HISTORY IS NOT STATE: CAUSAL DIAGNOSTICS FOR SUFFICIENT-STATE FAILURES IN LANGUAGE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

History records events; state captures what remains true. Endpoint accuracy cannot reveal whether a model can use correct state, whether that state repairs an error while conflicting history remains visible, or whether decisions respond selectively to state content. MarkovLens combines replay-verified equivalent histories with matched interventions on externally supplied task-sufficient states. In a Qwen2.5-7B-Instruct run, adding the correct state beside the unchanged history raises accuracy from 75% to 100%, repairing all 96 raw errors without harming any of 288 correct answers. Changing one decision-relevant field redirects 272/384 outputs (70.83%) toward the newly implied answer. By contrast, answer-preserving edits across ten fields produce a 2.66% group-cluster answer-change rate. The qualitative repair-and-selectivity pattern persists in full-BF16 Qwen, under semantically verified semi-natural dialogue, and in a cross-family DeepSeek replication. MarkovLens turns state into an intervention target rather than treating Gold-state accuracy as an endpoint.

1 INTRODUCTION

Consider a restaurant assistant. A user says, *Find an Italian restaurant in Boston*, then corrects the request: *Actually, make that Seattle*. The transcript contains both cities, but only Seattle is true now. Another user may ask for an Italian restaurant in Seattle from the start. The histories differ; the current state does not.

Language agents face this pattern whenever preferences are cancelled, requirements are overwritten, tools change an environment, or new evidence invalidates an earlier conclusion. A long transcript mixes what used to be true with what remains true. If two histories end at the same state, why should the route determine the answer?

A single accuracy score cannot locate where a state-related failure enters. Even Gold-state accuracy leaves an important ambiguity: did the model use the state content, or was it merely helped by a cleaner prompt? MarkovLens resolves this ambiguity by freezing the interaction and manipulating task-sufficient state in controlled counterfactual ways. It asks four increasingly informative questions:

1. **Screen:** Can the model answer from the correct current state?
2. **Repair:** Does that state still fix the decision when the contradictory history remains visible?
3. **Direction:** If only the supplied state changes so that another answer becomes correct, does the decision move with it?
4. **Selectivity:** If the state changes but the correct answer does not, does the decision mostly stay put?

The Seattle example makes the final question concrete. We keep the Boston-to-Seattle history, query, restaurants, instructions, state renderer, and state position fixed. Supplying *Destination: Seattle* tests the correct state. Replacing only Seattle with Boston tests whether the decision moves toward the Boston-consistent restaurant. Changing another state field while keeping the same restaurant optimal tests whether the model reacts to any edit or mainly to edits that change the task-implied decision. Our causal language refers to controlled interventions on externally supplied state representations and observable decisions, not identification of a unique internal model state.

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.

The experiments follow this sequence. A Raw-versus-Gold screen identifies candidate bottlenecks. In the primary matched Qwen2.5-7B study, correct state repairs every observed Raw error while the original history remains present. Decision-changing edits redirect 70.83% of outputs, whereas answer-preserving edits across ten fields change only 2.66% of decisions after group clustering. The same qualitative structure survives a full-BF16 execution, semantically verified semi-natural rewrites of the same underlying interactions, and a DeepSeek cross-family replication.

Our contributions are:

1. We make task-sufficient state an external intervention target for evolving interactions.
2. We construct replay-verified equivalent histories and matched state counterfactuals that separate the route, the supplied state content, and the resulting decision.
3. We introduce a decision-selectivity test: answer-changing edits strongly redirect decisions, while multi-field answer-preserving edits are far more stable in aggregate.
4. We show that the qualitative pattern persists in full-BF16 Qwen, under controlled semi-natural surface variation, and in a cross-family DeepSeek replication.

2 Related Work

Stateful agent evaluation. Multi-turn and agent benchmarks test changing goals, tools, databases, and applications (Laban et al., 2025; Liu et al., 2024; Ma et al., 2024; Lu et al., 2025; Yao et al., 2025; Trivedi et al., 2024; Drouin et al., 2024; Microsoft, 2026; Berkeley Function Calling Leaderboard Team, 2026). They measure whether an agent succeeds as an interaction evolves. MarkovLens instead freezes the interaction and makes the current state itself a controlled intervention target.

Dialogue state, memory, and belief revision. Dialogue-state tracking, memory, and belief-revision work builds representations of active goals and constraints (Budzianowski et al., 2018; Mrkšić et al., 2017; Wu et al., 2019; Rastogi et al., 2020; Heck et al., 2020; Das et al., 2024; Carranza & Rojas, 2025; Wilie et al., 2024; Zhong et al., 2024; Rozanov & Rei, 2025). MarkovLens does not propose another tracker. It uses known states and matched counterfactuals to ask how observable decisions respond to supplied state content.

State sufficiency and history condensation. Adequate state representations are classical in partially observable decision processes, reinforcement learning, and state abstraction (Kaelbling et al., 1998; Sutton & Barto, 2018; Li et al., 2006). Prompt compression and history condensation seek representations that retain what matters for later decisions (Li et al., 2023; Jiang et al., 2024; Singh et al., 2026). MarkovLens operationalizes this idea behaviorally: equivalent routes should support the same decision, answer-changing state edits should redirect it, and answer-preserving edits should usually leave it stable.

3 MarkovLens

3.1 From Histories to Controlled State Interventions

Figure 1 shows the logic before the notation. Two conversations can reach the same current state. We then freeze the history, query, candidates, renderer, and instructions, and change only the supplied state: correct, answer-preserving, or decision-changing.

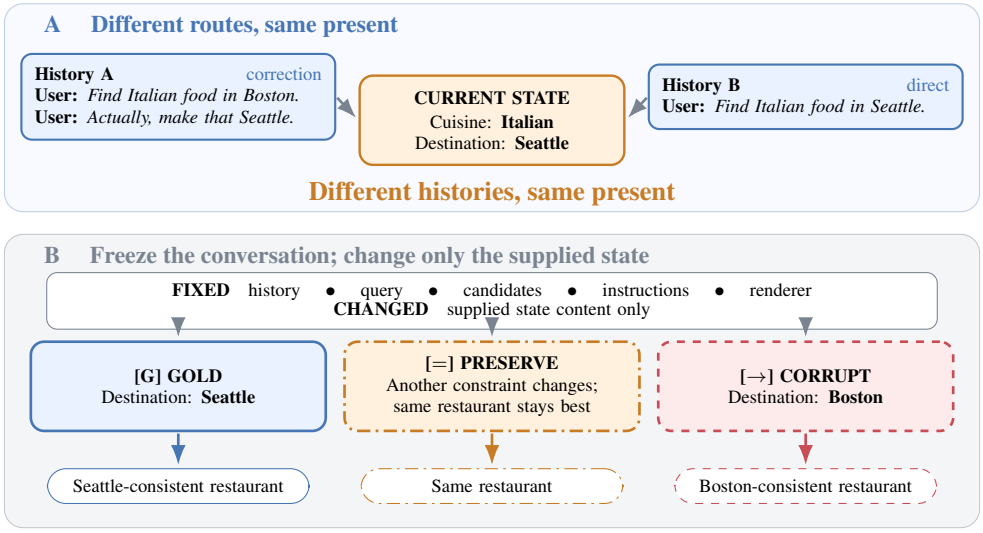

Figure 1: History records the route; state records what is true now. Panel A shows two replay-verified histories reaching the same present. Panel B freezes the surrounding interaction and changes only the supplied state. GOLD supplies the correct state. PRESERVE changes the state but keeps the correct restaurant fixed. CORRUPT changes the state so that another restaurant becomes correct. Symbols and border styles keep the conditions distinct in grayscale.

Let $H_t = (o_1, \ldots, o_t)$ be the observed history, q_t the current query, and C_t the decision context, such as restaurant candidates, unit tests, or available tool actions. We call S_t task-sufficient when it contains the information that is still valid and needed for the decision. Once S_t, q_t, and C_t are known, the route recorded in H_t should add nothing:

$$Y_t \perp H_t \mid S_t, q_t, C_t. \tag{1}$$

This is an operational property of our controlled tasks, not a claim that every real conversation has one unique state. The transition system computes S_t, and a deterministic task rule $y_t = f(S_t, q_t, C_t)$ computes the answer. Replay checks verify equivalent histories. Non-leakage checks reject states containing answer strings, candidate identifiers, rankings, generated code, or final actions. Candidate perturbation keeps the state fixed while changing the options, so the state cannot simply reveal an answer identifier.

3.2 From Screen to Decision Selectivity

1. Screen: can the model decide once the state is given? The **sufficient-state gap (SSG)** compares Raw-history accuracy with Gold-state-only accuracy:

$$\mathrm{SSG} = \mathrm{Acc}(S_t, q_t, C_t) - \mathrm{Acc}(H_t, q_t, C_t). \tag{2}$$

A positive gap identifies a candidate history-to-state bottleneck; a negative gap can expose difficulty using the state interface. Raw and Gold also differ in presentation, so SSG is a screen rather than a causal decomposition.

2. Repair: does correct state help while history remains visible? We add the Gold state beside the unchanged Raw history:

$$\hat{y}^H = M(H_t, q_t, C_t), \qquad \hat{y}^{H,S} = M(H_t, S_t, q_t, C_t). \tag{3}$$

Repair counts Raw errors corrected by the intervention; harm counts Raw-correct decisions made wrong. In the restaurant example, both Boston and Seattle remain in the transcript, but the explicit current state says Seattle. This matched comparison is stronger than removing or replacing the history.

3. Selectivity: does the decision follow what the state implies? Starting from the same history-plus-state prompt, we change only the supplied state content. A decision-changing **Corrupt** intervention satisfies

$$f(\widetilde{S}_t, q_t, C_t) \neq f(S_t, q_t, C_t). \tag{4}$$

A directional hit occurs when the output moves to the answer implied by $\widetilde{S}_t$. An answer-preserving **Preserve** intervention instead satisfies

$$\widetilde{S}_t \neq S_t, \qquad f(\widetilde{S}_t, q_t, C_t) = f(S_t, q_t, C_t). \tag{5}$$

Corrupt changes Seattle to Boston and therefore changes the correct restaurant. Preserve changes a state constraint while keeping the same restaurant optimal. The contrast tests whether decisions react selectively to their task consequence rather than to any textual edit.

Supporting diagnostics. Equivalent-history groups expose route dependence. We report pairwise disagreement ($\mathrm{MIV}_{\mathrm{pair}}$) and the fraction of groups containing any disagreement ($\mathrm{MIV}_{\mathrm{group}}$). Obsolete-state attraction (OSA) measures calibrated overlap between an error and choices that were correct earlier. State update error ($\mathrm{SUE} = 1 - \mathrm{F1}_{\mathrm{slot}}$) audits explicit state extraction. These measures support the intervention sequence; none substitutes for it.

3.3 Controlled Environment

MarkovLens starts from an initial state, applies labeled updates, renders those updates in natural language, and computes the final answer from the resulting state and decision context. It then creates several routes to that same endpoint: a direct request, an early mistake followed by a correction, a conflict, a cancellation, or an irrelevant detour. Every route is replay-verified.

We study three task families. **User Constraint Update** changes preferences, locations, cancellations, and uncertainty. **Code Requirement Drift** changes software specifications and uses executable-style decision contexts. **Tool/Environment Mutation** changes external resources after tool calls, failed actions, or corrections. Appendix D gives the generation algorithm.

4 Experimental Design

Evaluation sets. User Constraint Update contains 96 final-state groups and 384 histories, with four distinct routes to each state. Six update buckets cover retention, overwrite, deletion, uncertainty, conflict, and mixed changes. The canonical Raw-versus-Gold screen uses this grid; Code Requirement Drift supplies a cross-task screen, and Tool/Environment Mutation probes state rendering. All equivalent histories are replay-verified.

Primary matched study. The mechanism study uses Qwen2.5-7B-Instruct with local Ollama CUDA inference, GGUF Q4_K_M weights, deterministic decoding, and the same 96-group/384-history structure. RAW_GOLD adds the correct state beside RAW. RAW_CORRUPT changes exactly one decision-relevant field; all 96 group-level changes are mechanically verified to change the correct answer. The multi-field Preserve study covers ten field paths, 400 group–field combinations, and 1,600 histories. Every Preserve edit changes the state while mechanically preserving the deterministic answer, and no model output is used to choose an edit. Across matched state-content conditions, the history, query, candidates, instructions, renderer, and state placement remain fixed.

Inference and robustness. Confidence intervals resample final-state groups, and paired tests operate on group-level outcomes. Invalid outputs count as incorrect; valid-only rates are reported separately in the appendix. Simple truncation baselines change the history presentation and are therefore not matched interventions. A separate full-BF16 Qwen2.5-7B execution uses the frozen 96-group protocol. The same BF16 model also evaluates a stratified 48-group subset in both template form and semantically verified semi-natural dialogue, with latent transitions, final states, queries, candidates, and intervention objects held fixed. A DeepSeek-V4-Flash run provides cross-family evidence. These runs test qualitative robustness rather than quantization or capability differences. Appendix A gives construction details, configurations, and denominators.

5 RESULTS

5.1 A GOLD-STATE SCREEN FINDS CANDIDATE BOTTLENECKS

The screen separates candidate bottlenecks (Table 1). On User Constraint Update, Qwen rises from 0.833 with Raw to 1.000 with Gold. Mistral moves from 0.917 to 0.833, yet reaches 1.000 when the same gold content is rendered as dialogue on an interface-equalized slice. The state may contain the answer-relevant information and still be difficult for a model to use. Because Raw and Gold differ in presentation, this stage motivates rather than replaces the matched intervention.

Table 1: Canonical Raw-versus-Gold screen. CRD denotes Code Requirement Drift.

Task	Model	Raw	Gold	SSG [95% CI]
User constraints	`qwen2.5:7b`	.833	1.000	+.167 [.130,.206]
User constraints	`mistral:7b`	.917	.833	-.083 [-.151,-.023]
CRD	`qwen2.5:7b`	.755	1.000	+.245 [.135,.354]
CRD	`mistral:7b`	.708	.646	-.062 [-.214,.078]

5.2 CORRECT STATE REPAIRS HISTORY-CONDITIONED ERRORS

The Boston-to-Seattle transcript remains in the prompt; we simply add the correct current state, Seattle. RAW accuracy is 75.00%. Adding the state raises accuracy to 100.00%, a gain of 25.00 percentage points (95% group-bootstrap CI $[20.83, 29.17]$; paired randomization $p \approx 10^{-5}$). GOLD_ONLY also reaches 100.00%. Every one of the 96 raw errors is repaired, and none of the 288 raw-correct answers is harmed. Correct supplied state can therefore repair history-conditioned errors while the original history remains present.

Simply throwing away earlier context does not reproduce the repair: LAST_USER_TURN_ONLY reaches 8.33%, LAST_K=2 reaches 45.83%, and RAW_GOLD reaches 100.00%; the truncation conditions change history presentation and are therefore not matched interventions.

Pairwise disagreement falls from $\mathrm{MIV}_{\mathrm{pair}} = 0.4167$ to 0, and the fraction of groups with any disagreement falls from $\mathrm{MIV}_{\mathrm{group}} = 0.6667$ to 0. This is an important consistency consequence of complete repair: because RAW_GOLD is correct on every history, equivalent histories must agree. We therefore do not present zero MIV here as a separate mechanism finding.

5.3 COUNTERFACTUAL STATE EDITS REVEAL DECISION SELECTIVITY

Suppose the conversation ends with Seattle as the current destination. We keep that conversation, the restaurants, and the entire prompt unchanged. Changing the supplied destination from Seattle to Boston changes which restaurant should win. Changing another constraint while leaving the same restaurant optimal should not. This contrast asks whether the model follows decision-relevant state content or merely reacts to edited text.

When the one-field corruption changes the task-implied decision, Qwen follows the corruption-implied answer in 272/384 histories (70.83%). All 96 group-level corruptions are mechanically verified to be decision-effective. The paired effect is 0.7083, with 95% CI $[0.6380, 0.7734]$ and paired randomization $p \approx 10^{-5}$. Because the state block, its placement, and the surrounding prompt are identical, the result shows behavioral sensitivity to the information inside that block, not merely to the presence of structured text.

The multi-field Preserve condition provides the complementary test. It changes ten field paths over 400 group–field combinations and 1,600 histories while mechanically keeping the deterministic answer fixed. After clustering by final-state group, gold retention is 95.72% (95% CI $[94.81, 96.61]$), and the answer-change rate is 2.66% (95% CI $[2.11, 3.21]$). Behavior is not perfectly uniform across fields. The central contrast is nevertheless sharp: 70.83% directional following when the edit changes the correct answer versus 2.66% answer change when it does not. The response is decision selective in aggregate.

Table 2: Primary matched Qwen2.5-7B Q4 results. Preserve reports the group-cluster estimate across ten answer-preserving fields.

Question	Observed result	Uncertainty / calibration
Correct-state repair	75.00% → 100.00%; 96/96 repaired, 0/288 harmed	$[20.83, 29.17]$ pp
Decision-changing edit	272/384 = 70.83% directional hit	$[63.80, 77.34]\%$
Answer-preserving edits	2.66% group-cluster answer change	$[2.11, 3.21]\%$

5.4 The Pattern Survives Changes in Execution Stack and Surface Form

Full-BF16 execution. The pattern is not specific to the Q4_K_M execution stack. A separate full-BF16 Qwen2.5-7B run again yields 75.00% RAW and 100.00% RAW_GOLD accuracy, repairs all 96 RAW errors without harming any of 288 correct answers, redirects 304/384 predictions (79.17%) under decision-changing corruption, and changes none of 352 valid paired decisions under the answer-preserving intervention; 32 Preserve outputs are invalid. Precision and runtime differ together, so this is qualitative robustness evidence rather than an estimate of a quantization effect.

Controlled semi-natural surface robustness. We next keep the underlying task fixed while changing how its histories are expressed. DeepSeek-V4-Flash rewrites a stratified 48-group subset into more natural dialogue; all 192 rewrites pass conservative semantic-equivalence checks, and the same BF16 Qwen model evaluates both surfaces. RAW accuracy changes from 144/192 (75.00%) on the templates to 169/192 (88.02%) on the semi-natural histories, showing that wording changes difficulty. The intervention pattern remains recognizable: correct state repairs 23/23 semi-natural RAW errors while harming 1/169 correct cases, Corrupt produces the newly implied answer in 134/189 valid outputs (70.90%), and Preserve changes 1/184 valid paired decisions (0.54%). Thus the qualitative repair-and-selectivity pattern persists when the same underlying interactions are expressed through independently generated, semantically verified semi-natural dialogue.

5.5 Cross-Family Evidence

A DeepSeek-V4-Flash run on the frozen full protocol shows the same separation. RAW accuracy is 289/384 (75.26%) and RAW_GOLD is 383/384 (99.74%): 95/95 errors are repaired and 1/289 correct cases is harmed. Decision-changing corruption redirects 222/384 outputs (57.81%), while Preserve retains the gold answer in 383/384 cases and changes 1/383 valid paired decisions (0.26%). This is cross-family evidence for the qualitative contrast, not a model ranking.

5.6 State Rendering and Task Still Matter

The other task families make the same caution concrete: a state can be correct yet hard for a model to use. On Code Requirement Drift, Qwen has a positive SSG of $+0.245$, while Mistral's -0.062 gap is statistically unstable. On Tool/Environment Mutation, Qwen falls from 0.958 raw accuracy to 0.792 with compact gold state, then climbs to 1.000 when the same state is rendered as dialogue. Mistral reaches 0.875 with Raw, 0.917 with a JSONized event log, and 0.833 with either gold-state rendering. State content matters, but so do the container and the reader.

Appendix C collects the secondary controls. In-distribution supervised solvers confirm that the gold representation contains enough information to answer, while held-out variants and value shifts show that robust extraction remains unsolved. Complexity-controlled sweeps likewise associate failures with difficult updates rather than with a simple monotonic increase in transcript length.

6 Discussion and Limitations

Gold-state accuracy is only the first question. A stronger diagnosis keeps history visible and changes the supplied state counterfactually. Correct state can repair a history-conditioned error; a decision-changing state can redirect the output; and an answer-preserving state can test whether that response is selective. The Q4 study separates these properties cleanly. The qualitative pattern persists in a

full-BF16 execution, under semantically verified semi-natural rewrites evaluated by that same BF16 model, and in a different model family. Explicit state is therefore more than a helpful summary when the decision moves with what the state implies and remains comparatively stable when the implied decision is preserved.

The conclusion is behavioral. MarkovLens changes an external representation supplied in the prompt and observes the resulting decision. It neither reads nor edits an unobserved internal model state, proves that the model represents a unique latent state, nor turns every error into a state-construction failure. Rendering controls also show that correct content can be difficult to use: the state container and output interface remain part of the system being measured.

Controlled transitions make exact state, replay equivalence, and mechanically verified counterfactual answers possible. They also bound the claim. The semi-natural bridge shows that the pattern is not confined to the original templates, but it is controlled surface robustness rather than validation on real user logs or deployed agents. Natural interactions may contain implicit goals, uncertain evidence, social context, or several adequate state descriptions. Field-level Preserve behavior is heterogeneous, so the aggregate result is strong decision selectivity rather than universal invariance. Finally, Q4, BF16, and DeepSeek use different inference stacks; their numerical differences cannot be attributed to precision, backend, provider, scale, or capability in isolation.

7 Conclusion

History tells us how we got here. State tells us where we are. MarkovLens freezes the route and asks what happens when the supplied state is corrected, changed to imply another decision, or changed while preserving the same decision. In the Boston-to-Seattle example, that means keeping both cities in view while testing whether the model follows what is true now. Correct state repairs, decision-changing state redirects, and answer-preserving state usually leaves the decision unchanged. This turns state from a static benchmark annotation into an intervention target for diagnosing stateful language agents.

AI use statement

Generative AI tools were used to assist with language editing and manuscript organization. The authors reviewed the resulting text, scientific claims, citations, and numerical values and take responsibility for the final manuscript.

Reproducibility statement

Appendix A documents matched inference configurations, intervention conditions, semi-natural rewrite validation, denominators, and output categories. Appendix C reports secondary controls, and Appendix D describes episode generation, replay verification, non-leakage checks, and candidate perturbation. The supplementary materials contain the generation templates, prompts, parsing rules, and per-example outputs needed to reproduce the reported diagnostics.

References

Berkeley Function Calling Leaderboard Team. Berkeley function calling leaderboard v3: Multi-turn and multi-step function calling. https://gorilla.cs.berkeley.edu/blogs/13_bfcl_v3_multi_turn.html, 2026. Accessed 2026-05-25.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ: A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547/.

Rafael Carranza and Mateo Alejandro Rojas. Interpretable and robust dialogue state tracking via natural language summarization with LLMs. *arXiv preprint arXiv:2503.08857*, 2025. URL https://arxiv.org/abs/2503.08857.

Sarkar Snigdha Sarathi Das, Chirag Shah, Mengting Wan, Jennifer Neville, Longqi Yang, Reid Andersen, Georg Buscher, and Tara Safavi. S3-DST: Structured open-domain dialogue segmentation and state tracking in the era of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14996–15014, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.891. URL https://aclanthology.org/2024.findings-acl.891/.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. WorkArena: How capable are web agents at solving common knowledge work tasks? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11642–11662. PMLR, 2024. URL https://proceedings.mlr.press/v235/drouin24a.html.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 35–44, 1st virtual meeting, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigdial-1.4. URL https://aclanthology.org/2020.sigdial-1.4/.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1658–1677, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Leslie Packard Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

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

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for mdps. *Proceedings of the International Symposium on Artificial Intelligence and Mathematics*, 2006.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6342–6353, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.391. URL https://aclanthology.org/2023.emnlp-main.391/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. *arXiv preprint arXiv:2408.04682*, 2025.

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

Microsoft. State-bench: Stateful agent task evaluation benchmark. https://github.com/microsoft/STATE-Bench, 2026. Accessed 2026-05-25.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1777–1788, Vancouver,

Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1163. URL https://aclanthology.org/P17-1163/.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8689–8696, 2020. doi: 10.1609/aaai.v34i05.6394. URL https://arxiv.org/abs/1909.05855.

Nikolai Rozanov and Marek Rei. Stateact: Enhancing llm base agents via self-prompting and state-tracking. In *Proceedings of the First Workshop on Reasoning and Learning for Human-Machine Dialogues*, 2025.

Jyotika Singh, Fang Tu, Miguel Ballesteros, Weiyi Sun, Sandip Ghoshal, Michelle Yuan, Yassine Benajiba, Sujith Ravi, and Dan Roth. MT-OSC: Path for LLMs that get lost in multi-turn conversation. *arXiv preprint arXiv:2604.08782*, 2026. URL https://arxiv.org/abs/2604.08782.

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

Harsh Trivedi, Niranjan Subramanian, Tushar Khot Wang, Ashish Sabharwal, Ali Farhadi, and Aniruddha Kembhavi. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 16022–16076, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.850. URL https://aclanthology.org/2024.acl-long.850/.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10480–10496, 2024.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 808–819, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1078. URL https://aclanthology.org/P19-1078/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

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

A Matched Intervention Details

Execution setting and separation from the canonical study. The primary matched 7B experiment uses Qwen2.5-7B-Instruct with local Ollama CUDA inference, GGUF Q4_K_M quantization, and deterministic decoding. It contains 96 equivalent-state groups and 384 histories. The canonical screen used a different backend and quantization configuration. We therefore treat the screen and matched study as distinct experiments, not as repeated measurements from one inference stack.

Conditions. RAW contains the interaction history, query, and candidates. GOLD_ONLY replaces the history with the correct task-sufficient state. RAW_GOLD retains the history and adds that state. RAW_CORRUPT keeps RAW_GOLD's layout but changes one state value to imply a known alternative answer. The multi-field Preserve study changes one of ten field paths while mechanically preserving the deterministic answer; it contains 400 group–field combinations and 1,600 histories. No model output is used to select an intervention. RAW_OBSOLETE uses a previously valid state

and is evaluated on the 192 histories for which the matched obsolete intervention is defined. Between a gold-state prompt and its content intervention, the history, renderer, query, candidate set, state placement, and instructions are identical.

Table 3: Matched Qwen2.5-7B counts. Directional-hit rows evaluate agreement with the answer implied by the supplied manipulated state, not task accuracy.

Condition or outcome	Numerator	Denominator	Rate
RAW accuracy	288	384	75.00%
GOLD_ONLY accuracy	384	384	100.00%
RAW_GOLD accuracy	384	384	100.00%
RAW errors repaired by RAW_GOLD	96	96	100.00%
RAW-correct cases harmed by RAW_GOLD	0	288	0.00%
RAW_CORRUPT corrupt-answer hit	272	384	70.8333%
RAW_CORRUPT corrupt-answer hit, valid only	272	368	73.9130%
RAW_OBSOLETE obsolete-answer hit	32	192	16.67%
Matched candidate/position chance	32	192	16.67%

Corruption output audit. Each RAW_CORRUPT group changes exactly one decision-relevant field, and all 96 groups are mechanically verified to change the deterministic answer. The output partition is 272 corruption-consistent outputs, 48 gold outputs, 48 other valid outputs, and 16 invalid outputs. These categories sum to all 384 histories. The corresponding group-level paired delta is 0.708333, its 95% group bootstrap interval is $[0.6380208, 0.7734375]$, and the paired randomization test gives $p \approx 10^{-5}$.

History truncation baselines. LAST_USER_TURN_ONLY has 8.33% strict accuracy and a 91.67% valid-output rate. Relative to RAW, it repairs 0/96 errors and harms 256/288 correct decisions. LAST_K=2 reaches 45.83% strict accuracy with 100% valid outputs, repairing 32/96 errors and harming 144/288 correct decisions. These conditions rule out the simple explanation that correct-state repair is reproduced by discarding older context, but they are not matched causal controls because they change history presentation.

Multi-field answer-preserving intervention. Each intervention satisfies $S_{\text{preserve}} \neq S_{\text{gold}}$ while preserving $f(S_{\text{preserve}}, q, C) = f(S_{\text{gold}}, q, C)$. Across 400 group–field combinations, 1,600 histories, and ten distinct field paths, raw gold retention is 93.00%, the valid paired answer-change rate is 4.12%, and the invalid rate is 3.00%. The group-cluster gold-retention estimate is 95.72% with 95% bootstrap CI $[94.81\%, 96.61\%]$; the group-cluster answer-change estimate is 2.66% with CI $[2.11\%, 3.21\%]$. Field-specific behavior is not uniform: some destination and reliability or constraint edits cause errors or invalid outputs. The evidence therefore supports strong aggregate decision selectivity, not universal invariance.

Matched accuracy and path sensitivity. RAW_GOLD minus RAW is $+25.00$ percentage points with 95% group bootstrap CI $[20.83, 29.17]$ and paired randomization $p \approx 10^{-5}$. Pairwise path sensitivity is $\text{MIV}_{\text{pair}} : 0.4167 \to 0$, while the any-disagreement group rate is $\text{MIV}_{\text{group}} : 0.6667 \to 0$. These are distinct statistics. Since RAW_GOLD is correct on all 384 histories, both zero-MIV values follow from complete repair and are not treated as independent mechanism estimates.

Full-BF16 Qwen2.5-7B replication. The non-quantized replication uses Qwen2.5-7B-Instruct in full BF16 with Transformers 5.5.0, PyTorch 2.11.0+cu128, CUDA 12.8, and an NVIDIA RTX 4080 SUPER with 32 GB memory. It formats prompts with `tokenizer.apply_chat_template(add_generation_prompt=True)` and uses deterministic greedy decoding with `max_new_tokens=128` on the same frozen 96-group/384-history protocol. RAW is 288/384 (75.00%); GOLD_ONLY and RAW_GOLD are both 384/384 (100.00%). RAW_GOLD repairs 96/96 errors, harms 0/288 correct answers, and yields the same $+25.00$ point effect, 95% CI $[20.83, 29.17]$, and paired $p \approx 10^{-5}$. Pairwise and group MIV again move from 0.4167 and 0.6667 to zero. RAW_CORRUPT produces the corruption-implied answer in 304/384 cases (79.17%), with 368 valid and 16 invalid outputs. Under Preserve, 352/384 outputs are valid and 32 are invalid; none of the 352 valid paired decisions changes. Because runtime and precision change together, this run is qualitative robustness evidence rather than a controlled estimate of quantization's effect.

Controlled semi-natural surface robustness. The surface study uses a stratified 48-group subset containing 192 histories and all six operation buckets. DeepSeek-V4-Flash rewrites each template history into more natural dialogue. Conservative semantic-equivalence validation accepts all 192 rewrites. The rewrite changes only surface form: latent transitions, final states, queries, candidate sets, and Gold, Corrupt, and Preserve intervention objects remain unchanged. The same full-BF16 Qwen2.5-7B stack evaluates the template and semi-natural versions.

Table 4: Same-model surface robustness on a matched 48-group subset. Corrupt reports directional hits among valid outputs; Preserve reports valid paired answer changes.

Surface	RAW	RAW_GOLD	Repair	Harm	Corrupt valid	Preserve change
Template	144/192	192/192	48/48	0/144	152/184 (82.61%)	0/176
Semi-natural	169/192	191/192	23/23	1/169	134/189 (70.90%)	1/184 (0.54%)

Surface form changes difficulty: RAW rises from 75.00% on templates to 88.02% on semi-natural histories. The matched intervention pattern nevertheless remains: observed RAW errors are repaired, decision-changing corruption produces substantial directional movement, and valid paired decisions are highly stable under answer-preserving edits. This is controlled surface robustness, not validation on real interaction logs.

Obsolete intervention. The obsolete-answer hit rate is 32/192 (16.67%). The appropriate candidate-size and target-position-matched chance baseline is also 16.67%, giving zero lift and matched randomization $p = 1.000$. A uniform non-gold baseline is inappropriate because candidate set sizes and target positions are unbalanced. Thus active obsolete-state following is indistinguishable from matched chance, even though canonical raw errors show elevated passive overlap with obsolete answers.

Secondary Qwen2.5-0.5B contrast. The smaller-model run yields approximately 29.17% RAW accuracy, 83.33% GOLD_ONLY accuracy, and 91.67% RAW_GOLD accuracy. Its obsolete-answer and corruption-consistent directional hit rates are both 0%. Differences in backend, quantization, capability, and valid-output coverage prevent a scaling interpretation.

DeepSeek-V4-Flash. The SiliconFlow model `deepseek-ai/DeepSeek-V4-Flash` is evaluated with the same frozen protocol. Each full-grid condition has 384 outputs; the obsolete condition has 192. RAW is correct on 289/384 histories (75.26%), while RAW_GOLD is correct on 383/384 (99.74%), a gain of 24.48 percentage points (95% group-bootstrap CI $[16.93, 32.29]$ pp; paired $p = 10^{-5}$). Correct state repairs 95/95 RAW errors and harms 1/289 previously correct cases. RAW_CORRUPT produces the corruption-implied answer in 222/384 cases (57.81% overall; 222/333 = 66.67% among valid outputs), with 51 invalid outputs. RAW_OBSOLETE produces the obsolete answer in 52/192 cases (27.08% overall; 52/151 = 34.44% among valid outputs), with 41 invalid outputs. RAW_PRESERVE retains the gold answer in 383/384 cases, produces one valid paired answer change among 383 pairs (0.26%), and has no invalid outputs. We do not report a full-grid MIV change: RAW and RAW_GOLD share only 79 groups with four valid outputs.

Secondary Gemma 4 E4B cross-family evidence. Gemma 4 E4B is evaluated on the frozen MarkovLens 96-group, 384-history dataset with the same matched state-intervention protocol. Inference uses local llama.cpp CUDA, Q4_K_M quantization, and deterministic decoding. RAW_GOLD accuracy is 95.8%, an improvement of $+33.3$ percentage points with 95% group-bootstrap CI $[+25.3, +41.4]$ pp. The correct-state intervention repairs 144/144 RAW errors and harms 16/240 previously correct RAW cases. RAW_CORRUPT produces the corruption-implied answer in 224/384 cases (58.3%), while RAW_OBSOLETE produces the obsolete-state answer in 16/192 cases (8.3%). Output-format reliability is poor: invalid or malformed outputs are frequent, and all GOLD_ONLY generations fail the strict output-format validity criterion, although answer fields can often still be recovered. Recoverable fields do not make the required interface reliable. We do not report a Gemma MIV change because RAW and RAW_GOLD do not share enough groups with four valid outputs for a reliable estimate.

B Additional Experimental Details

All reported runs use canonical answer normalization before computing accuracy, MIV, and OSA. Invalid or unparsable outputs count as incorrect and as disagreements when their normalized forms differ. Equivalent histories are distinct state-transition paths rather than paraphrases and are replay-verified to induce the same final state. Non-leakage validation rejects states containing final choices, candidate identifiers, rankings, scores, generated code, or final actions. Candidate perturbation holds the state and query fixed while changing candidates so that the correct answer changes.

Confidence intervals resample final-state groups rather than individual histories because the four histories within a group are dependent. Paired tests operate on group-level outcomes evaluated under both conditions. Deterministic decoding prevents sampling variation from being counted as path sensitivity.

For an equivalent-history group $\mathcal{G}$ with K histories, $\mathrm{MIV}_{\mathrm{pair}}$ averages $\mathbf{1}(\hat{y}_i \neq \hat{y}_j)$ over all within-group pairs and then over groups; $\mathrm{MIV}_{\mathrm{group}}$ is the fraction of groups containing any disagreement. OSA is $\Pr(\hat{y}_t \in \mathcal{O}_t \mid \hat{y}_t \neq y_t)$, where $\mathcal{O}_t$ contains deduplicated answers implied by earlier states after removing the final gold answer. Its random calibration matches candidate-set size and target position when those opportunities are unbalanced.

C Additional Controls

C.1 Interface and Cross-Task Controls

Table 5: Interface-equalized control on a 48-group User Constraint Update slice. Compact and dialogue-style state renderings can produce different conclusions.

Model	Raw	Raw recap NL	Raw JSONized	Gold compact	Gold dialogue	Equalized SSG
`qwen2.5:7b`	0.833	0.797	0.875	1.000	1.000	+0.125
`mistral:7b`	0.917	–	0.833	0.833	1.000	+0.083

Table 6: Tool/Environment Mutation 48-group interface summary.

Model	Raw	Raw JSONized	Compact Gold	Dialogue Gold
`qwen2.5:7b`	0.958	0.958	0.792	1.000
`mistral:7b`	0.875	0.917	0.833	0.833

Table 7: Small frontier/API interface control. These slices are interface probes, not population-level scaling results.

Task family	Raw	Raw JSONized	Gold JSON	Gold NL/dialogue	Task-aware	Ledger
constraint_update	0.833	–	0.500	0.667	0.792	0.708
code_requirement_drift	1.000	–	1.000	–	–	–
tool_env_mutation	0.917	1.000	1.000	0.833	–	–

C.2 State Construction and Construct Validity

Table 8: Prompted state-extraction snapshot. These results audit state construction but are not primary causal evidence.

Model	Raw	Gold	Task-aware	Ledger	0-shot exact	0-shot F1	Few exact	Few F1
`qwen2.5:7b`	0.833	1.000	0.875	0.836	0.000	0.415	0.375	0.626
`mistral:7b`	0.917	0.833	–	–	–	–	0.542	0.641

Table 9: Supervised construct-validity controls.

Task family	Control	Train	Eval	State exact	State F1	Downstream acc.
constraint_update	state-to-answer solver	240	96	–	–	1.000
code_requirement_drift	state-to-answer solver	120	24	–	–	1.000
constraint_update	structured extractor	240	96	1.000	1.000	1.000
code_requirement_drift	structured extractor	120	24	1.000	1.000	1.000

Table 10: Robustness ladder for mediated state extraction. In-distribution supervision succeeds, while held-out variants and unseen values expose brittleness.

Task family / split	Baseline	State exact	State F1	Downstream acc.
constraint / variant holdout	structured extractor	0.500	0.931	1.000
constraint / variant holdout	nearest neighbor	0.417	0.786	0.917
code / variant holdout	structured extractor	0.250	0.833	0.500
code / variant holdout	nearest neighbor	0.500	0.832	0.583
constraint / value shift	structured extractor	0.000	0.305	0.000
constraint / value shift	nearest neighbor	0.000	0.305	0.000
code / value shift	structured extractor	0.000	0.343	0.333
code / value shift	nearest neighbor	0.000	0.343	0.333

C.3 OSA Calibration

Table 11: OSA calibration against size-matched random non-gold answer sets.

Model	OSA	Random baseline	Lift	Ratio
`qwen2.5:7b`	1.000	0.533	+0.467	1.875
`mistral:7b`	0.500	0.500	0.000	1.000

C.4 Update Type, Complexity, and Length

Table 12: Operation-conditioned raw-history error rates in the 96-group balanced grid.

Model	Neg.	Conflict	Uncert.	Delete	Overwrite	Retain
`qwen2.5:7b`	0.500	0.375	0.333	0.250	0.250	0.000
`mistral:7b`	0.250	0.125	0.083	0.167	0.083	0.000

Table 13: Complexity-versus-length summary. State-update complexity (SUC) is the stronger predictor in these sweeps.

Model	Examples	$r(y, \mathrm{SUC})$	$r(y, \mathrm{tok})$	β_{SUC}	β_{tok}	R^2
`qwen2.5:7b`	168	-0.388	+0.254	-0.0229	+0.0018	0.177
`mistral:7b`	168	-0.317	+0.019	-0.0263	-0.0008	0.104

Table 14: Condition-level complexity-versus-length results. Complexity rows approximately fix token length; length rows fix update complexity.

Family	Condition	Tokens	SUC	Qwen Acc.	Mistral Acc.	Notes
Complexity	simple add/retain	130.0	4.0	1.000	1.000	easy update
Complexity	delete	126.8	6.0	1.000	0.750	cancellation
Complexity	overwrite	128.8	8.0	1.000	1.000	replacement
Complexity	uncertainty	123.2	8.5	1.000	0.750	uncertainty
Complexity	conflict	126.5	17.5	0.750	0.500	conflict
Complexity	mixed hard	127.0	21.5	0.500	0.500	dense updates
Length	short	38.5	17.5	0.750	0.750	fixed complexity
Length	medium	97.1	17.5	0.500	0.500	fixed complexity
Length	long	155.9	17.5	1.000	0.625	fixed complexity

D EPISODE GENERATION ALGORITHM

Algorithm 1: MarkovLens episode generation

1. Sample task family F, operation budget T, equivalent variants K, and initial state S_0.
2. For $t = 1, \ldots, T$, sample an operation a_t, update $S_t = U_F(S_{t-1}, a_t)$, and render observation $o_t = R_F(a_t, S_{t-1}, S_t)$.
3. Generate decision context C_T and query q_T.
4. Compute gold answer $y_T = f_F(S_T, q_T, C_T)$.
5. Generate obsolete answer set $\mathcal{O}_T$ from previous states $\{S_0, \ldots, S_{T-1}\}$.
6. Render equivalent histories $H_T^{(1)}, \ldots, H_T^{(K)}$ such that all induce S_T.
7. Validate non-leakage and candidate perturbation.
8. Return the gold state, answer, operation labels, obsolete states, and equivalent-history group identifier.