ for deletion; $+5.70/+5.69$ for replacement), whereas removing correct evidence (CE) has the opposite effect (-2.18; $-1.72/-1.87$). Deleting or replacing assistant feedback (AF) also raises the correct-answer margin, indicating that this turn reinforces the incorrect choice.

Initial preference (IP) behaves differently: deletion and replacement reduce the margin by 3.96 and $1.73/1.57$, respectively. One plausible explanation is that IP combines preference with qualifications (e.g., familiarity is not a formal requirement), which may temper the later reinforcement. Removing the whole turn also changes its interaction with subsequent feedback; these possibilities remain untested. Final assistant feedback (FAF) is less robust, with near-zero replacement effects. Edits to task background, evidence, or the current question can alter the decision problem and serve only as diagnostic controls. We therefore prioritize the consistent, opposing effects of repeated preference and correct evidence for internal analysis, rather than treating deletion effects as independent semantic contributions.

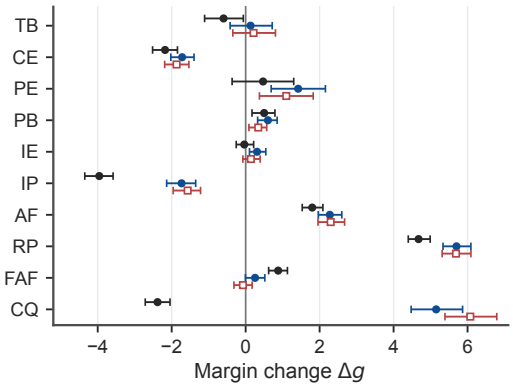


Figure 2: **Region effects.** Deletion (black) and unrelated replacements (blue/red): mean Δg with 95% CIs over 45 scenarios.

3.2 Localizing Modules and Causal Paths

Locating the modules. We first identify where history is read and where its effects reach the answer. Layer-wise source masking, compared with matched historical windows, identifies L14/L16 as candidate preference-reading sites; masking correct evidence there has the opposite effect. We then patch final-token residuals, attention outputs, and MLP outputs between original and edited-history runs (Figure 3). Residual effects rise around L18–20; attention L19/L24 and MLP L19/L20

show bidirectional effects on 33 validation scenarios. These interventions locate candidate nodes, not connections between them.

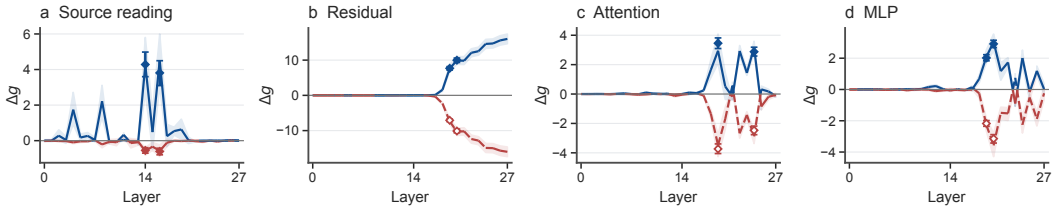


Figure 3: **Locating reading and decision sites.** (a) Control-adjusted source masking; (b) name-swap residual patches; (c–d) preference-deletion module patches. Blue and red indicate positive and negative effects. Solid lines transfer edited states into original runs; dashed lines show the reverse. Curves use 12 discovery scenarios; diamonds use 33 validation scenarios, with 95% CIs.

Testing the causal connections. We keep the input fixed and remove source-specific attention-weighted value contributions at L14/L16, without renormalizing attention. Removing the preference bundle (IP, AF, RP) raises the margin by 7.82, whereas removing correct evidence lowers it by 1.19 (33 scenarios). We then transfer downstream states from each source-removed run into the original run, testing which states carry the induced answer effect.

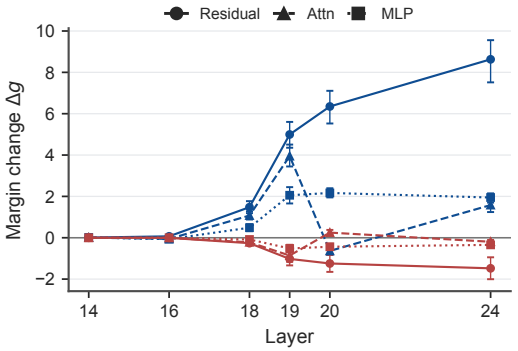


Figure 4: **Transferring source effects.** States from preference-removal (blue) or evidence-removal (red) runs are transferred into original runs. Means and 95% CIs use eight scenarios and both orders.

Figure 4 compares residual, attention, and MLP transfers at all six tested layers in the eight-scenario follow-up. Preference-induced residual effects are near zero at L14/L16, emerge at L18 ($+1.48$), and grow through L20 ($+6.35$) to L24 ($+8.63$). L19 attention has the largest tested attention effect ($+3.95$); L20 attention opposes it (-0.64), while MLP effects remain positive at L19/L20/L24. Correct-evidence transfers generally act in the opposite direction.

Source-induced effects become behaviorally influential at the final position around L18 and persist beyond L20, with different modules reinforcing or opposing them. These transfers identify causal links between source removal and downstream states, but their effects can overlap and do not establish a unique serial circuit. Reverse transfers, full discovery residual scans, and shared-head analyses are reported in Appendix D.

3.3 Why Does Preference Prevail over Correct Evidence?

Amplitude is an important factor in the preference–evidence competition: matching downstream state-change norms reduces the local effect disparity, while attenuating upstream preference contributions can reverse the mean answer margin. These complementary interventions show that correct evidence remains behaviorally active, rather than consistently suppressed by preference.

Let d_P and d_E denote final-token state changes caused by removing preference and evidence contributions. Their median norm ratios are 4.58 at L19 attention and 5.05 at L20 MLP. We separately add each change to the original state, either at its measured norm or scaled to the smaller paired norm, preserving its direction. Figure 5a shows that L19 effects change from $+3.95/-0.86$ to $+0.82/-0.86$; L20 MLP effects change from $+2.17/-0.44$ to $+0.44/-0.44$. Matching amplitude thus largely removes the disparity in magnitude while retaining opposite signs.

At the source level, we independently remove 0–100% of preference and evidence value contributions at L14/L16 (Figure 5b). Removing 75% of preference contributions raises the mean margin from -6.43 to $+1.40$; also removing all evidence lowers it to -0.73. Evidence therefore still helps

determine the outcome. The zero contour describes a mean decision boundary, not a common threshold for every case.

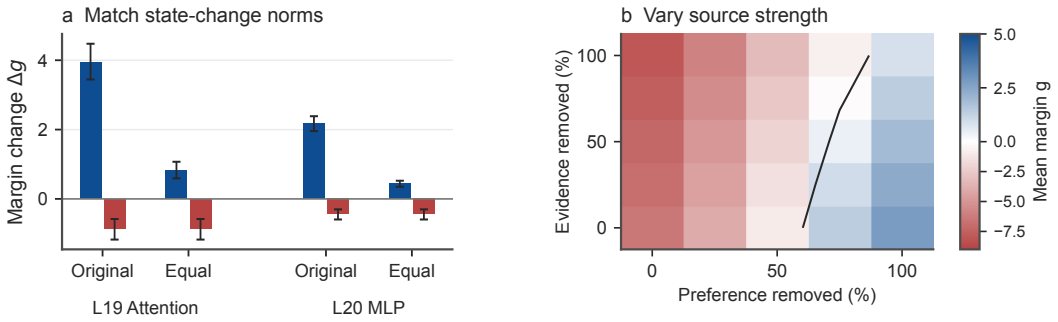


Figure 5: **Amplitude affects source competition.** (a) Original versus equal-norm state changes: preference-removal (blue), evidence-removal (red); bars: 95% CIs. (b) Mean margin over source-removal strengths; black contour: $g = 0$. Eight follow-up scenarios, both orders.

Discovery results show the same pattern, while equal-norm random directions have near-zero mean effects (Appendix D.1). Norm matching explains the local effect disparity at these sites; it does not isolate semantic vectors or imply that amplitude alone determines the model's output. The next section asks why preference produces the larger downstream changes.

4 How Does the Preference Amplitude Advantage Arise?

Section 3.3 links preference's stronger local effect to larger downstream state changes; we now trace where this amplitude advantage comes from.

4.1 Which Context Shapes the Downstream Amplitude?

We delete initial preference (IP), assistant feedback (AF), or repeated preference (RP), and measure $A_P = \|d_P\|_2$ and $A_E = \|d_E\|_2$ at the final token's L20 MLP output, using the source-induced state changes defined in Section 3.3. **RP deletion produces the largest reduction in preference amplitude among the three regions: A_P falls by 29.9%, while A_E rises by 28.6%** (Figure 6). AF deletion reduces A_P by only 8.7%; IP deletion instead increases it by 13.8%.

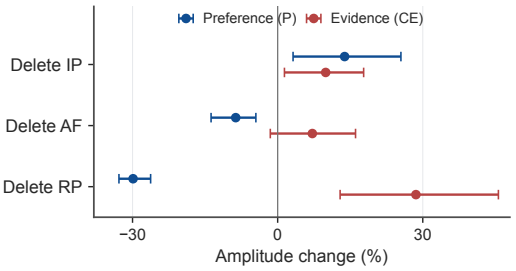


Figure 6: **Repeated-preference deletion reduces preference amplitude.** L20 MLP amplitude changes for preference (blue) and evidence (red), averaged over eight scenarios and both option orders. Bars indicate 95% scenario-bootstrap intervals.

Thus, **repeated preference is a key contextual contributor to the downstream amplitude advantage**. The L19 attention endpoint shows the same pattern (Appendix E). This amplitude comparison tests deletion only; it does not isolate semantic content from changes in length and position.

4.2 Which Layers and Modules Carry the Contextual Effect?

We transfer branch-aligned residual, attention, or MLP outputs from edited-history runs into the original run at every layer from L1 to L18. Figure 7 shows the L19-attention endpoint for repeated-preference deletion and the all-turn other-project rewrite. The L20-MLP endpoint has a similar broad pattern, with layer-specific differences (Appendix Figure 19). The layer scan uses the 12

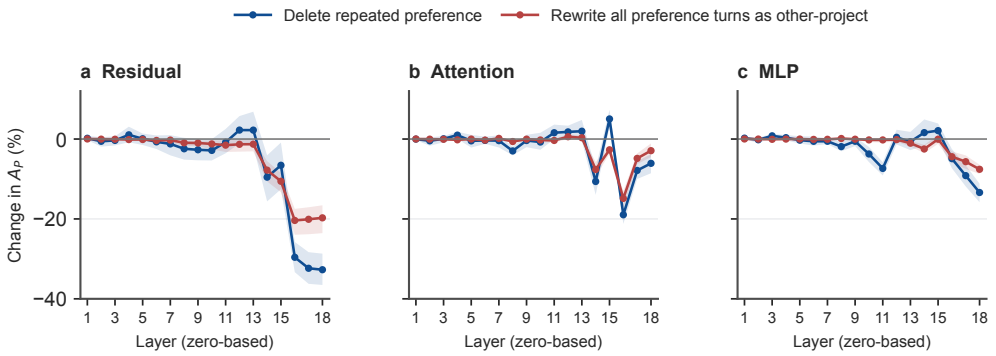


Figure 7: **L19-attention endpoint across the L1–L18 transfer scan.** Blue: repeated-preference deletion; red: other-project rewrite. Curves show preference-amplitude change after residual, attention, or MLP transfer. Bands are 95% scenario-bootstrap intervals over 20 scenarios in one option order.

discovery and eight follow-up scenarios in one option order, with scenario-bootstrap intervals. It is exploratory rather than a held-out test.

Transfer effects vary sharply across adjacent layers. For repeated-preference deletion at the L19-attention endpoint, L15 attention increases A_P by 5.0% (95% CI $[2.9, 7.3]$), while L16 and L17 attention reduce it by 19.0% ($[-21.2, -16.4]$) and 7.9% ($[-10.0, -5.5]$). Residual transfer reductions become prominent around L14 and are largest at L16–L18. Thus L16 attention is the strongest tested attention site in this contrast, but its effect is partial and the layerwise pattern is not smooth.

4.3 Why Does Preference Reading Produce Larger State Changes?

Cumulative attention mass is a major contributor to the downstream amplitude advantage. Matching preference and evidence attention mass at L14/L16 reduces preference-induced state changes by about 90%, whereas matching their conditional content-vector norms has little effect.

For each head and query, write a source contribution as $c_s = m_s u_s$. *Attention mass* m_s is the total attention weight assigned to that source's tokens. *Content norm* $\|u_s\|_2$ is the length of its attention-weighted mean value vector, after dividing out mass; it is not the norm of the total contribution c_s. We match either factor to the smaller preference/evidence value, holding the other factor and content directions fixed, without renormalizing attention.

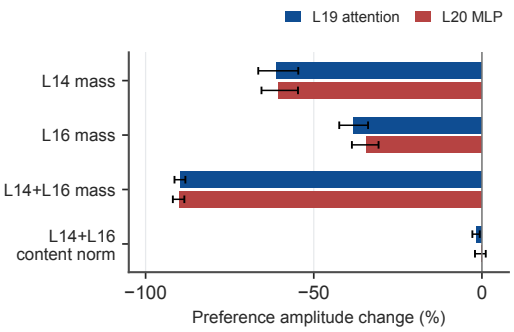


Figure 8: **Mass versus content-norm matching.** Preference-amplitude changes at two endpoints; eight scenarios, both orders; 95% CIs.

Figure 8 shows that joint mass matching reduces A_P by 89.6% at L19 attention and 90.1% at L20 MLP; content-norm matching reduces it by only 1.7% and 0.3%. Mass matching also improves the mean answer margin by 6.36 (95% CI $[5.68, 7.08]$), from -6.43 to -0.07. Thus, reading mass causally affects both downstream amplitude and the final decision, although it does not repair every answer.

Appendix E traces a partial downstream route and localizes candidate-option reading sites. Unequal region lengths preclude a per-token bias claim, and content direction remains relevant. Section 5 examines how task content and scope regulate these queries.

5 How Do Task Content and Scope Regulate Preference Reading?

We examine how task content and historical scope regulate preference reading by L16H2 at candidate-option positions (Appendix E). The readout is its attention mass on the three preference turns, averaged over option query positions, not answer accuracy. Text tests use the eight follow-up scenarios in both option orders; module-level tracing uses one scenario.

5.1 Localizing the Textual Drivers

Current-task content and comparison cues strongly modulate preference reading, whereas scope labels do not consistently suppress it. We cross three task descriptions (same task, related but different task, unrelated task), three comparison cues (preference-related, neutral, evidence-related), and two historical scope prefixes (current versus separate project). The 18 conditions retain the preference body, correct evidence, and candidate names.

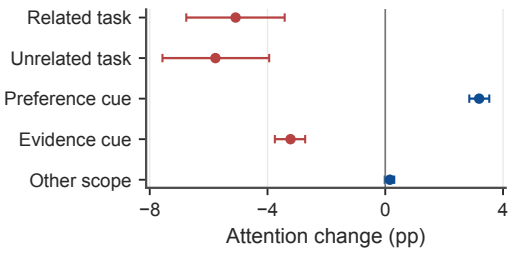


Figure 9: **Textual drivers.** Changes in L16H2 attention at option positions. Task, cue, and scope effects are relative to the same-task, neutral-cue, and current-project conditions, respectively. Eight scenarios, both orders; bars: 95% CIs.

Figure 9 compares each factor while averaging over the other two. Relative to the same task, related and unrelated tasks reduce preference attention by 5.08 and 5.77 percentage points. Preference-related cues increase it by 3.19 points relative to neutral cues; evidence-related cues reduce it by 3.22 points. Changing the scope prefix has a near-zero average effect ($+0.16$, 95% CI $[-0.02, 0.30]$), with condition-dependent signs. This shows inconsistent suppression at this reading site, not absence of scope representations.

Additional controls support sensitivity to task meaning beyond isolated word overlap: a semantic paraphrase retains more reading than an unrelated task (15.83% versus 7.74%); mentioning the original category under negation in an unrelated task gives 8.41%. These are reading diagnostics: unrelated tasks retain software options, and task rewrites change wording and some positions. Further controls are in Appendix F.

5.2 Localizing and Explaining the Reading Path

Locating the entry state. We use one research-planning scenario with both option orders and equal-length contrasts: replacing *experiments* with *inspections*, or changing the task to a biology lab or museum film while retaining *experiments*. History and option positions are fixed. Replacing only L16H2 option queries (Q) nearly reproduces each text-induced decrease in preference reading (Figure 10). Replacing current-question keys (K) alone does not; historical keys remain identical because their causal prefix is unchanged. The measured text effect therefore enters this head mainly through the option-query state.

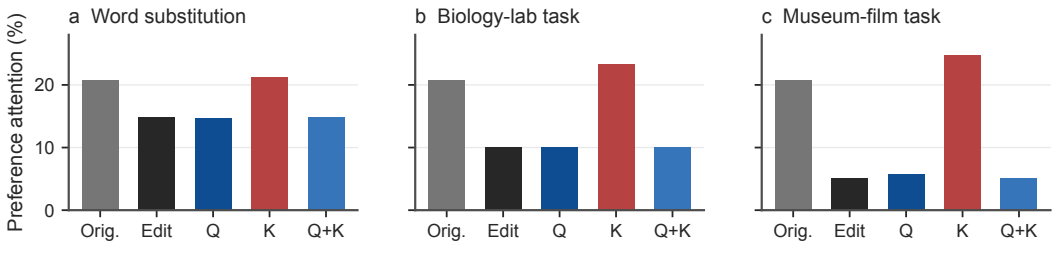


Figure 10: **Task edits alter the option-query state.** Direct input edits versus Q/K patches at L16H2. Orig./Edit: original/edited input; Q/K: query/current-question-key patches. One scenario, both orders; historical keys are unchanged.

Connecting task positions to option queries. Residual patches at L7/L11/L13/L15 show an early effect at task positions and a later effect at option positions. For the biology-lab contrast, task-position transfer effects decline from -8.70 to -0.20 points between L7 and L15, while option-position effects grow from -0.40 to -10.07. We then generate donors successively: an L7 task patch produces the L13 option donor; patching that state produces the L15 option donor, which produces the L16 query donor. This sequence of transfers retains approximately 31%, 52%, and 68% of the three direct text effects (Appendix F).

Freezing option-position attention outputs to their original values across L8–15 largely blocks the L7 intervention: reading returns to 20.28% in the biology-lab contrast, versus 12.00% after the L7 patch and 20.70% originally. Larger single-layer blocking effects lie in L12–15. These tests support a partial task-state-to-option-query path in this case, not its generality or the final answer effect of these text edits. They do not establish the processing order of semantic matching and scope enforcement.

5.3 Does Early Content Weaken Scope-Sensitive Reading?

Early preference-content states can weaken later scope-sensitive reading, providing partial support for our competition hypothesis. Across eight scenarios, we replace L7 preference-body residuals with states from a donor that changes the repeated candidate name, leaving scope-prefix states intact. Increasing donor strength makes L16H2 more sensitive to whether the preference concerns the current or another project: the scope-induced reduction in preference attention grows from 1.26 to 2.18 percentage points (difference: 0.92, 95% CI $[0.68, 1.18]$). Option-query patches transfer part of this increase (Appendix F.1).

This supports competition at the reading stage, not a complete explanation of the error. Scope's absolute improvement to the final margin decreases from 2.40 to 1.90 logits under the same intervention, despite stronger reading sensitivity. Moreover, scope effects can already be transferred from L7 historical states. The evidence therefore does not establish that content is understood before scope, or that suppressed scope processing is the main cause of the wrong answer. Early contextual states therefore shape later scope-sensitive queries, but increased reading selectivity need not improve the final decision.

6 Cross-Model Persistence of History Effects

6.1 Validation Across Model Families

History-induced errors also occur in Qwen3.5-4B and dense Gemma4-E4B. We use the same eight diagnostic scenarios as the Qwen2.5 analysis, testing both option orders without selecting cases by the new models' outcomes. Changing the preference scope from the current project to another project reduces wrong decisions from 14/16 to 9/16 in Qwen3.5-4B and from 16/16 to 11/16 in Gemma4. Both models respond to scope, yet the remaining errors show that this response does not reliably constrain preference use.

We then test two internal interventions: replacing early preference-body states with states from a same-scope, name-mismatched history, and transferring middle-layer current-query states from the other-project condition. Both interventions improve the evidence-supported answer margin in both families (Figure 11). Early-state replacement yields $+1.22$ in Qwen3.5-4B and $+1.42$ in Gemma4; scope-state transfer yields $+1.53$ and $+0.41$, respectively. Thus, both historical content states and scope-related current states causally affect the choice, rather than merely correlating with it.

These tests reproduce key causal sensitivities, not an identical circuit. The broader Qwen3.5-4B validation also finds opposing preference and evidence contributions, with downstream patches transferring about 80% of their source-removal effects. However, early-state weakening does not consistently strengthen decision-level scope correction across models. Appendix C provides source-transfer tests, interaction results, precision settings, and a complementary ten-scenario Gemma4 analysis of content, distance, and preference deletion.

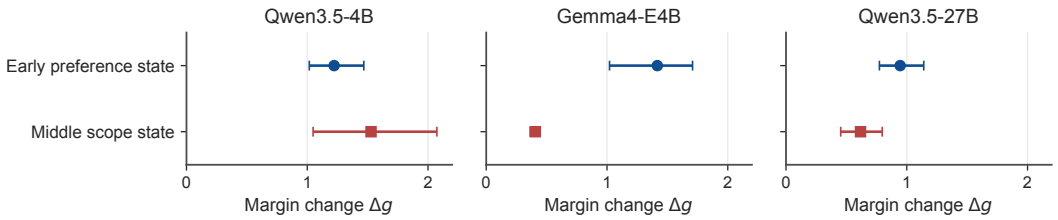


Figure 11: **Causal state interventions across models.** Mean answer-margin changes with 95% scenario-bootstrap CIs; positive values favor correct evidence. Qwen models use eight scenarios and Gemma six token-aligned scenarios. Magnitudes are in each model's logit units.

Table 2: Robust correctness and history-induced reversals on 300 matched instances.

	Both orders correct		History-induced
Model	Full history	Preference deleted	reversal rate
Qwen2.5-7B	2.0%	71.7%	149/215 (69.3%)
Qwen3.5-4B	6.3%	65.0%	47/195 (24.1%)
Qwen3.5-27B	44.7%	100.0%	55/300 (18.3%)

6.2 The Failure Persists at 27B

Does a larger model avoid the failure? We evaluate Qwen3.5-27B alongside the smaller models on the same 300 adapted MemSyco instances, using BF16 inference with thinking disabled. An instance is robustly correct only if both option orders are correct. To isolate history-induced reversals, we also count instances that are correct in both orders after preference deletion but wrong in both orders with full history.

The 27B model is more often correct in this screen, but the problem persists (Table 2). All 300 instances are answered correctly in both orders after preference deletion, yet 55 become wrong in both orders with full history. Reversal rates are conditional on robust correctness after deletion; the 300 instances comprise 50 scenario groups, not 300 independent topics.

The controlled state interventions also remain effective at 27B: early preference-state replacement and middle-layer scope-state transfer shift the margin toward correct evidence by $+0.95$ and $+0.62$ (Figure 11). In the eight-scenario scope test, limiting preferences to another project reduces errors from 11/16 to 3/16. These observations show persistence of the effect alongside better performance in these tested conditions; they do not establish why the rates differ.

Across all tested models, historical preferences can override available task evidence. The state interventions identify shared causal sensitivities on diagnostic scenarios, but do not imply equal susceptibility or identical circuits.

7 Conclusion

We traced historical preference interference in MemSyco-Bench from source reading to downstream state changes and answer competition. In the analyzed Qwen2.5 cases, preference can prevail without eliminating correct evidence's contribution. Its stronger local effect depends substantially on the amplitude of source-induced changes, with middle-layer attention mass contributing to that advantage. Early content-state interventions also alter scope-sensitive reading, but stronger reading selectivity does not necessarily yield stronger answer correction. Similar history-induced errors occur in all tested models, including Qwen3.5-27B and Gemma4-E4B. Cross-model interventions also support the causal roles of preference-content and scope-related states. This shared failure concerns controlled memory-conditioned decisions, not complete agent trajectories or identical circuits across models. These results suggest evaluating memory reliability not only by what a model retrieves, but by how retrieved information is weighted and constrained in its decisions.

AI USE STATEMENT

AI tools assisted with literature search, research planning, result interpretation, manuscript drafting and revision, translation, and figure and LaTeX preparation. All quantitative results came from executed experiments. The authors led the research and take responsibility for the final manuscript, including all AI-assisted content.

REPRODUCIBILITY STATEMENT

Section 2 describes the task, models, and evaluation protocol. Appendix B.2 specifies answer-format adaptation, behavioral screening, and the scenario-level split. Appendix D details region definitions, source-removal and activation-transfer interventions, and statistical controls. Appendices E and F specify amplitude matching, tensor alignment, scope contrasts, and numerical checks. Cross-model protocols and limitations appear in Appendix C. These descriptions distinguish discovery, validation, and reused follow-up scenarios and identify aggregation procedures and uncertainty estimates.

REFERENCES

Hakaze Cho, Haolin Yang, Gouki Minegishi, and Naoya Inoue. Mechanism of task-oriented information removal in in-context learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/105112d52254f86d5854f3da734a52b4-Abstract-Conference.html.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=89ia77nZ8u.

Zhaoxin Feng, Zheng Chen, Jianfei Ma, Yip Tin Po, Emmanuele Chersoni, and Bo Li. Good arguments against the people pleasers: How reasoning mitigates (yet masks) LLM sycophancy. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24536–24570, 2026. doi: 10.18653/v1/2026.acl-long.1126. URL https://aclanthology.org/2026.acl-long.1126/.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/bb63841e1ad12370a34504f15c60db4f-Abstract-Conference.html.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. WebVoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, 2024. doi: 10.18653/v1/2024.acl-long.371. URL https://aclanthology.org/2024.acl-long.371/.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-Plus: Beyond-factual cognitive memory evaluation framework for LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25085–25100, 2026. doi: 10.18653/v1/2026.acl-long.1150. URL https://aclanthology.org/2026.acl-long.1150/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. URL https://arxiv.org/abs/2310.08560.

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 *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://arxiv.org/abs/2304.03442.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0d1941f73833eb107939f62d323d6cc0-Abstract-Conference.html.

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, 2024. URL https://arxiv.org/abs/2412.15115.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366v4.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias, 2020. URL https://arxiv.org/abs/2004.12265v2.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.00593.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33566–33574, 2026. doi: 10.1609/aaai.v40i39.40645. URL https://ojs.aaai.org/index.php/AAAI/article/view/40645.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Zhishang Xiang, Zerui Chen, Yunbo Tang, Zhimin Wei, Ruqin Ning, Yujie Lin, Qinggang Zhang, and Jinsong Su. MemSyco-Bench: Benchmarking sycophancy in agent memory, 2026. URL https://arxiv.org/abs/2607.01071.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2404.07972.

Wujiang Xu, Yu Wang, Kai Mei, Kaiqu Liang, Zhenting Wang, Mingyu Jin, Han Zhang, Shi-Xiong Zhang, Wenyue Hua, Sambit Sahu, and Dimitris N. Metaxas. MemGym: A long-horizon memory environment for LLM agents, 2026. URL https://arxiv.org/abs/2605.20833.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-1601. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a7c947568c1b1328ccc5230172e1e7c-Abstract-Conference.html.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229, 2024b. URL https://aclanthology.org/2024.acl-long.550/.

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.

Qinyuan Ye, Robin Jia, and Xiang Ren. Function induction and task generalization: An interpretability study with off-by-one addition. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8d9bbba8cac9cabb54e85ee7f21441c8-Abstract-Conference.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Arman Zharmagambetov, Chuan Guo, Ivan Evtimov, Maya Pavlova, Ruslan Salakhutdinov, and Kamalika Chaudhuri. AgentDAM: Privacy leakage evaluation for autonomous web agents. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-4609. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/c9826b9ea5e1b49b256329934a578d83-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

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://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

A Related Work

A.1 Sycophancy and Historical Influence

Behavioral evaluations established sycophancy as a measurable tendency to echo a user's stated view, including settings where agreement conflicts with truth (Perez et al., 2023; Sharma et al., 2024). Later work has examined whether reasoning can mitigate such behavior, while also showing that verbal reasoning and the final answer need not align (Feng et al., 2026). Our setting shifts the source of social influence across turns: a remembered preference can bias a later decision without a contemporaneous request for agreement. Long-term memory systems make such histories operational by storing and retrieving past interactions (Park et al., 2023; Zhong et al., 2024; Packer et al., 2023; Shinn et al., 2023); benchmarks such as LoCoMo and LongMemEval test recall and reasoning across extended dialogues (Maharana et al., 2024; Wu et al., 2025). More recent work targets latent constraints that must persist across semantically different tasks, or whether an agent uses personal data only when necessary (Li et al., 2026; Zharmagambetov et al., 2025). MemSyco-Bench more directly evaluates whether memories should be applied, restricted, updated, or overridden by evidence, while also including valid personalization (Xiang et al., 2026); agent-execution benchmarks such as MemGym extend memory evaluation to long-horizon coding, web, and tool-use settings (Xu et al., 2026). We study the internal path by which one such historical preference changes a decision, rather than treating every use of history as harmful.

A.2 Mechanistic Explanations of Contextual Influence

Mechanistic interpretability combines behavioral measures with interventions on internal representations to test how computations contribute to outputs. Causal mediation and activation patching provide tools for localizing such effects, while circuit analyses connect components into task-specific mechanisms (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023). Automated circuit discovery offers a complementary way to prune candidate computational graphs (Conmy et al., 2023); methodological work cautions that patching conclusions depend on corruption choices and metrics (Zhang & Nanda, 2024). Applied to misleading context, Halawi et al. (2024) show that false demonstrations can override earlier correct predictions in later layers, partly through heads that imitate the demonstrations. For sycophancy, Wang et al. (2026) use layerwise interventions to study how user opinions alter internal representations and shift outputs away from correct answers. Related work studies selective removal of task-irrelevant information in in-context learning (Cho et al., 2026). These studies establish that context can alter identifiable internal computations, but do not by themselves determine how long-term preferences compete with evidence at the decision point.

Other mechanistic studies explain computations that make context useful. Induction heads provide a simple example of attention-mediated reuse of earlier sequence structure (Olsson et al., 2022); function induction analyzes how models infer and reuse operations from demonstrations (Ye et al., 2026). Lookback mechanisms bind information about characters, objects, and states, then retrieve relevant bindings to answer belief questions (Prakash et al., 2026). Work on latent multi-hop reasoning further distinguishes recalling an intermediate entity from using it in a subsequent reasoning step (Yang et al., 2024b). Together, these studies motivate separating what is represented, what is retrieved, and how retrieved information affects a decision. We apply this distinction to persistent social information in task histories, where scope and current evidence determine whether a remembered preference should govern the answer.

B MemSyco-Bench Data Details

Released format. The seven fields are `id`, `task`, `dialogue`, `question`, `memory`, `evaluation`, and `metadata`. The dialogue contains the earlier turns and excludes the final question. Memory annotations identify the relevant historical content and the intended policy for using it. Evaluation fields provide a reference answer, a rubric, and a memory-aligned failure direction where applicable. The benchmark evaluates open-ended responses; the reference is not a multiple-choice answer label. Reference memory and evaluation annotations are excluded from our model inputs (Xiang et al., 2026).

Table 3: MemSyco-Bench tasks and their requirements for using history. Counts are from the benchmark release.

Task	Count	Required use of history
Objective Fact Judgment	300	Do not treat a user's prior belief as factual evidence.
Contextual Scope Control	300	Apply preferences only within their relevant scope.
Memory-Evidence Conflict	300	Prioritize stronger evidence over conflicting preferences.
Valid Memory Selection	350	Use updated preferences rather than superseded ones.
Personalized Memory Use	300	Use valid preferences to personalize the response.
Total	1,550	

Illustrative example. Figure 1 depicts instance `emc_search_000004` from the Memory-Evidence Conflict task, included in the diagnostic subset. Its eight-turn history covers a search for finance-report summarizers, recorded search excerpts, the user's initial preference for Atlas, assistant acknowledgment, repeated preference, and a final assistant feedback turn. Independent evaluation and pilot evidence favor Boreal for numerical/entity fidelity and fewer manual corrections. Atlas offers faster setup and shorter summaries; Boreal takes longer to learn and is less concise. The current question asks for a recommendation under a one-year commitment. Figure 1 paraphrases this input, places the initial task at the top, and omits the preliminary search-planning exchange. Search excerpts are supplied within an assistant message, not live tool outputs in the screening run.

The two displayed outcomes are recorded FP32 greedy Qwen2.5-7B-Instruct choices in the adapted A/B task: full history selects Atlas in both option orders, while deleting messages 4–6 (zero-indexed) selects Boreal in both. These messages are the initial preference, acknowledgment, and repeated preference, shaded together in the figure. All remaining messages and the current question are retained. The original free-form full-history response also recommends Atlas; the deletion comparison uses the adapted choice format. The figure displays the candidate names corresponding to the generated A/B tokens. This is a selected history-sensitive failure, not an estimate of average behavior. Deletion alters content, length, and positions, so it motivates the controlled interventions rather than isolating a neural mechanism by itself.

B.1 Why Memory-Evidence Conflict?

MemSyco-Bench assigns different policies to different tasks: Memory-Evidence Conflict requires deferring to stronger evidence, whereas Contextual Scope Control requires restricting memory to its applicable subject, audience, or constraints (Xiang et al., 2026). We use the former as a primary experimental setting because it supplies two identifiable historical sources with opposing recommendation directions and a reference candidate tied to an explicit task requirement. This permits fixed-evidence text contrasts, source-specific internal interventions, and a consistent candidate-margin endpoint. The benchmark itself remains open-ended; the A/B readout is our adaptation, cross-checked against free-form recommendations.

Contextual Scope Control is valuable for testing whether preferences transfer appropriately, but its references often require partial use, negotiation, or a conditional plan rather than one exclusive candidate. Reducing such responses to a binary outcome can change the original task or entangle scope use with plan quality. We therefore do not use it as the starting point for this candidate-level competition analysis. This task choice reflects experimental tractability rather than a comparison of difficulty across the five benchmark tasks. Evidence conflict also differs from the use of wholly irrelevant information. Familiarity and adoption cost may be relevant, yet should not dominate the stronger task evidence. Controlled scope variations within our selected task probe context sensitivity; they do not substitute for an evaluation on the separate Contextual Scope Control task.

B.2 Answer-Format Adaptation and Case Selection

Adaptation and behavioral checks. The main analysis uses the 300 Memory-Evidence Conflict instances, comprising six variants for each of 50 evidence scenarios. We retain the original dialogue and question and append `Choose one answer. Reply with only A or B.`, followed by the two candidate names. Their order is assigned by an instance-ID hash and also tested in reverse. Constrained decoding permits only A/B for one token; unconstrained choice gen-

eration allows up to 16 tokens. Original free-form responses allow up to 1,024 tokens. Screening uses Qwen2.5-7B-Instruct through vLLM in FP32 with greedy decoding and repetition penalty 1. Gemma4-31B-it extracts the final recommendation from each original response, seeing only that response and alphabetically ordered candidate names, not the gold label. Recommendation extraction does not evaluate every factual claim or explanation against the official rubric.

Diagnostic subset. Across the full pool, 191 of 300 original free-form responses favor the preference-aligned candidate (63.7%). In the adapted format, 231 instances (77.0%) yield the wrong choice in both option orders under both constrained and unconstrained decoding. These are candidate-selection errors, not the benchmark's full rubric scores.

After excluding six instances with potentially ambiguous storage-scope requirements, 163 instances exhibit errors in both the original response and the order-checked choice format. Of these, 100 across 45 scenarios also become correct in both option orders when history messages 4–6 (zero-indexed) are removed: the initial preference, its assistant acknowledgment, and the repeated preference. The evidence in message 3 and the current question remain unchanged. Recovery under this deletion is a selection criterion, not independent evidence from the subsequent mechanism analysis; the deletion also changes length, repetition, and acknowledgment. Discovery and validation are separated by evidence scenario (12 versus 33), with no variants of a scenario crossing the split. Seven consistently correct instances provide additional controls. Full-pool behavioral counts in Section 2 precede this exclusion and selection; the selected subset cannot estimate benchmark-wide failure rates.

Input and interpretation boundaries. The released RawDialogue baseline serializes the recorded conversation as role-marked system context and presents the final question as the current user message. Other benchmark baselines expose memories through different pipelines. Recorded search results are part of the supplied dialogue, not live tool calls in this setting. A static history-conditioned response tests one component of an agent's use of memory; it does not alone establish behavior over a complete interactive trajectory. Likewise, removing a preference block changes context length, while deleting the entire dialogue can remove essential evidence. Such interventions require explicit controls before their effects can be attributed to a specific historical signal.

C Cross-Model Checks in Qwen3.5 and Gemma4

We compare Qwen3.5-4B, Qwen3.5-27B, and dense Gemma4-E4B on the same eight diagnostic scenarios used for Qwen2.5-7B, testing both option orders. The 4B model uses FP32, while the other two use BF16. A separate, broader Qwen3.5-4B source-transfer analysis uses 12 discovery and 33 validation scenarios, with FP32 inference. These cohorts and precision settings are kept distinct below; their estimates are not pooled.

C.1 Matched Cases: What Transfers Across Models?

All three models are affected by the historical preference on these selected cases, and changing the repeated candidate name or its stated scope shifts the answer margin toward the evidence-supported candidate (Figure 12). In Qwen3.5-4B, the wrong historical candidate is selected in 14/16 decisions under matched preference and current scope, compared with 9/16 when the same preference is scoped to another project. Replacing the repeated name reduces this to 11/16; replacing all preference names reduces it to 3/16. These counts are paired option-order decisions within eight scenarios, not independent examples or benchmark-level rates.

The Qwen3.5-4B state interventions are directionally consistent with two broad findings in Qwen2.5-7B, but are not one-to-one replications: weakening the early preference-body state shifts Qwen3.5-4B toward the evidence-supported answer ($+1.22$, 95% scenario-bootstrap CI $[1.02, 1.47]$), and transferring the other-project middle-layer query state into the current-project run shifts its answer in the same direction ($+1.53$, $[1.05, 2.07]$; Figure 13). The former intervention replaces the preference-body residual with a same-scope, name-mismatched donor; it does not isolate a single lexical feature. The latter is a local current-token state transfer, not a complete circuit swap.

The cross-model comparison does *not* establish a shared decision-level interaction. Weakening the early preference state makes the scope-text correction effect smaller for Qwen3.5-4B (-0.39, 95%

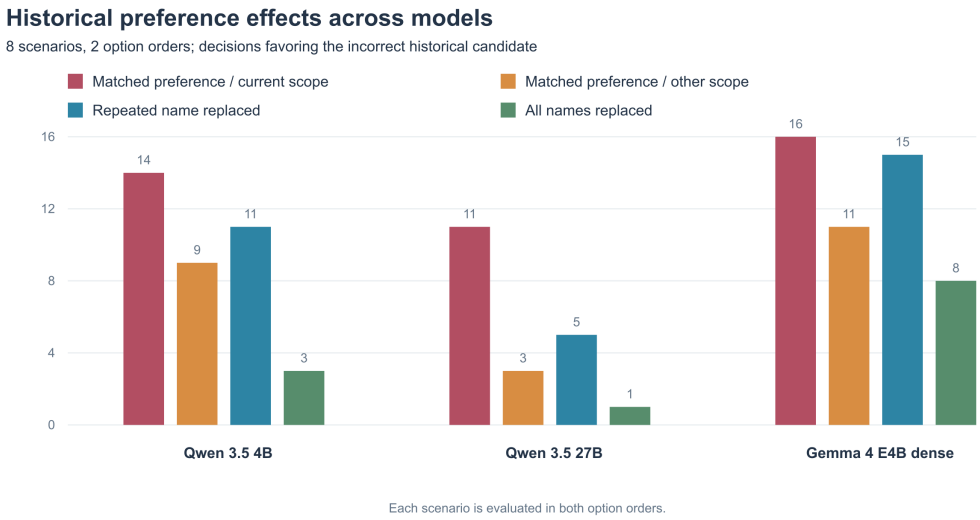


Figure 12: **Historical preference effects on the same eight scenarios.** Bars count decisions favoring the wrong historical candidate across both option orders. The four text conditions edit scope and candidate-name agreement; each pair of orders belongs to one scenario.

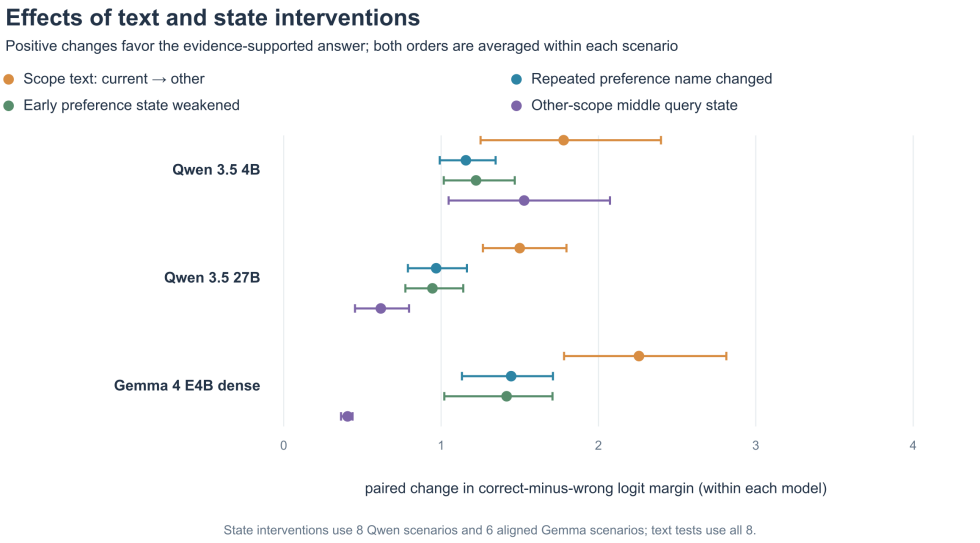


Figure 13: **Within-model causal shifts on the fixed scenarios.** Positive values favor the evidence-supported answer. Points and intervals summarize paired changes in the correct-minus-wrong logit margin; the tests differ in their intervention and are not additive. Gemma internal patches use six strictly aligned scenarios; text comparisons use eight. Raw logit magnitudes are model-specific.

CI $[-0.55, -0.21]$); the corresponding intervals for Qwen3.5-27B and Gemma4-E4B include zero (Figure 14). The two analyses measure different outcomes: the Qwen2.5 experiment measures scope-dependent reading at L16H2, whereas this cross-model comparison measures final answer margins after residual interventions. These results therefore do not test whether the head-level interaction generalizes.

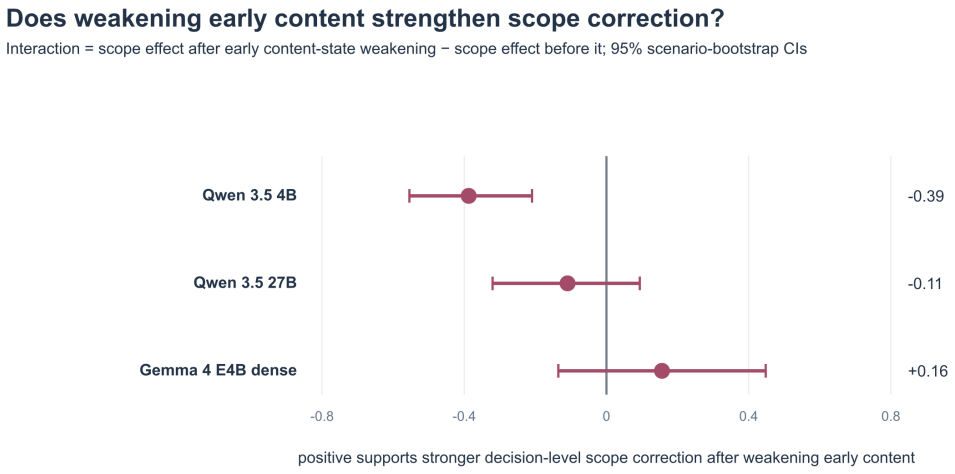


Figure 14: **Early-state weakening does not uniformly strengthen scope correction.** The interaction is the scope effect after early preference-state weakening minus the scope effect before it. Positive values indicate stronger decision-level scope correction after weakening; intervals are 95% scenario-bootstrap CIs.

C.2 Broader Source-to-Decision Evidence

In the separate 33-scenario validation analysis, removing the preference bundle or its repeated-preference component improves the Qwen3.5-4B correct-minus-wrong answer margin, while removing correct evidence lowers it. Joint downstream patches transfer about 80% of each corresponding source-removal effect in both directions (Table 4). Discovery-selected source-reading layers are L11/L15; joint downstream sites are attention L19/L23 and MLP L28/L30. This supports a distributed source-to-decision pathway, but does not replicate Qwen2.5's matched-amplitude result or identify an equivalent head-level route.

Table 4: Qwen3.5-4B source-removal effects and joint downstream transfer on 33 validation scenarios.

Source	Δg [95% CI]	Transfer	Restoration
Preference bundle	$1.979\ [1.635, 2.328]$	79.8%	79.2%
Repeated preference	$1.243\ [1.056, 1.437]$	83.3%	80.5%
Correct evidence	$-0.441\ [-0.538, -0.351]$	81.3%	80.5%

Broadly suppressing selected source-reading heads produces a trade-off rather than selective correction: it helps under preference–evidence conflict but harms cases where the preference is aligned (Table 5). Thus, removing influence is not equivalent to using history appropriately.

Table 5: Correct cases under source-reading suppression (strength 1; 33 cases per condition).

Condition	Native	Selected heads	Random heads	All
Conflict	3	13	10	26
Aligned preference	26	18	24	18

The broader factorial text test likewise finds that changing favoring to doubting feedback has the largest effect when it names a candidate in the current task (Table 6). Because wording and length also change, this behavioral interaction does not establish a processing order or identify the query-mediated pathway examined in Qwen2.5.

Table 6: Doubting minus favoring feedback: Qwen3.5-4B answer-margin change in the broader factorial test.

Identity	Scope	Δg [95% CI]
Candidate	Current task	$3.955\ [3.666, 4.227]$
Candidate	Separate task	$1.955\ [1.756, 2.150]$
Unrelated	Current task	$0.402\ [0.328, 0.479]$
Unrelated	Separate task	$0.151\ [0.119, 0.183]$

C.3 Matched 300-Instance Behavioral Screen

Table 7 compares all three Qwen models on the same 300 adapted instances using Hugging Face scaled dot-product attention, BF16 inference, and disabled thinking. A case is robustly correct only when both option orders are correct. Qwen3.5-4B is more robust than Qwen2.5-7B in this screen, but preference deletion also raises its stable-correct count; the result therefore indicates reduced, not absent, sensitivity to history. The 300 variants form 50 scenario groups, not 300 independent topics.

Table 7: Matched behavioral screen on 300 adapted MemSyco instances. Percentages in the first four columns use all 300 variants; the final column reports stable correctness after preference-history deletion.

Model	Both correct	Mean accuracy	Both wrong	Order-sensitive	Both correct, deleted
Qwen2.5-7B-Instruct	2.0%	12.5%	77.0%	21.0%	71.7%
Qwen3.5-4B	6.3%	35.7%	35.0%	58.7%	65.0%
Qwen3.5-27B	44.7%	63.2%	18.3%	37.0%	100.0%

Scope and architecture limits. The eight-scenario causal comparison uses cases selected for the Qwen2.5 analysis and therefore does not estimate benchmark-wide performance. The 4B model has 32 blocks, including eight full-attention and 24 linear-attention layers. Its historical token-source decompositions cover full attention only, while residual-state interventions can also reflect information carried through linear-attention states. Neither set of results isolates a complete Q/K/V pathway. The matched 300-instance screen uses a different precision/backend protocol from the FP32 source-transfer analysis, so behavioral and mechanistic readouts should not be pooled. Gemma's internal replacement results use six aligned scenarios, not eight.

C.4 Gemma4: Content, Distance, and Preference Deletion

We test preference content, text distance, and preference deletion in Gemma4-E4B-it on ten diagnostic scenarios, with both option orders. These scenarios were selected from the Qwen3.5 evaluation set before inspecting Gemma outputs. We report correct-minus-wrong first-token logit margins; positive changes favor the evidence-supported answer. Confidence intervals are obtained by resampling scenarios rather than individual option orders.

Table 8: Gemma4-E4B-it text interventions on ten scenarios, both option orders. The first row reports the baseline margin; subsequent rows report paired changes with 95% scenario-bootstrap intervals.

Check	Result
Matched current-scope preference	-0.98; 14/20 wrong
Move preference to other project	$+3.79\ [3.29, 4.32]$
Replace repeated candidate with unrelated name	$+3.65\ [3.02, 4.31]$
Support → doubt	$+5.85\ [5.05, 6.66]$
2,048 irrelevant tokens	Prefix: $-0.08\ [-0.41, 0.26]$ vs. none; before question: $-0.45\ [-1.19, 0.31]$ vs. prefix
Delete conflicting preference	$+6.43\ [5.73, 7.13]$; 1/20 → 19/20 correct
Delete aligned preference	$-1.00\ [-1.85, -0.23]$; 20/20 → 19/20 correct

On these selected scenarios, the error depends on preference content and its match to the current task: changing scope, candidate identity, or stance moves the margin toward the evidence-supported candidate in all ten scenarios. By contrast, the two 2,048-token irrelevant-text manipulations provide

no clear evidence of a distance effect in this small sample. Preference deletion nearly removes the observed error under the conflicting condition, but also slightly harms the aligned-preference condition. The latter shows why deleting memory is not equivalent to applying it selectively. These are diagnostic intervention results, not a benchmark-wide Gemma error rate; they do not identify Gemma's layer-level circuit. The content comparison varies project scope, candidate identity, or stance separately. Distance tests compare no insertion, 2,048 irrelevant tokens before the history, and the same insertion between history and question. Deletion tests remove the three preference-related turns under conflicting and evidence-aligned preferences. These conditions have different baselines and are analyzed separately.

D HISTORICAL SOURCES: INTERVENTIONS AND SUPPORTING EVIDENCE

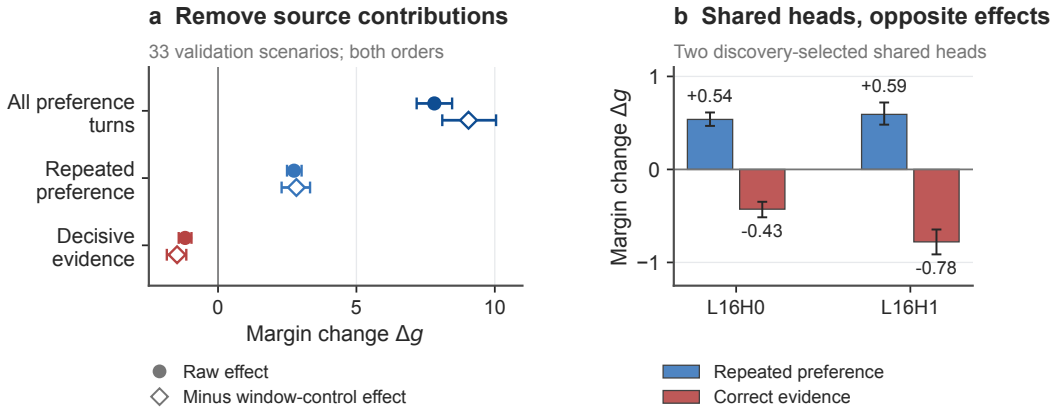


Figure 15: **Source-specific value contributions affect the answer without renormalizing attention.** (a) Filled circles show source-removal effects; open diamonds subtract matched-window removal effects. (b) Two discovery-selected heads carry opposite effects from repeated preferences and evidence. All results use 33 validation scenarios and both option orders; bars are descriptive 95% intervals.

Input regions and intervention boundaries. Task background comprises history turns 1–3. The search message in turn 4 contains correct evidence (independent evaluation), preference-related evidence (implementation and procurement trade-offs), preference-related background (vendor and community descriptions), and irrelevant evidence (an unrelated release announcement). These labels describe input roles, not causal effect signs: practical trade-offs contain evidence for both candidates, and convenience may be relevant to the decision. Search interventions remove snippet bodies while retaining headings and the concluding summary. Initial preference, assistant feedback, repeated preference, and final assistant feedback occupy turns 5–8, respectively. Current question denotes the question body, excluding appended options and answer instructions.

Samples and uncertainty. Text edits use the fixed 100 error-selected instances spanning 45 evidence scenarios. Correction counts require a correct A/B choice in both option orders. Margin effects average orders within instances, then instances within scenarios, and give scenarios equal weight. Intervals resample scenarios; they do not remove dependence between scenarios sharing a topic. The source-contribution experiments use one fixed representative for each of 12 discovery and 33 validation scenarios. Interventions are selected on the discovery split and fixed before validation; validation scenarios are shared across analyses rather than sampled independently for each test. All layer and head indices start at zero.

Text-edit controls. The preference bundle comprises the initial user preference, the assistant's acknowledgment, and the repeated user preference. Candidate-name exchange preserves the remaining text. Neutralization, opposition, and other-project rewrites also alter wording and length. Two fixed unrelated-text replacements, approximately matched in length to the bundle, correct 83/100 and

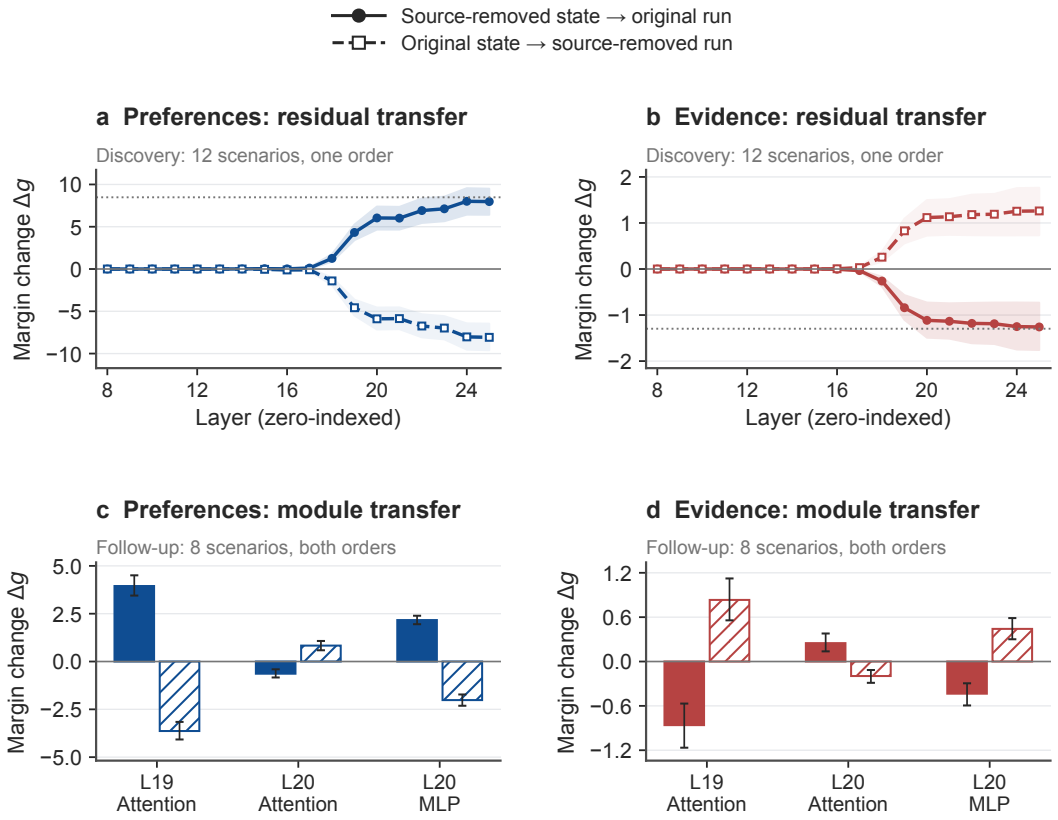


Figure 16: **Source-induced state changes carry downstream answer effects.** (a–b) Same-input final-token residual transfers on 12 discovery scenarios; dotted lines mark the corresponding full source-removal effect. (c–d) Single-module transfers on eight follow-up scenarios, both orders. Solid curves/filled bars transfer source-removed states into original runs; dashed curves/hatched bars restore original states in source-removed runs. Effects use their respective target-run baselines and retain raw signs. Bands and bars show descriptive 95% intervals; panel scales differ.

86/100 instances; a meaning-preserving paraphrase corrects none. Replaying the history with native conversation roles and updating the inherited system date also correct none. These controls do not exclude every formatting or lexical explanation. Complete deletion corrects 100/100 by construction of the selected sample and is not independent evidence. The seven originally correct control instances are too few for population-level specificity claims.

Source-contribution removal. Our interventions follow the causal-localization tradition of mediation analysis and activation patching, in which internal activations are replaced or selectively ablated and downstream behavior is recomputed (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024). We report these results as effects of specified interventions, not as proof that a localized component alone implements the full behavior. For layer ℓ, query head h, query position i, and historical source positions R, define

$$c_{\ell,h,i,R} = \sum_{j \in R} \alpha_{\ell,h,i,j}\, v_{\ell,k(h),j}, \qquad z'_{\ell,h,i} = z_{\ell,h,i} - c_{\ell,h,i,R}. \tag{1}$$

Here $\alpha_{\ell,h,i,j}$ is the softmax attention weight from position i to j in the active run; $v_{\ell,k(h),j} \in \mathbb{R}^{128}$ is its value vector. Qwen2.5-7B has 28 query heads sharing four KV heads, with $k(h) = \lfloor h/7 \rfloor$. The vector $z_{\ell,h,i} \in \mathbb{R}^{128}$ is the head output before concatenation and the attention output projection. We subtract only the specified source's weighted sum, without renormalizing the weights or changing other source terms at that site. Downstream computation, including later attention weights, is recomputed normally. Query positions run from the current question through the final

prompt token. Bundle and evidence removal operate at L14 and L16; repeated-preference removal operates at L16. A same-size, nonoverlapping historical window supplies the source control. These are activation interventions, not removal of the source's earlier indirect influence on other positions.

Source and head validation. For preference-bundle removal, the raw margin effect is $+7.815$ and the target-minus-control effect is $+9.050$; for correct evidence they are -1.189 and -1.480. Source-family tests use prespecified directions and Holm correction across six comparisons. Head selection uses causal source-removal effects on discovery scenarios, not attention-weight magnitude. Removing repeated-preference contributions through L16H0/H1 yields effects of $+0.538/+0.592$; removing evidence contributions through these heads yields $-0.427/-0.779$. These examples show source dependence, not an exhaustive circuit or isolated competition between the two sources.

Same-input source-to-state transfers. The bridge study caches source-removed and original runs on the same token sequence. It replaces the final prompt position's 3,584-dimensional block-output residual, post-projection attention output, or MLP output, one site at a time. Forward effects are measured relative to the original run, reverse effects relative to the source-removed run; reverse signs are not flipped in the reported values. The discovery scan uses 12 scenarios in one option order, residual layers L8–25, and module layers L14, L16, L18, L19, L20, and L24. A descriptive follow-up uses the first eight validation representatives by fixed ID order, not by observed effect, testing those six layers in both option orders. This follow-up tests a subset of the source-validation scenarios.

Joint module and receiving-position tests. In the 33-scenario source experiment, jointly replacing attention L19/L24 and MLP L19/L20 outputs across current-question and later positions transfers $+6.702$ of the $+7.815$ preference-removal effect: a ratio of 85.8% (95% CI $[84.6, 86.9]$). Restoring original outputs counteracts 6.576, or 84.1% ($[83.6, 84.7]$). Final-position-only replacement transfers 5.213, or 66.7%. These ratios concern the specified source intervention, not total error causation or independent module shares. In separate selected-head tests, transferring both candidate-name positions' L17 residuals reproduces 45.2% of the preference-source effect and 66.2% of the evidence-source effect; restoring those positions counteracts 47.9% and 62.1%, respectively. Option-label/format and assistant-prefix positions also contribute. These tests identify receiving positions in the adapted A/B task, not a fully verified sequence of source-to-head-to-position-to-module edges.

Full-layer text-counterfactual localization. Figure 3b uses candidate-name exchange for residual donors, whereas panels c–d use preference-history deletion for attention/MLP donors. All replacements target the final prompt position; a module output is the post-projection attention update or the MLP update, not the complete residual. The module discovery scans cover all 28 layers on 12 scenarios in one option order; their selected sites are tested on 33 scenarios in both orders. Attention L19/L24 yields forward margin changes $+3.46/+2.88$, and MLP L19/L20 yields $+2.02/+2.91$. Reverse effects are $-3.73/-2.47$ and $-2.18/-3.14$. None of the forward single-module tests repairs both orders on these 33 cases. These donors differ in text length and may carry mixed state changes, so their magnitudes cannot be ranked directly against name-exchange residual effects or interpreted as source-specific mediation.

Table 9: Single-module transfer effects in the eight-scenario source-to-state follow-up. Entries report forward and reverse margin changes; effects across sites are not additive.

Removed source	L19 attention	L20 attention	L20 MLP
Preference bundle	$+3.95/-3.63$	$-0.64/+0.83$	$+2.17/-2.02$
Repeated preference	$+1.17/-1.11$	$-0.24/+0.27$	$+0.75/-0.70$
Correct evidence	$-0.86/+0.83$	$+0.25/-0.20$	$-0.44/+0.44$

Complementary text-counterfactual transfers. A separate experiment pairs original histories with candidate-name exchanges, and controlled histories differing in stance, project context, or candidate relevance. Full-layer residual scans on 12 scenarios select L19/L20 from the prespecified eligible range L0–20, excluding terminal readout layers. The selected layers are therefore not evidence for a unique peak or the onset of information. In 33 other scenarios, natural name-exchange

donors correct 7/33 original errors at L19 and 12/33 at L20; reverse transfers disrupt 14/33 and 26/33 initially correct counterfactual choices. Residual effects exceed same-norm random perturbations in the prespecified directions. Unlike the same-input bridge, these donors arise from text edits, and whole-state transfers can carry mixed information. They cannot be concatenated into a single proven serial pathway.

What remains unresolved. The experiments in this section establish causal effects of specified history edits, source contributions, and downstream states in the adapted A/B task. The competition tests below provide a local amplitude account; Section 4 and Appendix E investigate its upstream origin. The localization and source-transfer tests do not separately identify query/key matching or establish all source-to-head-to-position-to-module edges. Nor does a successful A/B intervention establish the same computation in unrestricted generation.

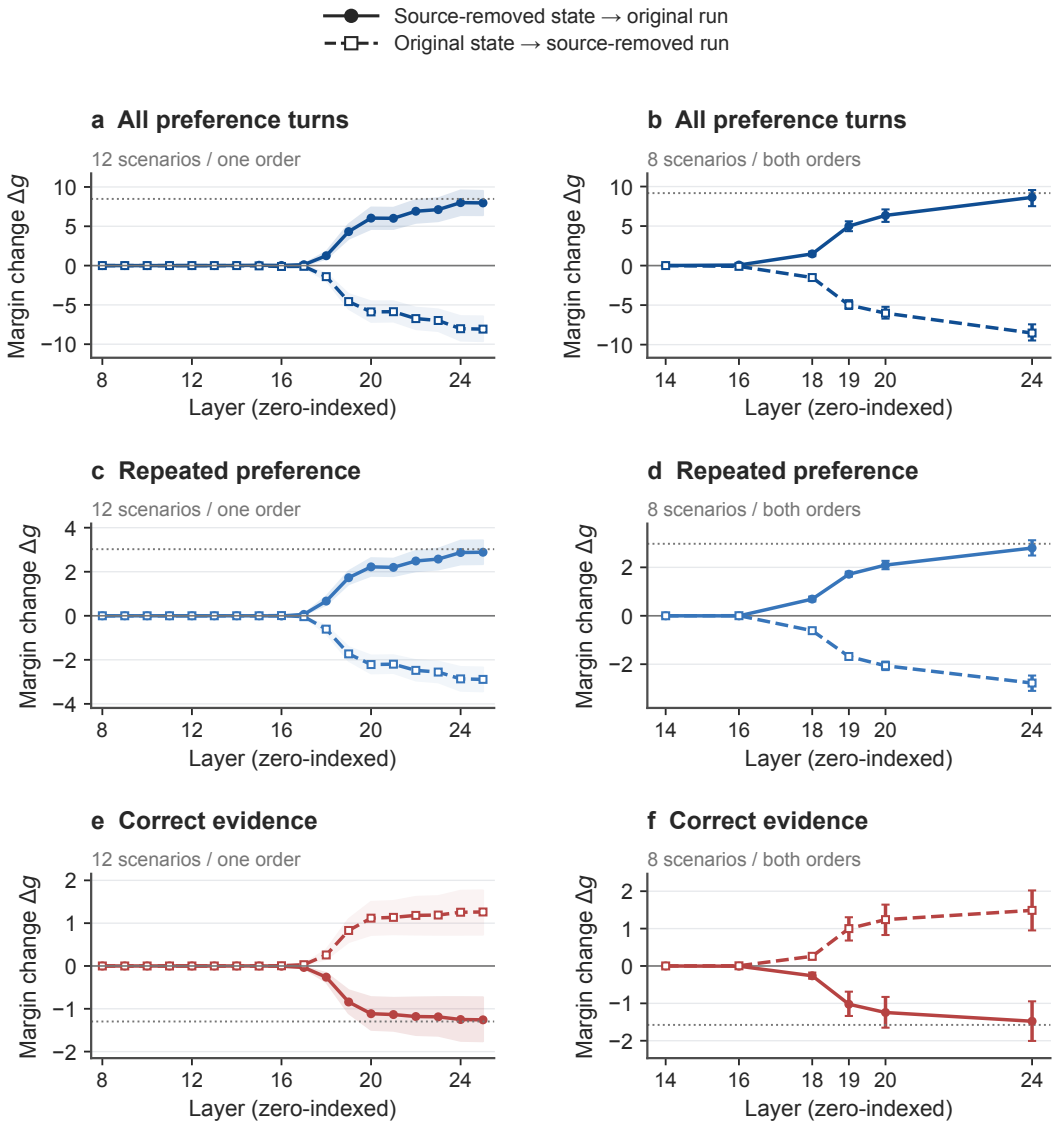


Figure 17: **Complete measured residual-transfer scans.** Left: L8–25 on 12 discovery scenarios in one option order. Right: six selected layers on eight follow-up scenarios in both orders; connecting lines join measured points only. Each intervention replaces the final prompt token's block-output residual at one layer. Solid and dashed curves retain forward and reverse raw margin changes, respectively. Dotted lines give the full source-removal effect for the corresponding source and sample. Bands and error bars show descriptive 95% scenario-bootstrap intervals. Axis scales vary across panels.

D.1 Competition and State-Change Amplitude

Samples and factorial intervention. Both experiments use 12 discovery scenarios in one option order and eight follow-up scenarios in both orders, yielding 28 scenario-order observations from 20 scenarios. Let B denote the original run, P preference removal, E evidence removal, and PE joint removal. The input and historical KV cache are unchanged. At L14/L16, for current-question and later queries, attenuation replaces the pre-projection head output by

$$z' = z - a_P c_P - a_E c_E, \qquad a_P, a_E \in \{0, .25, .5, .75, 1\}. \tag{2}$$

Here c_P and c_E are the active run's weighted value sums for the preference bundle and correct evidence, and a is the fraction removed. Weights are not renormalized at that site; subsequent

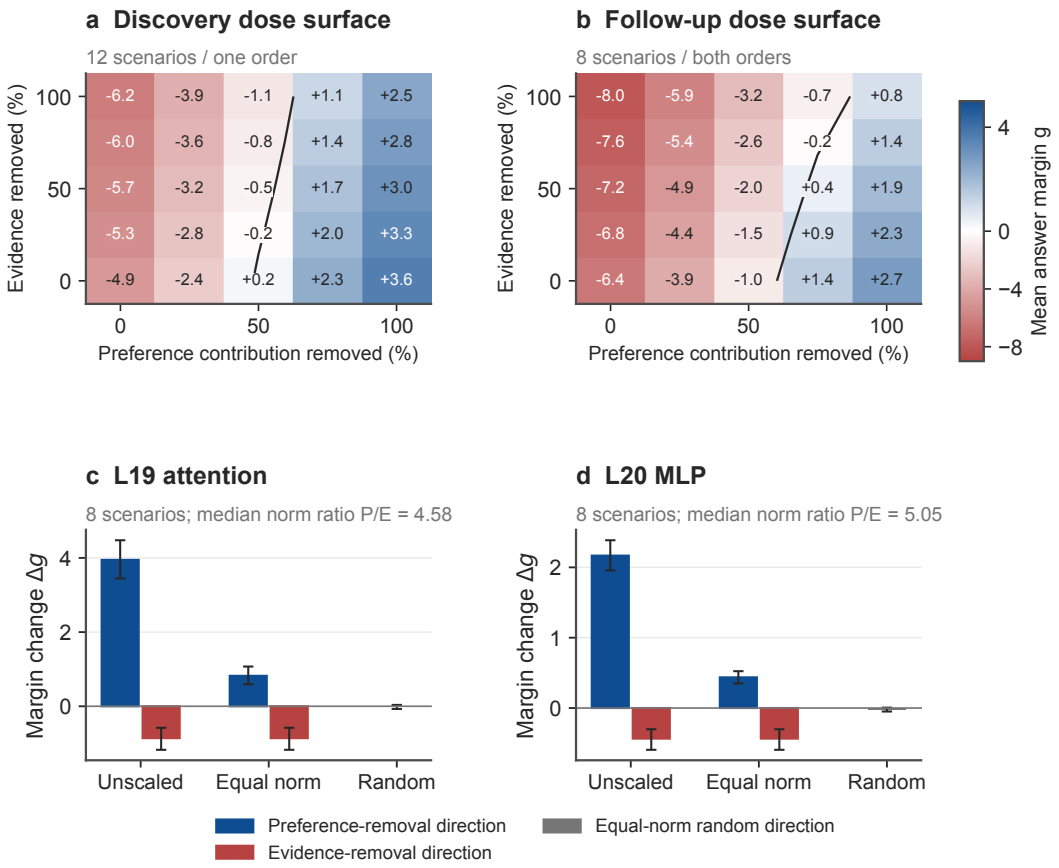


Figure 18: **Preference advantage is sensitive to source strength and local state-change amplitude.** (a–b) Mean correct-minus-alternative margins over a 5×5 source-attenuation grid; black contours interpolate the mean zero-margin boundary. (c–d) Local patches on eight follow-up scenarios, averaging both orders. Blue/red bars apply preference-removal/evidence-removal directions to the original state; equal-norm patches use the smaller paired L2 norm. Random directions use that same norm. Bars retain signed effects; error bars are descriptive 95% scenario-bootstrap intervals.

computation is recomputed. The 2×2 experiment uses the four endpoint conditions. Evidence benefits are $g_B - g_E$ with preference and $g_P - g_{PE}$ without it. Their difference,

$$I = (g_P - g_{PE}) - (g_B - g_E), \tag{3}$$

is positive if evidence's marginal benefit increases after preference removal. Its sign differs between the two samples (Table 10). The inconsistent interaction signs do not support a general evidence-suppression mechanism.

Table 10: Factorial source removal. Entries are scenario-mean margins, with both orders averaged in follow-up. I is the paired interaction with a descriptive 95% bootstrap interval.

Sample	B	P	E	PE	I [95% CI]
Discovery (12)	-4.92	$+3.56$	-6.22	$+2.50$	-0.23 $[-0.60, 0.14]$
Follow-up (8)	-6.43	$+2.74$	-8.01	$+0.82$	$+0.34$ $[0.01, 0.62]$

Evidence has a positive endpoint marginal effect in 26/28 scenario-order observations; preference removal improves the margin more than evidence removal worsens it in all 28. Follow-up both-order corrections are 0/8 for B, 3/8 for P, and 1/8 for PE. These conditional effects are not universal semantic weights. Evidence-state transfers compare $B \leftrightarrow E$ with $P \leftrightarrow PE$ at L14/L16/L18/L19/L20/L24 residuals (final token and candidate names), and final-token L19 attention, L20 attention, and L20 MLP. No consistently enhanced evidence-transfer node emerges: for

example, the conditional difference in L19 attention is -0.26 in discovery versus $+0.18$ in follow-up. Individual site effects overlap and cannot be added.

Equal-norm state interventions. For each scenario, order, and site, cache $h_B, h_P, h_E \in \mathbb{R}^{3584}$ at the final prompt token. Sites are L18/L20 block-output residuals, L19 post-projection attention output, and L20 MLP output. Define

$$d_s = h_s - h_B, \quad r = \min(\|d_P\|_2, \|d_E\|_2), \quad h'_s = h_B + \frac{r}{\|d_s\|_2} d_s, \quad s \in \{P, E\}. \tag{4}$$

Run the original input with only this local state replaced. Both directions are scaled down or left unchanged, never amplified beyond their measured source-removal differences. Separate unscaled and half-scaled patches use $h_B + d_s$ and $h_B + 0.5d_s$. A deterministic Gaussian random direction per scenario/order/site is normalized to r and added to h_B as a control; this is not an exhaustive random-direction null distribution. We compare downstream margin changes, not merely vector similarity.

Table 11 reports the main local results and the L20 residual check. The norm ratio is the median over scenario-order pairs; effects first average orders within each scenario. For L19 attention, the paired equal-norm imbalance $\Delta g_P + \Delta g_E$ is -0.019 ($[-0.116, 0.078]$) in discovery and -0.039 ($[-0.105, 0.020]$) in follow-up. L20 MLP gives $+0.005$ ($[-0.095, 0.098]$) and $+0.001$ ($[-0.095, 0.068]$). These descriptive intervals do not establish equivalence. Intervals use percentile bootstrapping over scenarios with 10,000 resamples and random seed 20260926.

Table 11: Local amplitude controls. P/E columns give signed margin changes for preference-removal and evidence-removal directions. Norms are intervention-induced state differences, not norms of isolated preference/evidence representations.

Sample	Site	Norm ratio	Unscaled P/E	Equal-norm P/E
Discovery	L19 attention	4.05	$+3.35/-0.75$	$+0.73/-0.75$
	L20 MLP	5.20	$+2.25/-0.40$	$+0.41/-0.40$
	L20 residual	5.11	$+6.04/-1.11$	$+1.06/-1.11$
Follow-up	L19 attention	4.58	$+3.95/-0.86$	$+0.82/-0.86$
	L20 MLP	5.05	$+2.17/-0.44$	$+0.44/-0.44$
	L20 residual	5.37	$+6.35/-1.24$	$+1.23/-1.24$

Joint unscaled L19 attention/L20 MLP replacement yields preference-removal effects of $+4.43/+4.80$ in discovery/follow-up, and evidence-removal effects of $-0.86/-0.99$. These are partial, non-additive effects. Source lengths and token counts are not matched between preference and evidence, so the upstream amplitude disparity may involve repetition, reading strength, cancellation between heads, or downstream transformations. Neither norm matching nor the source-attenuation surface establishes globally linear competition or identifies its upstream origin.

Numerical audit. Baseline and single-source margins agree between the factorial and source-removal experiments within 3.06×10^{-5}; all four attenuation-grid corners match the corresponding factorial conditions exactly. FP32 inference is used throughout. These experiments concern selected A/B errors, not unmodified agent behavior or population-level error rates.

E Amplitude Formation: Protocols and Supporting Results

Sample, aggregation, and uncertainty. All experiments use Qwen2.5-7B-Instruct in FP32. The context, factor-matching, and pathway experiments use 12 discovery scenarios in one option order and eight follow-up scenarios in both orders. The full layer scan uses these 20 scenarios in one option order and is exploratory rather than held out. For two-order experiments, we average orders within each scenario, then weight scenarios equally. Percentage changes are computed against the matched baseline within each observation before aggregation. Mean per-observation amplitude ratios therefore differ from ratios of mean norms. Reported intervals are 95% scenario-bootstrap intervals (4,000 resamples for context and route experiments; 1,000 for the dense scan), not simultaneous intervals corrected for all exploratory comparisons. These diagnostic scenarios do not support estimates of benchmark-wide prevalence.

E.1 Text Contrasts and Branch-Aligned Transfers

At the final prompt token's L19 attention or L20 MLP output, we measure $A_s = \|h^{(-s)} - h^{(B)}\|_2$, where B is the baseline run and $-s$ removes source s's weighted-value contributions at L14/L16. This is a source-induced state difference, not an isolated semantic-vector norm. Each edited input has its own baseline and source-removal runs; deleting a preference turn also changes the remaining preference source.

Preference P includes the initial preference, assistant acknowledgment, and repeated preference; evidence E is the correct evidence region. Deletion experiments remove one preference turn and use the unedited history as reference. Scope experiments prepend "For the current project:" to each of the three turns; the counterfactual replaces one or all prefixes with "For a separate project:". These pairs retain the underlying turn content and candidate names and have equal total token lengths. They are not the same intervention as the broader other-project rewrite described in Appendix D. Deletion changes length and positions; it does not isolate a single semantic attribute.

Table 12: Text-induced changes in source amplitudes (percent), eight follow-up scenarios. The scope row uses the current-project reference; deletion rows use the original history.

	L19 attention		L20 MLP	
Edit	A_P	A_E	A_P	A_E
Delete initial preference	$+15.8$	$+12.3$	$+13.8$	$+9.9$
Delete assistant feedback	-9.1	$+11.8$	-8.7	$+7.2$
Delete repeated preference	-31.2	$+33.6$	-29.9	$+28.6$
All three turns: other project	-15.7	$+8.4$	-14.7	$+3.3$

The repeated-preference deletion moves the mean margin from -6.43 to -0.97. Changing only that turn's scope reduces A_P by 12.2%/11.2%, compared with 15.7%/14.7% when changing all three turns. These effects are not additive. In the latter contrast, mean margin moves from -5.01 to -2.61, and mean per-observation A_P/A_E remains 3.34/3.51; every follow-up scenario's order-averaged ratio remains above one. Context affects source use, but these results do not establish an ordering of semantic matching and scope enforcement. The retained text may still contain relevant convenience or migration-cost information; not all remaining influence is automatically inappropriate.

Patch tensors and alignment. Each input has three runs: baseline B, preference removed P, and evidence removed E. Source removal subtracts the active run's weighted-value contributions at L14/L16, without softmax renormalization. A text donor is patched into its matching branch: $B \to B$, $P \to P$, $E \to E$. Transfers replace a tensor of shape $1 \times T_q \times 3584$, where T_q covers the identical current-question suffix, choice instructions, options, and assistant prefix. The implementation checks exact suffix token equality; historical absolute positions need not match. Historical KV caches remain those of the target input. Residual transfers replace block outputs at L8/12/14/16/18. Module transfers replace post-projection attention or MLP outputs at L14/16/18. Both final-token endpoints are then measured anew. This is not a single patch to B with unchanged P/E comparisons.

L16 attention transfers repeated-deletion reductions of 20.2% ($[-23.3, -17.3]$) at L19 and 18.6% ($[-21.2, -16.2]$) at L20 in the original eight-scenario follow-up. The full scan tests residual, attention, and MLP outputs at every layer from L1 to L18 on all 20 scenarios in one option order. Figure 19 shows the dense L20-MLP endpoint; the corresponding L19-attention scan is in the main text.

E.2 Factor Matching and Serial Transmission

For one head and query, the decomposition used in the main text is

$$m_s = \sum_{j \in s} \alpha_j, \qquad u_s = \frac{\sum_{j \in s} \alpha_j v_j}{m_s}, \qquad c_s = m_s u_s. \tag{5}$$

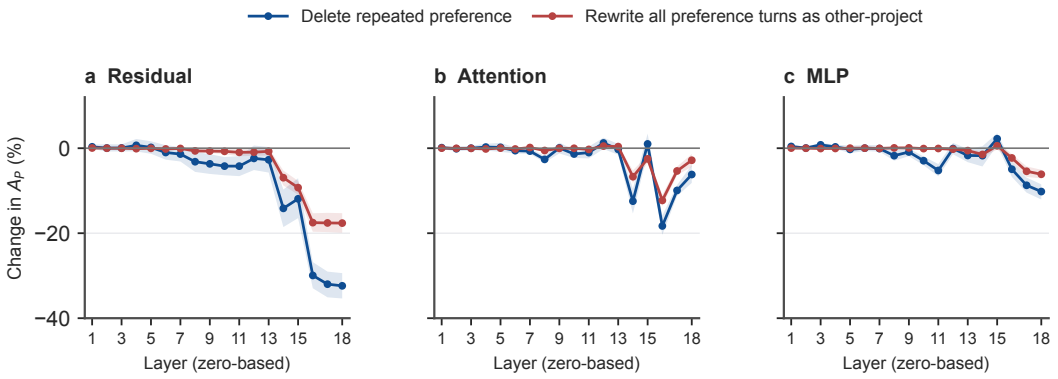


Figure 19: **L20-MLP endpoint across the L1–L18 transfer scan.** Blue: repeated-preference deletion; red: other-project rewrite. Curves show preference-amplitude change after residual, attention, or MLP transfer. Bands are 95% scenario-bootstrap intervals over 20 scenarios in one option order.

Here α_j is the attention weight and $v_j \in \mathbb{R}^{128}$ the value vector. Cumulative mass sums over source tokens, not over layers or time. Mean preference/evidence masses at L14 and L16 are 0.098/0.005 and 0.084/0.006, respectively. These descriptive differences alone do not establish a per-token bias.

For source s, head h, and query q, $c_s = m_s u_s \in \mathbb{R}^{128}$ is the source-weighted value sum before the output projection. Let $m_* = \min(m_P, m_E)$ and $n_* = \min(\|u_P\|_2, \|u_E\|_2)$. Mass matching replaces c_s by $(m_*/m_s)c_s$; content-norm matching replaces it by $(n_*/\|u_s\|_2)c_s$; joint matching multiplies both factors. Denominators are clamped at 10^{-12}. Non-source contributions remain unchanged at that site. P/E removal branches subtract the *adjusted* contributions, so amplitudes are measured relative to the factor-matched baseline. These interventions are performed per head and query, not on globally averaged masses. They are output-term attenuations, not naturally normalized Q/K edits.

Table 13: Single-layer versus joint-layer mass matching, and the content-norm control. Changes are relative to original computation; mean original margin is -6.431. Eight follow-up scenarios.

Operation	L19 A_P change (%)	L20 A_P change (%)
L14 mass matching	-61.1	-60.4
L16 mass matching	-38.2	-34.3
L14/L16 mass matching	-89.6	-90.1
L14/L16 content-norm matching	-1.7	-0.3

Joint mass matching also reduces evidence amplitudes by 25.1%/27.4%, and mean per-observation amplitude ratios from 5.05/5.00 to 0.77/0.76. Content-norm matching leaves mean margin at -6.959, versus -0.072 for mass matching. The two source regions differ in length (about 279 versus 28 tokens); these experiments do not separate region size from mean per-token attention. Mean conditional content norms are smaller for preference (5.063/4.363 at L14/L16) than for evidence (6.385/5.479). Within-source and cross-head directional coherence is not greater for preference either. None of these averaged diagnostics can be multiplied to reconstruct an aggregate model effect, and content direction remains intact in the matching interventions.

Pathway selection and sequential transfers. The L16 branch is a prespecified focus from the text-transfer experiments, not the best layer under every intervention: L14 mass matching has the larger single-layer amplitude effect. Within L16, discovery selects between mass and content-norm matching by mean reduction in $\log(A_P/A_E)$ averaged across both endpoints. The selected mass intervention supplies three branch-aligned donor runs. Each serial step transfers only the preceding step's actual state into fresh original-input $B/P/E$ target runs: L16 block output at all current-query positions, then L19 attention output at those positions, then L20 MLP output at the final token. Natural intervening computation is retained. Thus the named route is not a claim of adjacent architectural edges.

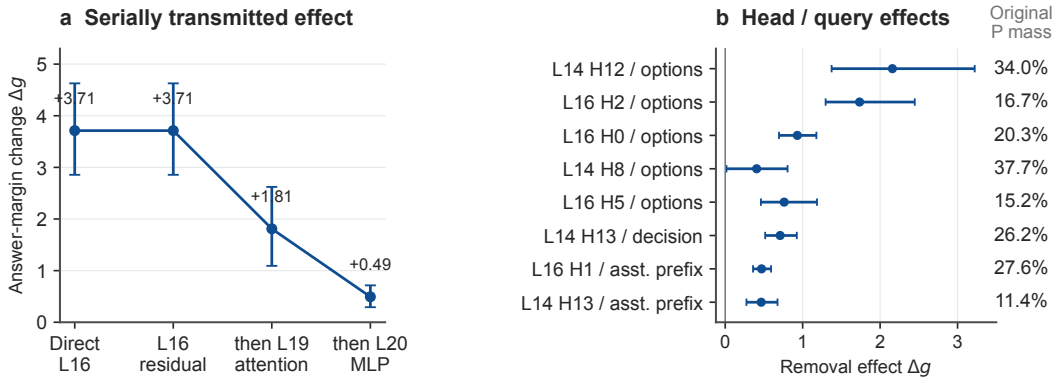


Figure 20: **Sequential state transfers and query-position effects.** (a) Successive serial transfers from an L16 mass-matched run; each point reports the resulting answer-margin change, not a module's independent contribution. L16 residual/L19 attention transfers cover all current-query positions; L20 MLP uses the final token. (b) All eight discovery-selected head–position combinations, tested by single-combination preference removal. The right column reports original attention mass, not an intervention effect. Both panels use eight follow-up scenarios and 95% confidence intervals. The selected heads have not been serially linked to the route in (a).

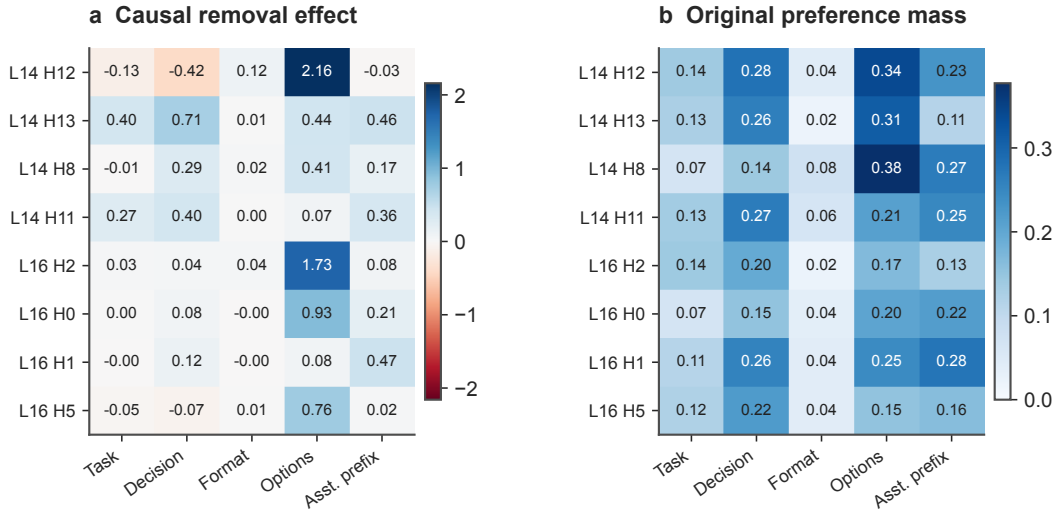


Figure 21: **Causal importance differs from attention magnitude.** All 40 combinations of discovery-selected heads and predefined current-query groups, measured on the eight follow-up scenarios. Left: answer-margin change after preference-source removal. Right: original attention mass on the preference region. Cells are scenario means; each intervention is separate and effects cannot be summed.

Direct intervention and L16-residual transfer both yield $\Delta g = 3.712$. Passing only the resulting L19 attention state yields 1.810; passing only its resulting L20 MLP final-token state yields 0.493 ($[0.292, 0.715]$). The last transfer also retains a 16.0% reduction in L20 A_P. Discovery has the same sign, with final effect 0.532. In a complementary knockout, fixing P/E runs' L19 attention outputs to their matching B outputs, while retaining residual bypasses, reduces L20 A_P/A_E amplitudes by 30.8%/32.5%. L19 therefore carries both source effects, not an exclusively preference-amplifying signal. The serial path does not account for the entire direct intervention.

E.3 Head and Query-Position Localization

Head selection on discovery scenarios. At L14 and L16 we scan all 28 query heads by removing only one head's preference-source contribution across all current-query positions. On discovery data, eligible heads must have positive mean margin improvement and positive mean projection of the L20 state change onto that layer's all-head preference-removal state change. Projection fraction is $\langle d_h, d_{\text{all}} \rangle / \|d_{\text{all}}\|_2^2$, not cosine similarity or output recovery. The four highest-margin eligible heads per layer are retained: L14H12/H13/H8/H11 and L16H2/H0/H1/H5. Follow-up single-head effects are all positive, but L16H5's mean L20 projection is -0.02; it is not equally strong evidence for the same downstream amplitude direction.

Current-query positions are divided into five fixed, nonoverlapping groups: task description (about 28 tokens), decision wording (41), answer-format instruction (11), candidate options (10), and assistant prefix including the final prompt token (5). For the selected heads, all 40 head–group combinations are tested. Eight combinations are selected using discovery margin and projection criteria; Figure 20b shows all eight, rather than choosing combinations by follow-up effect. Group interventions do not identify the most important individual token.

Intervention versus description. For the chosen head and query group, we subtract $\sum_{j \in P} \alpha_{h,q,j} v_{k(h),j}$ from the 128-dimensional pre-projection head output. We neither remove the whole head nor renormalize attention. Other queries and heads are unchanged at the intervention site. Figure 21 reports all 40 follow-up effects and their original masses. L14H12/L16H2 option-position effects are 2.16 ($[1.37, 3.22]$) and 1.73 ($[1.30, 2.45]$); their original weighted-V norms are 2.39/1.10. L14H8 at options has greater mass and weighted-V norm (37.7%, 2.77), but a smaller effect, 0.41. Direction, projection, and downstream use matter in addition to mass. Single-combination effects are not additive, and no joint coverage or single-head-to-downstream serial circuit is established here.

Numerical checks and scope. Unmodified-run and self-patch amplitude errors are zero for context transfers and at most 3.34×10^{-5} for route tests; headwise output-projection reconstruction error is at most 8.94×10^{-8}, and full-source-removal reproduction error in head localization is 2.67×10^{-5}. All are below the prespecified 0.002 tolerance. The findings concern history-sensitive errors in the adapted A/B task, rather than unrestricted generation.

F Task Content and Scope: Supplementary Analyses

Text design and limitations. The 18-condition design uses the eight follow-up scenarios in both option orders (288 responses). These scenarios are shared with the source-transfer analysis rather than held out independently. Scope prefixes are token-length matched. Task and comparison-cue rewrites vary wording and slightly change positions. Orders are averaged within scenarios before descriptive bootstrap intervals are computed; intervals are not corrected for multiple comparisons. Unrelated-task conditions retain software candidates and are attention diagnostics, not valid accuracy tests. Alternative scope phrasings also yield head- and wording-dependent effects; a near-zero average effect does not demonstrate absence of scope representations.

Historical-region and lexical controls. Replacing two preference turns with approximately length-matched unrelated text leaves 19.14% attention when only RP is retained, 15.12% for AF, and 10.44% for IP, versus 17.10% originally and 2.96% when all are replaced. These values include attention to replacement slots and are not additive contributions. With a controlled question template, exact task-category wording yields 21.28%, a semantic paraphrase 15.83%, an unrelated task 7.74%, and an unrelated task mentioning the original category under negation 8.41%. The last contrast is $+0.67$ percentage points with a 95% CI $[-0.35, 1.74]$. These templates differ from the factorial design and must not be treated as paired baselines across experiments.

Single-case module protocol. The module study uses `emc_search_000070`; two option orders are not independent scenarios. Six task variants have identical token counts and option positions. Historical K is identical across each pair, Q+current-K replacement reproduces the changed-input readout, and V replacement leaves same-layer attention mass unchanged. Forward patches insert

changed-input states into the original run; reverse patches restore original states. Module freezing retains the L7 task intervention while resetting specified option-position module outputs to the original run.

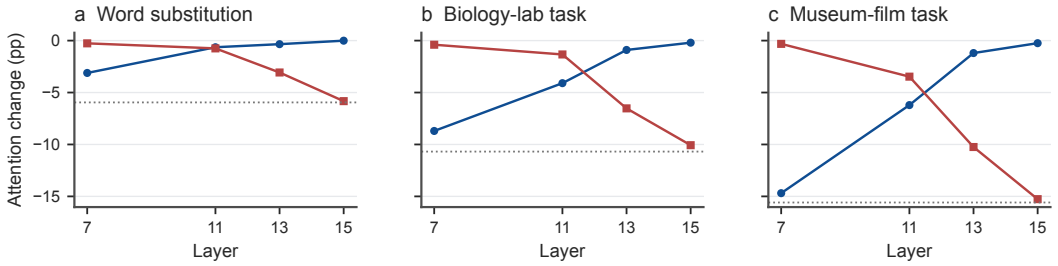


Figure 22: Task-position (blue) versus option-position (red) residual transfers; dotted lines mark direct text effects. One scenario, both orders. Each point is an independent single-site intervention, not a serial path.

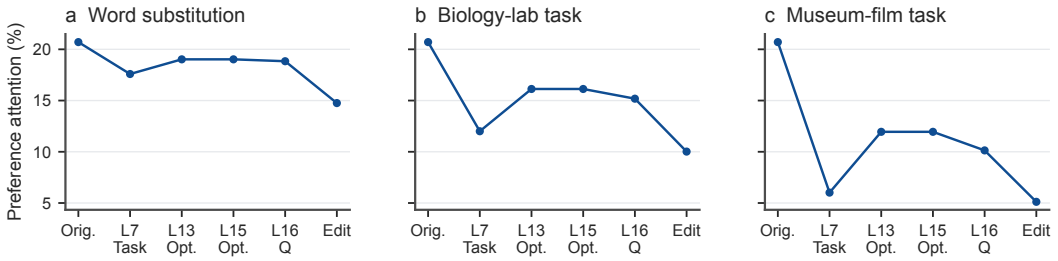


Figure 23: Sequential state transfer through L7 task residual, L13/L15 option residuals, and L16 option Q. Each donor is generated by the preceding patched run. Original and changed inputs are references, not relay stages. Only part of each direct text effect survives.

For word, biology-lab, and museum contrasts, sequential transfer yields at 18.84%, 15.18%, and 10.14%, compared with original 20.70% and direct-edit values 14.75%, 10.02%, and 5.12%. Freezing L8–15 option attention restores reading to 20.67%, 20.28%, and 20.41%; freezing MLP outputs instead gives 16.84%, 13.32%, and 9.05%. This does not imply MLPs are irrelevant. Grouped interventions may create unnatural states; these tests do not establish a complete causal path to final answers or establish semantic-versus-scope processing order.

F.1 Early Content and Scope Competition

Two complementary tests use the eight follow-up scenarios with both option orders. Results are averaged within scenario before 10,000-resample scenario bootstrapping. In the first, three equal-length preference prefixes specify the current versus a separate project, holding preference bodies, current question, evidence, and option positions fixed. Separate-project scope improves the answer margin by 2.399 and reduces the preference-removal effect from 7.976 to 6.121 logits; the evidence-removal effect changes from -1.752 to -1.940. Preference therefore remains influential even when its scope is restricted. L7 historical-state patches transfer 2.348 logits of the scope effect; current-position patches transfer 0.109 at L13 and 1.204 at L15. These interventions identify sites that transmit the scope effect, but do not establish when scope is understood.

The second follow-up changes only the wrong-candidate name in repeated preference to create a weaker-content donor under each scope condition. At L7, we interpolate its contextualized preference-body residuals into the original run with $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. Scope-prefix states are not patched. This changes a contextual state, not an isolated name feature. At each strength, the scope contrast is separate minus current project.

The option-Q intervention replaces only L16H2 queries at candidate positions from the separate-project run into the current-project run at the same α. Its final-margin effect is near zero. L15

Table 14: L7 content intervention: scope-sensitive reading versus decision correction.

Readout	$\alpha = 0$	$\alpha = 1$	Change [95% CI]
Scope-induced attention reduction (pp)	1.26	2.18	$0.92\ [0.68, 1.18]$
Option-Q transfer: reduction (pp)	0.27	0.81	$0.55\ [0.41, 0.69]$
Scope-induced margin improvement	2.40	1.90	$-0.50\ [-0.88, -0.06]$

current-position residual swaps transfer 0.38/0.92 pp of reading reduction and 1.20/0.95 logits of margin improvement at $\alpha = 0/1$, respectively. Differences are calculated before rounding. Thus, greater scope sensitivity enters current states without increasing absolute decision correction. Scope and content may interact in reading, but these results do not establish a serial semantic-before-scope mechanism or its sufficiency for errors. Interpolation can create states that do not occur in unmodified runs, the scenarios are shared across analyses, and scope manipulations differ between preference and evidence.