When Context Sticks: Hysteresis from Repeated Semantic Framing in LLMs

Anonymous authors
Paper under double-blind review

Abstract

Large language models increasingly operate over long multi-turn histories, yet it remains unclear whether repeated semantic framing acts only as a local prompt effect or can establish a persistent history-dependent internal state. We study this question using fact-matched dialogue trajectories, dose-response controls, residual-stream measurements, a Same Present, Different Past protocol, reversal tests, and activation-level interventions across eight open-weight models. Behavioral framing preference increased from 0.155 after one exposure to 0.596 after five exposures while the most recent framed turn was held fixed. Frame-related residual projections accumulated with exposure and remained separable after token-identical neutral washout; after eight neutral turns, 11.6% of the behavioral and 19.8% of the representational A-versus-B separation remained. Adding, removing, or transferring the identified frame component produced directionally consistent behavioral changes, whereas random and orthogonal controls were near zero. The pattern generalized across framing axes and appeared in observational and counterfactual replay analyses of real-world conversation structures.

1 Introduction

Large language models increasingly operate in extended multi-turn settings in which earlier context can influence decisions many turns later (Laban et al., 2026; Hong et al., 2025). In such interactions, semantically similar cues are often encountered repeatedly: an assistant may repeatedly describe evidence in terms of risks rather than benefits, emphasize short-term rather than long-term consequences, or consistently organize information around one explanatory perspective. These repeated signals are typically treated as additional pieces of context, but it is unclear whether their effects are simply local and transient. Repeated exposure may instead alter the model's internal state in a cumulative manner, causing later generations to remain aligned with an established frame even when that frame is no longer explicitly present. This distinction matters for long-horizon assistants and reasoning systems, where models continuously integrate prior interaction history rather than responding to isolated prompts (Wu et al., 2025; Bai et al., 2025). More broadly, it raises a basic question about contextual computation in language models: does repeated semantic framing merely influence the next prediction, or can it progressively establish a history-dependent internal state without any parameter update?

Answering this question is difficult because several simpler mechanisms can produce superficially similar behavior. A model that appears increasingly aligned with a repeated frame may simply be responding to the most recent framed turn, repeating recently observed vocabulary, or continuing the dominant topic of the context (Yan et al., 2024). Longer contexts may also change generation independently of semantic framing (Bai et al., 2024; 2025). Moreover, observing a behavioral shift alone does not reveal what has changed internally. Conversely, finding that a frame can be decoded from hidden activations does not establish that the corresponding representation is accumulated across turns or functionally used during generation (Lepori et al., 2026). Distinguishing these possibilities therefore requires more than comparing model outputs before and after repeated prompting. We need to separate semantic accumulation from recency and lexical repetition, determine whether repeated exposure leaves a measurable trace in internal representations, and test whether that trace continues to matter when the current input is held fixed. A particularly strong signature of history dependence would arise when two trajectories reach the same present context through different framing histories yet remain behaviorally or representationally distinct.

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.

In this work, we study repeated semantic framing through a mechanism-isolating multi-turn benchmark. We construct fact-matched dialogue trajectories in which the underlying information and final probe are held constant while only the framing history varies. Exposure is manipulated in a dose-response design, allowing us to compare models that encounter the same most recent framed evidence but differ in the number of earlier exposures. This isolates cumulative effects from simple recency. We then track frame-related residual-stream representations at a fixed generation point and ask whether their strength changes systematically with exposure. To directly test history dependence, we introduce a Same Present, Different Past paradigm: trajectories begin with different framing histories and then converge onto an identical neutral continuation before receiving the same final probe. If their later internal states and generation preferences remain separable, the difference must be attributable to prior context rather than the immediate input. We further examine reversal dynamics by exposing a previously established frame to an opposing frame, asking whether changing the model's direction depends on how strongly the earlier frame was established. Finally, we intervene directly on the identified frame-related representation by adding, removing, or transferring its activation component, providing a causal test of whether the representation contributes to subsequent generation rather than merely correlating with it.

Our study connects three levels of analysis that are often examined separately: observable generation behavior, internal representation dynamics, and causal manipulation of model states. We evaluate the core effects across both standard instruction-tuned and reasoning-oriented open-weight models, and replicate the main analyses across multiple semantic framing axes to assess whether the phenomenon is specific to a single framing construction. We further examine whether these dynamics extend to naturally occurring multi-turn interactions by combining real-world conversation analysis with matched counterfactual replay. Overall, the paper makes three methodological contributions. First, we introduce a factorized multi-turn benchmark that distinguishes repeated semantic exposure from recency, lexical repetition, and generic context effects. Second, we connect exposure-dependent behavioral changes to the evolution and persistence of frame-related internal representations under matched current contexts. Third, we use representation-level interventions to test whether these internal states are functionally involved in later generation. Together, these analyses characterize when repeated contextual signals remain transient and when they become persistent, history-dependent components of an LLM's internal computation.

2 Related Work

Recent work has increasingly shown that language-model behavior depends not only on the current prompt but also on the trajectory through which the model arrives at it. Laban et al. (2026) systematically compare single-turn and multi-turn interaction and find that models become substantially less reliable when information is revealed incrementally; in particular, models often commit to early assumptions and fail to recover from them later in the conversation. This result highlights a form of conversational path dependence in which early context can shape behavior many turns downstream. Related work on persuasive dialogue has similarly shown that repeated interaction can progressively alter model responses (Hong et al., 2025). Xu et al. (2024), for example, demonstrate that factual beliefs expressed by LLMs can be shifted over multi-turn persuasive conversations containing misinformation, while Tan et al. (2025) study how models change stance under sustained corrective and misleading persuasion.

Activation-space steering, activation patching, repetition, and prompt-sensitivity work most directly related to our representation and control analyses is reviewed in the supplementary material (Rimsky et al., 2024; Arditi et al., 2024; Jiang et al., 2025; Zhang and Nanda, 2024; Yan et al., 2024; Sclar et al., 2024).

3 Method

3.1 Models, Data, and Experimental Splits

We construct fact-matched framing trajectories and separate representation discovery, hyperparameter selection, and final evaluation.

We evaluate eight open-weight models: Qwen3-8B, Llama-3.1-8B-Instruct, Gemma-3-4B-it, Ministral-8B-Instruct-2410, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-0528-Qwen3-8B, and Phi-4-mini-reasoning. All model weights remain frozen throughout the experiments. All eight models are used for behavioral experiments; Qwen3-8B, Llama-3.1-8B-Instruct, DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Llama-8B are additionally used for full representation and intervention analyses. We run inference in BF16/FP16 without weight quantization for mechanistic experiments. Each model uses its native tokenizer and chat template. Log-likelihood experiments use deterministic forward passes, while free-generation validation uses greedy decoding.

The primary factorized benchmark suite contains 96 scenarios drawn from 12 semantic topics, with eight scenarios per topic. Each scenario contains five factual evidence packets P1,...,P5. Every packet has three content-matched realizations: Frame A (Ai), Frame B (Bi), and neutral (Ni). All realizations preserve the same entities, numerical quantities, temporal information, and factual propositions; only semantic emphasis and framing vary. We use one primary framing axis, gain vs. loss, for the full mechanistic study. Two additional axes, risk vs. benefit and short-term vs. long-term, are used for replication. Each realization has three lexical paraphrases. Length differences between matched A/B/N realizations are constrained to no more than 10%.

Topics are divided into disjoint Calibration, Validation, and Test splits, with four topics assigned to each split. All scenarios, paraphrases, and trajectories derived from the same topic remain in the same split. Calibration data are used only to estimate frame directions and normalization statistics. Validation data are used to select intervention layers, probe hyperparameters, and intervention magnitudes. Test data are accessed only after all analysis choices are frozen.

3.2 Experiment 1: Does Repeated Framing Accumulate?

We test whether repeated semantic framing produces a dose-dependent behavioral effect beyond recency, lexical repetition, and generic context repetition.

Each trajectory contains five evidence slots followed by an identical neutral probe Q. For Frame A, the primary exposure conditions are

$$
\begin{aligned}
E0 &= [N_1, N_2, N_3, N_4, N_5] \to Q, \\
E1 &= [N_1, N_2, N_3, N_4, A_5] \to Q, \\
E3 &= [A_1, N_2, A_3, N_4, A_5] \to Q, \\
E5 &= [A_1, A_2, A_3, A_4, A_5] \to Q.
\end{aligned}
$$

Frame B uses the symmetric construction. The same A5 or B5 occupies the final evidence position in the E1, E3, and E5 conditions. Therefore, the distance between the most recent framed exposure and Q is identical across these conditions. Comparisons among E1, E3, and E5 consequently isolate the contribution of earlier repeated exposures rather than simple recency.

The primary behavioral endpoint is paired continuation log-likelihood. For each neutral probe, we construct matched frame-consistent continuations yA and yB, and compute

$$
B = \log P(y_A \mid X, Q) - \log P(y_B \mid X, Q), \qquad \Delta B_e = B_e - B_0.
$$

The primary test compares E1, E3, and E5 and evaluates whether framing preference increases monotonically with exposure count.

Control construction and free-form validation details are provided in the supplementary material.

3.3 Experiment 2: Representation Accumulation and History Dependence

We test whether behavioral accumulation corresponds to an internal frame-related representation and whether that representation persists when the recent context becomes identical.

For each transformer layer l, we record the post-block residual-stream activation at a fixed generation-anchor position: the final token of the identical neutral probe immediately before prediction of the first answer token. The same token position is used across all exposure conditions.

Frame directions are estimated exclusively from Calibration data. For each matched A/B calibration pair,

$$\delta_{i,l} = h^A_{i,l} - h^B_{i,l}, \qquad d_l = \frac{\frac{1}{N}\sum_i \delta_{i,l}}{\left\|\frac{1}{N}\sum_i \delta_{i,l}\right\|_2}, \qquad R_l = h_l^\top d_l.$$

History dependence is evaluated using a Same Present, Different Past protocol:

$$H_A = A^5 \mid N^k \to Q, \qquad H_B = B^5 \mid N^k \to Q, \qquad H_N = N^5 \mid N^k \to Q, \quad k \in \{0, 1, 2, 4, 8\}.$$

Persistence is quantified as the remaining difference in both ΔB and Rl after each washout length. Because the recent context is held constant, residual differences between HA and HB provide evidence of dependence on earlier interaction history.

To test reversal dynamics, we compare

$$A^5 \to B^m \to Q \quad \text{vs.} \quad N^5 \to B^m \to Q, \qquad m \in \{1, 2, 3, 4, 5\}.$$

together with the mirrored B→A construction. The reversal threshold is defined as the smallest m for which the behavioral preference changes sign. We use the term hysteresis only when trajectories exposed to matched current framing strength remain distinguishable because of their different prior histories; otherwise, we refer to the effect more conservatively as persistence or history dependence.

3.4 Experiment 3: Causal Intervention on Frame Representations

We test whether the identified frame-related representation contributes causally to persistent generation bias rather than merely correlating with it.

We perform three subspace-specific interventions. Addition increases activation along the learned frame direction:

$$\begin{aligned} h'_l &= h_l + \lambda d_l && \text{(addition)}, \\ h'_l &= h_l - [(h_l^\top d_l) - \mu_l^{\text{neutral}}] d_l && \text{(removal)}, \\ h'_{\text{low}} &= h_{\text{low}} + [h_{\text{high}}^\top d_l - h_{\text{low}}^\top d_l] d_l && \text{(transfer)}. \end{aligned}$$

The natural intervention unit is defined on Validation data as

$$\lambda^* = \mathbb{E}[R_l \mid E5] - \mathbb{E}[R_l \mid E0], \qquad \{0.5\lambda^*, 1\lambda^*, 2\lambda^*\}.$$

with $1\lambda^*$ prespecified as the primary intervention magnitude.

Negative controls use a random unit direction and a unit direction orthogonal to dl, matched to the same intervention magnitude. Causal support requires directional consistency: addition should increase frame-consistent ΔB, removal should reduce it, and transfer should shift low-exposure behavior toward the corresponding high-exposure state. Random and orthogonal interventions should not produce comparable systematic effects.

3.5 Experiment 4: Generalization and Statistical Evaluation

We test whether the effect generalizes across model families, reasoning styles, framing axes, and naturally occurring long-horizon conversations.

Table 1: Primary behavioral dose-response.

Exposure	Mean ΔB	Approx. 95% CI half-width
E0	-0.006	$\pm$0.014
E1	0.155	$\pm$0.014
E3	0.380	$\pm$0.014
E5	0.596	$\pm$0.014

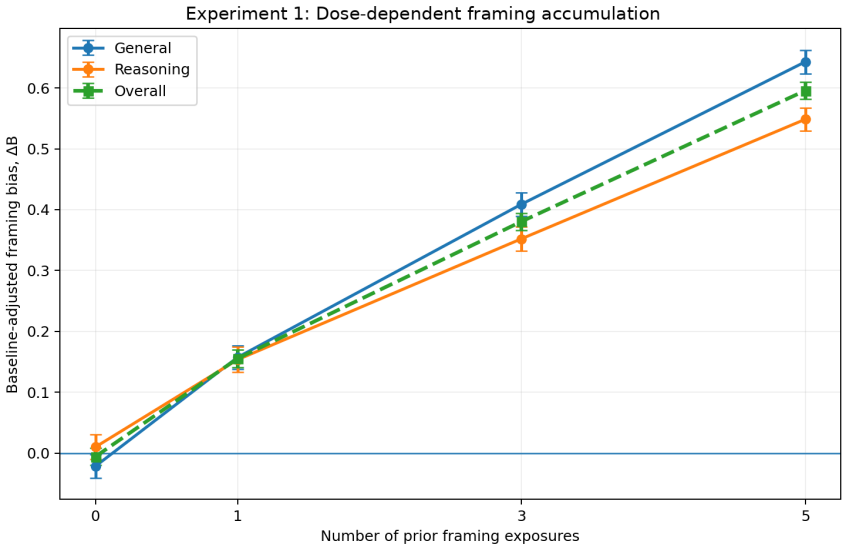


Figure 1: Experiment 1 primary dose-response.

All eight models are evaluated on the primary E0/E1/E3/E5 dose-response experiment and the Same Present, Different Past experiment. The four mechanistic models additionally undergo the full residual-stream and intervention analyses. The gain/loss axis is evaluated across all eight models. Risk/benefit and short-term/long-term framing are used as replication axes for the four mechanistic models. These replication axes test the core dose-response and history-dependence effects but do not repeat every causal intervention.

External-validity analyses use observational framing histories and matched counterfactual replay on real-world conversation structures; full selection, replay, and statistical details are provided in the supplementary material.

4 Results

4.1 Repeated Framing Produces Dose-Dependent Behavioral Accumulation

Table 1 and Figure 1 summarize the primary dose-response.

Repeated semantic framing produced a clear dose-dependent shift in subsequent model behavior. Across the eight models, baseline-adjusted framing preference increased from 0.155 after one exposure to 0.380 after three exposures and 0.596 after five exposures. Because the most recent framed evidence occupied the same position in the E1, E3, and E5 trajectories, the increase cannot be explained solely by the most recent framed turn. The pattern was directionally consistent across all evaluated models, although effect magnitude varied across model families.

Control conditions further separated semantic accumulation from simpler forms of repetition. The dose-response remained largely intact when each exposure used a different lexical paraphrase,

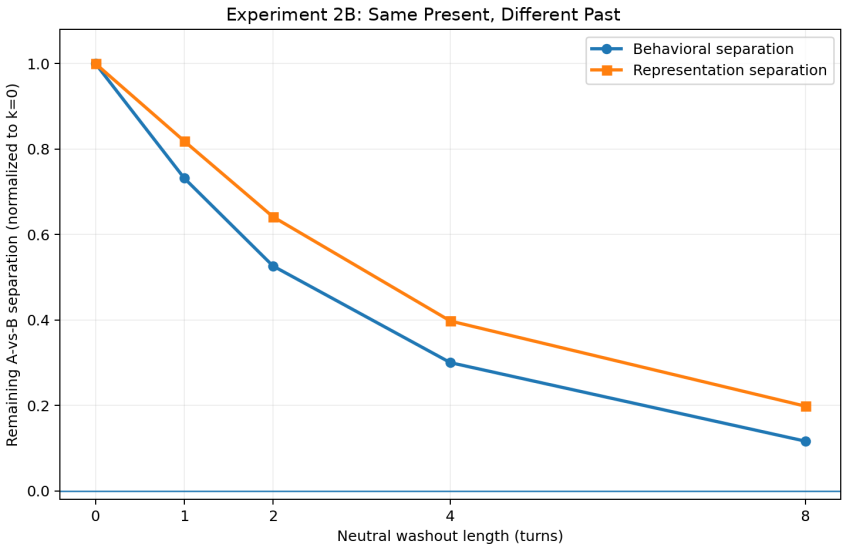


Figure 2: Persistence under Same Present, Different Past neutral washout.

reaching 0.552 at five exposures compared with 0.618 under the primary semantic-framing condition. In contrast, lexical repetition without the corresponding semantic frame produced only 0.071 bias at five exposures, while matched neutral repetition produced 0.033. Temporal arrangement modulated the magnitude of the effect but did not eliminate it: three distributed exposures yielded 0.387, compared with 0.347 for early-clustered and 0.420 for late-clustered exposure. Together, these results support behavioral accumulation beyond simple lexical repetition or recency.

4.2 Frame Representations Accumulate and Remain History-Dependent

Behavioral accumulation was accompanied by a corresponding increase in frame-related residual-stream representations. Across all four mechanistic models, projection onto the independently estimated frame direction increased monotonically from one to three to five exposures. For example, Qwen3-8B increased from 0.303 at E1 to 0.684 at E3 and 1.025 at E5; the same qualitative pattern held across the remaining models. Representation strength was also strongly associated with behavioral framing preference, with trajectory-level correlations ranging from r = 0.783 to 0.845.

Figure 2 shows persistence after the recent context becomes token-identical.

These differences persisted after the recent context became identical. In the Same Present, Different Past experiment, the behavioral separation between A-history and B-history trajectories decreased with neutral washout but remained at 30.0% of its initial magnitude after four neutral turns and 11.6% after eight. Representational separation decayed more slowly, retaining 39.8% after four neutral turns and 19.8% after eight. Thus, histories that converged onto token-identical recent contexts remained distinguishable both behaviorally and internally.

Reversal dynamics showed the same path dependence. Without a prior opposing history, one B-frame exposure was sufficient to shift the average behavioral preference toward B (-0.307). By contrast, after five A-frame exposures, preference remained A-consistent after one opposing exposure (0.345), remained near neutral after two (0.046), and reversed only after additional B exposure (-0.228 at m = 3). Across models, A-history trajectories typically required two to three opposing exposures to reverse, compared with one for neutral-history trajectories. These asymmetric reversal trajectories are consistent with hysteresis-like dynamics rather than purely local contextual responsiveness.

4.3 Frame-Related Representations Causally Contribute to Generation

Figure 3 summarizes the primary intervention effects.

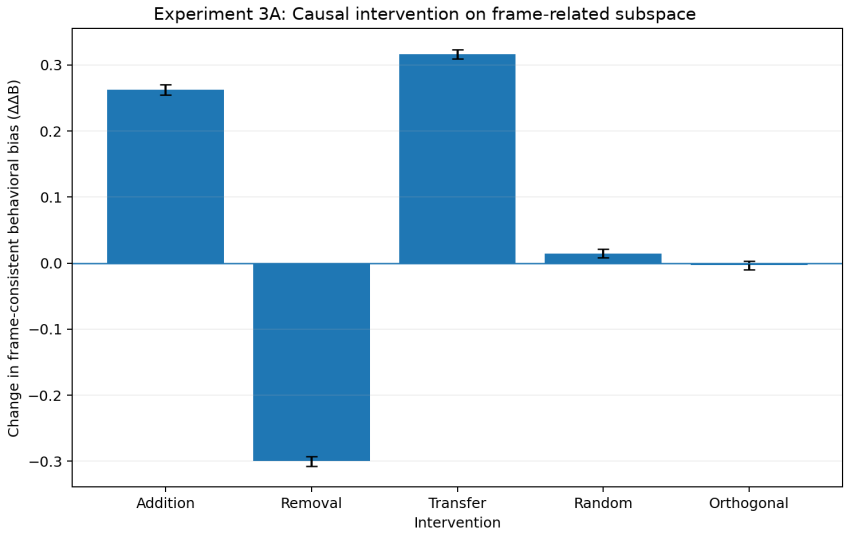


Figure 3: Causal intervention effects on frame-consistent behavioral bias.

Direct manipulation of the identified frame-related subspace produced directionally consistent behavioral changes. Adding the frame component increased frame-consistent bias by 0.263 on average, removing it reduced bias by 0.300, and transferring the high-exposure frame component into a low-exposure state increased bias by 0.317. In contrast, magnitude-matched random and orthogonal interventions produced near-zero changes (0.015 and -0.003, respectively), indicating that the effect was specific to the identified frame-related direction rather than a generic consequence of perturbing residual activations.

The behavioral effect also scaled with intervention magnitude. Frame-direction addition increased bias by 0.136, 0.256, and 0.444 at $0.5\lambda^*$, $1\lambda^*$, and $2\lambda^*$, respectively, while removal produced the corresponding decreases of -0.144, -0.287, and -0.467. This monotonic intervention response, together with the negative-direction controls, supports a causal contribution of the identified frame-related subspace to subsequent generation.

4.4 Effects Generalize Across Frames, Models, and Real-World Conversations

The accumulation effect generalized beyond the primary gain/loss manipulation. At five exposures, behavioral bias reached 0.666 for gain/loss framing, 0.586 for risk/benefit framing, and 0.488 for short-/long-term framing. History dependence also persisted across all three semantic axes: after eight neutral turns, A-versus-B separation remained at approximately 11%-12% of its initial magnitude. Across model families, all eight evaluated models exhibited measurable Same Present, Different Past effects, with approximately 29% of the initial behavioral separation remaining after four neutral turns and roughly 10% after eight.

Table 2 gives the aggregate counterfactual-replay scores; additional generalization and observational plots are reported in the supplement.

The same qualitative pattern appeared in the real-world conversation analysis. In the observational analysis, earlier-history framing consistency was moderately associated with later response framing tendency ($r = 0.441$); A-oriented histories were followed by an average score of +0.290, compared with -0.310 following B-oriented histories. Counterfactual replay provided a matched-history test using the same real-world conversation seed and identical final user request. Replaying the conversation under A-history produced a final framing score of +0.241, compared with -0.213 under B-history and -0.001 under neutral history, yielding an average A-versus-B gap of 0.454. The direction of this effect was consistent across all eight models. Together, the benchmark and real-world analyses indicate that repeated semantic framing can produce persistent history-dependent effects across multiple semantic dimensions, model families, and conversational settings.

Table 2: Counterfactual replay under identical final user request.

Counterfactual history	Overall final framing score
B-history	-0.213
Neutral-history	-0.001
A-history	+0.241

5 Discussion

Our results suggest that repeated semantic framing is not well characterized as a sequence of independent prompt effects. Behavioral preference increased from 0.155 after one exposure to 0.596 after five exposures even when the most recent framed turn was held fixed, while paraphrasing preserved most of the effect and lexical repetition produced only a small shift (0.071 at five exposures). This distinction connects to prior work on repetition: Yan et al. (2024) show that repeated surface co-occurrence can itself reinforce token associations during in-context learning. Our controls suggest a different phenomenon may coexist with this surface reinforcement: the accumulated effect remains when wording changes and is substantially stronger than repetition of frame-associated vocabulary alone. In this sense, repeated framing appears to alter not only which tokens are locally salient, but also the semantic state from which later predictions are made.

The internal analyses further suggest that this history dependence is representational rather than purely behavioral. Frame-related residual-stream projections increased monotonically with exposure, correlated with behavioral bias ($r = 0.783$–0.845), and retained 19.8% of their initial A-versus-B separation after eight neutral turns. Prior work has established that relatively high-level properties can often be represented as directions in activation space. Contrastive Activation Addition, for example, shows that behavioral attributes can be extracted from residual-stream differences and used to steer generation, while PERSONA reports approximately separable activation directions for personality traits (Rimsky et al., 2024; Feng et al., 2026). However, the presence of a decodable representation does not by itself imply that the model functionally uses it. Lepori et al. (2026) make this distinction explicit, showing cases in which models encode novel in-context semantics internally but fail to reliably deploy them downstream. Our intervention results address this gap directly: addition, removal, and transfer changed behavior by +0.263, -0.300, and +0.317, respectively, while random and orthogonal perturbations remained near zero. Together with the intervention-strength response, these findings are more consistent with the frame-related subspace participating in generation than with it merely recording past context. This conclusion should nevertheless remain localized: activation-patching results are known to depend on methodological choices, motivating our use of validation-only layer selection, magnitude-matched controls, and multiple intervention types rather than interpreting a single successful patch as a complete mechanism (Zhang and Nanda, 2024).

Extended discussion of path-dependent updating, agentic feedback, real-world implications, and trajectory robustness is provided in the supplementary material.

6 Limitations and Future Directions

Our study characterizes history-dependent framing in a factorized benchmark setting, but several questions remain open. First, the mechanistic analysis is restricted to open-weight models in the approximately 4–8B parameter range. This scale enables direct access to residual-stream activations, but it remains unclear whether the same representational dynamics persist in substantially larger models or change systematically with training regime. Activation-level interventions do not transfer uniformly across architectures: recent evaluations of steering methods across models from 1.5B to 70B report substantial variation in their effectiveness and reliability (Da Silva et al., 2025). Future work should therefore separate model scale from training effects using matched model families where possible, while testing behavioral history dependence in closed-weight systems even when internal interventions are unavailable.

Additional limitations concerning representation choice, explicit agent memory, richer framing dimensions, and complete hysteresis loops are discussed in the supplementary material.

7 Conclusion

This work asks whether repeated semantic framing remains a sequence of transient prompt effects or progressively alters the internal state from which a language model generates. By combining a factorized exposure design, matched-recency and repetition controls, residual-stream analysis, Same Present–Different Past trajectories, reversal tests, and activation-level intervention, we connect three phenomena that have largely been studied separately: contextual behavioral change, internal representation dynamics, and causal use of those representations.

The central perspective that emerges is that long-context LLM behavior may be trajectory-dependent. Two models facing the same current input can behave differently because they arrived there through different contextual histories, and repeated signals can leave internal traces that persist after those signals disappear from the immediate context. Real-world conversation analysis further motivates studying these dynamics beyond synthetic trajectories.

Accordingly, the main implication is broader than framing itself: evaluating a long-horizon model requires understanding not only how it responds to the information currently available, but also how its internal state evolves, persists, and recovers over the course of interaction. What a model does may depend not only on where it is in context, but on how it got there.

AI use statement

In this work, we have not used generative AI tools for any tasks with required disclosure. Additionally, we used generative AI tools solely to improve the readability of the manuscript, specifically for grammar correction, sentence-level restructuring, and wording refinement. We have reviewed all AI-assisted edits and verified that they did not alter the scientific content, claims, or conclusions of the paper. We take responsibility for the final content of this work.

Reproducibility statement

The model list, benchmark data construction, experimental splits, primary behavioral and representational endpoints, intervention-selection rule, and test protocol are described in Section 3. Detailed per-model tables, control conditions, diagnostic analyses, and statistical summaries are included in the supplementary material.

References

Arditi, Andy, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in Language Models Is Mediated by a Single Direction. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024). https://doi.org/10.52202/079017-4322.

Bai, Yushi, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3639–3664. https://doi.org/10.18653/v1/2025.acl-long.183.

Bai, Yushi, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3119–3137. https://doi.org/10.18653/v1/2024.acl-long.172.

Da Silva, Patrick Queiroz, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off Course: Reliability Challenges in Steering Language Models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 19856–19882. https://doi.org/10.18653/v1/2025.acl-long.974.

Feng, Xiachong, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. 2026. PERSONA: Dynamic and Compositional Inference-Time

Personality Control via Activation Vector Algebra. In International Conference on Learning Representations (ICLR).

Hong, Jiseung, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring Sycophancy of Language Models in Multi-turn Dialogues. In Findings of the Association for Computational Linguistics: EMNLP 2025, 2239–2259. https://doi.org/10.18653/v1/2025.findings-emnlp.121.

Jiang, Chonghe, Bao Nguyen, Anthony Man-Cho So, and Viet Anh Nguyen. 2025. Probe-Free Low-Rank Activation Intervention. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2812–2824. https://doi.org/10.18653/v1/2025.naacl-long.143.

Laban, Philippe, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs Get Lost In Multi-Turn Conversation. In International Conference on Learning Representations (ICLR).

Leask, Patrick, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. 2025. Sparse Autoencoders Do Not Find Canonical Units of Analysis. In International Conference on Learning Representations (ICLR).

Lepori, Michael A., Tal Linzen, Ann Yuan, and Katja Filippova. 2026. Language Models Struggle to Use Representations Learned In-Context. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 14841–14857. https://doi.org/10.18653/v1/2026.acl-long.676.

Poonia, Ansh, and Maeghal Jain. 2025. Dissecting Persona-Driven Reasoning in Language Models via Activation Patching. In Findings of the Association for Computational Linguistics: EMNLP 2025, 24553–24566. https://doi.org/10.18653/v1/2025.findings-emnlp.1335.

Rimsky, Nina, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 15504–15522. https://doi.org/10.18653/v1/2024.acl-long.828.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting. In International Conference on Learning Representations (ICLR).

Srikanth, Neha, Marine Carpuat, and Rachel Rudinger. 2024. How Often Are Errors in Natural Language Reasoning Due to Paraphrastic Variability? Transactions of the Association for Computational Linguistics 12: 1143–1162. https://doi.org/10.1162/tacl_a_00692.

Tan, Bryan Chen Zhengyu, Daniel Wai Kit Chin, Zhengyuan Liu, Nancy F. Chen, and Roy Ka-Wei Lee. 2025. Persuasion Dynamics in LLMs: Investigating Robustness and Adaptability in Knowledge and Safety with DuET-PD. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 1550–1575. https://doi.org/10.18653/v1/2025.emnlp-main.81.

Tan, Haoran, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025. MemBench: Towards More Comprehensive Evaluation on the Memory of LLM-based Agents. In Findings of the Association for Computational Linguistics: ACL 2025, 19336–19352. https://doi.org/10.18653/v1/2025.findings-acl.989.

Wu, Di, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In International Conference on Learning Representations (ICLR).

Xu, Rongwu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The Earth is Flat because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 16259–16303. https://doi.org/10.18653/v1/2024.acl-long.858.

Yan, Jianhao (Elliott), Jin Xu, Chiyu Song, Chenming Wu, Yafu Li, and Yue Zhang. 2024. Understanding In-Context Learning from Repetitions. In International Conference on Learning Representations (ICLR).

Zhang, Fred, and Neel Nanda. 2024. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In International Conference on Learning Representations (ICLR).

Zhao, Wenting, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In International Conference on Learning Representations (ICLR).

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In International Conference on Learning Representations (ICLR).

A Additional Related-Work Context

These studies establish that conversational history can produce persistent behavioral consequences, but they primarily characterize changes at the level of task performance, factual belief, or persuasion susceptibility. The underlying contexts also typically introduce new evidence or argumentative content across turns, making it difficult to distinguish the effect of repeated semantic framing from the effect of accumulating information. Our work focuses on a narrower setting in which factual content is matched across conditions and only the framing history varies. This allows us to ask whether repeated exposure itself produces a dose-dependent and history-dependent state, rather than merely whether earlier conversational content can influence later answers. Large-scale real-world conversation corpora provide a complementary setting for examining whether such history effects also arise in naturally occurring interactions (Zhao et al., 2024; Zheng et al., 2024). We therefore combine mechanism-isolating experiments with analysis of real multi-turn conversations and counterfactual replay of matched framing histories. This connects mechanism-isolating benchmark experiments with interaction patterns drawn from real conversational settings.

A complementary line of work investigates how contextual information and high-level behavioral attributes are represented inside language models. Lepori et al. (2026) show that LLMs can internally encode novel semantics introduced in context even when they fail to reliably use those representations for downstream prediction. Their findings distinguish between information being represented in hidden states and being functionally deployed during generation. Other work has shown that interpretable behavioral properties can often be associated with directions or subspaces in activation space. Contrastive Activation Addition (CAA), for instance, constructs steering vectors from mean residual-stream differences between contrastive examples and demonstrates that adding these directions during inference can systematically alter model behavior (Rimsky et al., 2024). More recent work extends this perspective to higher-level properties such as personality: PERSONA identifies approximately separable personality directions in activation space and manipulates them compositionally at inference time (Feng et al., 2026).

Activation patching and related causal interventions provide tools for moving beyond purely correlational representation analysis. Zhang and Nanda (2024) show, however, that activation-patching conclusions can depend strongly on methodological choices such as intervention construction and evaluation metric, motivating careful controls when attributing causal roles to internal components. Persona-oriented studies have likewise used activation patching to trace how contextual persona information propagates through model components and influences later reasoning (Poonia and Jain, 2025). Collectively, this literature demonstrates that semantic and behavioral information can be decoded and manipulated inside LLMs, but it usually treats the target representation as a relatively static object: a direction is identified from contrastive examples and then analyzed or steered. Our question is instead dynamic: whether repeated exposure causes a frame-related representation to strengthen across turns, remain after the immediate signal disappears, and retain functional influence because of the model's interaction history.

Repetition itself can alter language-model behavior through mechanisms that are simpler than semantic accumulation. Yan et al. (2024) study in-context learning through surface repetition and identify token co-occurrence reinforcement, showing that repeated contextual co-occurrence can strengthen associations between tokens during generation. This finding is particularly relevant to repeated-framing experiments because an apparent accumulation effect could arise from repeated lexical material rather than from a persistent semantic representation. More generally, prompt-sensitivity studies show that even meaning-preserving reformulations can substantially change LLM performance, emphasizing the need to separate semantic framing from surface-form variation (Srikanth et al., 2024; Sclar et al., 2024). Multi-turn persuasion research further demonstrates that sustained contextual signals can progressively shift model responses, but these signals typically differ in argumentative evidence as well as framing (Xu et al., 2024; Tan et al., 2025).

Our study therefore treats recency, lexical repetition, and semantic repetition as distinct explanations rather than interchangeable forms of context dependence. We use fact-matched framing, paraphrased exposures, lexical-repetition controls, and matched recent context to isolate repeated semantic framing from these simpler effects. We additionally ask a question that is largely orthogonal to prior repetition studies: whether two trajectories with the same present context but different framing histories can remain internally and behaviorally distinct. Finally, rather than inferring persistence solely from

output behavior, we connect exposure-dependent behavior to residual-stream representations and directly manipulate the corresponding subspace. In this sense, our work lies at the intersection of multi-turn behavioral dynamics, in-context representation analysis, and activation-level causal intervention, with the specific goal of determining when repeated contextual framing becomes a history-dependent internal state rather than a transient contextual cue.

B Extended Statistical and Evaluation Details

For external-validity analysis, we additionally use WildChat as a source of naturally occurring multi-turn conversations (Zhao et al., 2024). We retain conversations with sufficiently long interaction histories and a later user request that is approximately neutral with respect to the target framing axis. A separate frame-annotation procedure is used to identify trajectories with consistently A-oriented, B-oriented, or neutral prior histories. These real-world conversations are analyzed separately from the factorized benchmark suite and are never used for frame-direction estimation, layer selection, intervention tuning, or mechanistic inference. Where suitable multi-turn trajectories are available, LMSYS-Chat-1M is used as an optional supplementary replication source (Zheng et al., 2024).

Four controls use the same trajectory structure. In the paraphrased-frame control, each framed exposure uses a different lexical realization of the same semantic frame. In the lexical-repetition control, frame-associated vocabulary is repeated while the corresponding semantic framing is neutralized. In the neutral-repetition control, matched neutral material is repeated at the same positions and frequency. Finally, the order control compares three-exposure trajectories arranged as distributed [1,3,5], early-clustered [1,2,3], or late-clustered [3,4,5]. Together, these controls distinguish semantic accumulation from surface-form repetition, generic repetition, and temporal arrangement effects. Free-form generation is analyzed only as a secondary validation measure.

We test whether Rl changes systematically across E0, E1, E3, E5, and whether Rl predicts ΔB on held-out Test topics. Logistic-regression linear probes are trained as a secondary decodability analysis using Calibration data and evaluated cross-topic. Probe weights are not used as the primary frame direction, and probe performance is interpreted only as evidence of decodability rather than functional use.

Intervention layers are selected using Validation data only. For each layer, we compute cross-topic A/B separability together with the exposure-dependent representational change. We select the earliest layer whose Validation separability reaches at least 95% of the maximum layer-wise value. This rule favors an early stable representation rather than a potentially noisy peak and is fixed before Test evaluation. Full layer-wise curves are reported separately.

We additionally evaluate behavioral persistence using real-world multi-turn conversations drawn from WildChat. We first conduct an observational analysis in which the framing consistency of earlier conversation history is compared with the framing tendency of a later response after the conversation has transitioned to an approximately frame-neutral user request. This analysis tests whether naturally occurring repeated framing is associated with later response tendencies in realistic interaction histories, but it is treated as correlational because conversational content is not experimentally matched.

To reduce this confounding, we complement the observational analysis with counterfactual replay. For each selected real-world conversation, we construct matched A-history, B-history, and neutral-history versions of the earlier context while preserving the underlying factual propositions, conversational structure, and final user request as closely as possible. The final request is held identical across the three versions and replayed to each evaluated model. This analysis tests whether the history-dependent framing effects observed in the factorized benchmark suite remain detectable when the underlying topics, conversational structure, and user requests originate from real interactions. Where suitable trajectories are available, the same behavioral analysis is replicated on LMSYS-Chat-1M. These real-world analyses are used only to assess external validity; all mechanistic claims are based exclusively on the factorized benchmark suite.

The primary behavioral outcome is ΔB, and the primary representation outcome is Rl. Mixed-effects regression is used to account for repeated measurements within shared topics and scenarios. Exposure level, frame direction, and model type are included as fixed effects, together with prespecified interactions where appropriate; topic, scenario, and paraphrase are modeled as random effects. Exposure level is treated categorically in the primary analysis, with a separate monotonic-trend test

across 0, 1, 3, and 5 exposures. For washout experiments, washout length is modeled as an ordered predictor. Layer-wise exploratory analyses use Benjamini-Hochberg false-discovery-rate correction. We report effect sizes, 95% confidence intervals, and corrected p-values rather than treating individual trajectories as independent observations.

All frame directions, normalization statistics, probe settings, layer-selection rules, intervention magnitudes, and primary evaluation metrics are frozen before Test evaluation. Real-world analyses are reported separately from the held-out Test split and are not used to tune any experimental choices.

C Detailed Experimental Results

C.1 Experiment 1: Behavioral accumulation and controls

Table 3: Per-model behavioral dose-response.

Model	**E1**	**E3**	**E5**
Qwen3-8B	0.190	0.460	0.709
Llama-3.1-8B-Instruct	0.174	0.424	0.624
Gemma-3-4B-it	0.156	0.373	0.605
Ministral-8B-Instruct-2410	0.109	0.378	0.633
DeepSeek-R1-Distill-Qwen-7B	0.112	0.340	0.538
DeepSeek-R1-Distill-Llama-8B	0.184	0.359	0.589
DeepSeek-R1-0528-Qwen3-8B	0.169	0.400	0.590
Phi-4-mini-reasoning	0.149	0.309	0.476

Table 4: Semantic, paraphrase, lexical, and neutral controls.

Condition	**E1**	**E3**	**E5**
Semantic framing	0.154	0.391	0.618
Paraphrased framing	0.144	0.355	0.552
Lexical repetition	0.019	0.060	0.071
Neutral repetition	0.013	0.007	0.033

Table 5: Temporal arrangement control.

Three-exposure schedule	**Mean ΔB**	**95% CI half-width**
Distributed [1,3,5]	0.387	$\pm$0.015
Early [1,2,3]	0.347	$\pm$0.015
Late [3,4,5]	0.420	$\pm$0.014

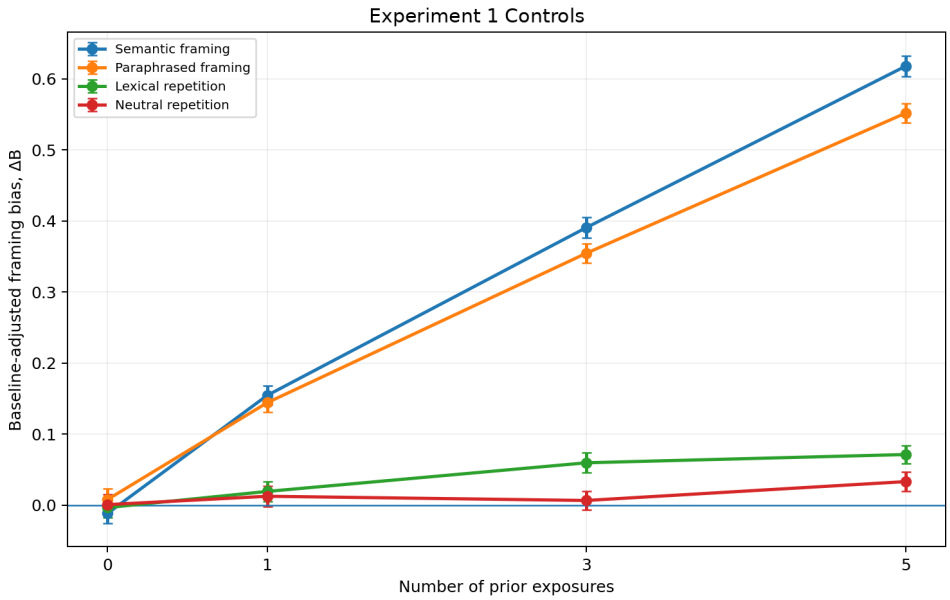


Figure 4: Control comparison. Paraphrased framing preserves the accumulation pattern, while lexical and neutral repetition are much weaker.

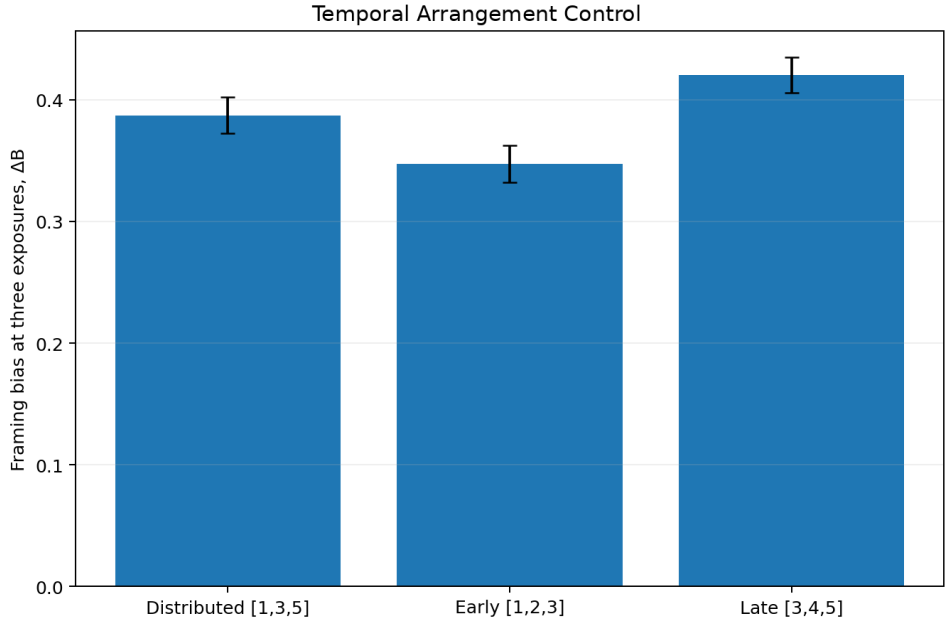


Figure 5: Order control: late exposure is somewhat stronger, but early and distributed schedules remain clearly above neutral.

C.2 Experiment 2A: Representation accumulation

Table 6: Representation accumulation and representation-behavior association.

Model	Layer	Rel. depth	E1 R	E3 R	E5 R	r(R,ΔB)
Qwen3-8B	22	0.611	0.303	0.684	1.025	0.845
Llama-3.1-8B-Instruct	20	0.625	0.237	0.578	0.898	0.833
DeepSeek-R1-Distill-Qwen-7B	21	0.583	0.191	0.504	0.804	0.802
DeepSeek-R1-Distill-Llama-8B	19	0.594	0.178	0.462	0.757	0.783

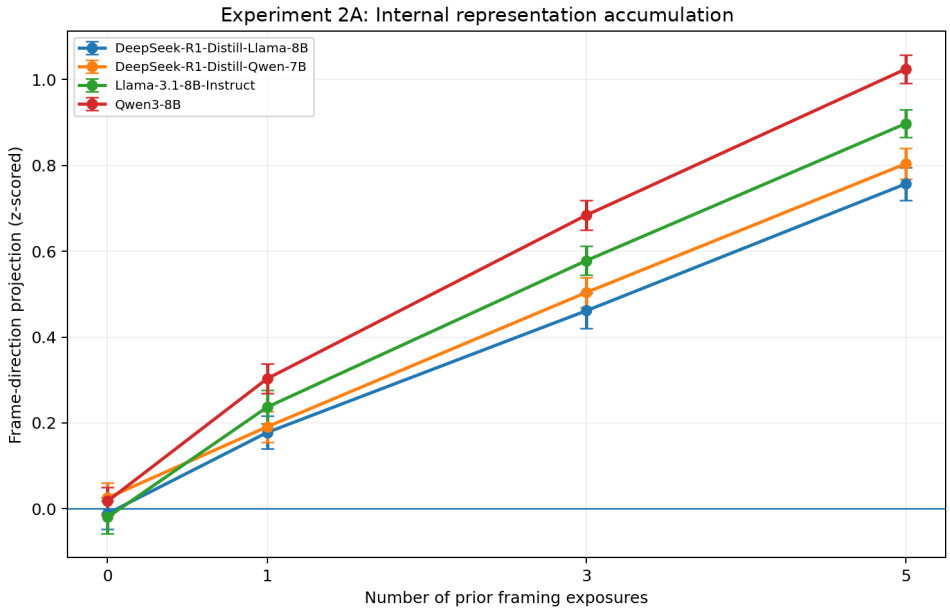


Figure 6: Residual-stream representation accumulation across framing exposures.

C.3 Experiment 2B: Same Present, Different Past

Table 7: Behavioral and representational persistence under neutral washout.

Neutral washout k	Behavior gap	Behavior retained	Representation gap	Representation retained
0	1.166	100.0%	1.834	100.0%
1	0.854	73.2%	1.502	81.9%
2	0.614	52.6%	1.177	64.2%
4	0.350	30.0%	0.729	39.8%
8	0.136	11.6%	0.364	19.8%

Table 8: Per-model behavioral separation under neutral washout.

Model	k0	k1	k2	k4	k8
Qwen3-8B	1.340	0.966	0.688	0.409	0.172
Llama-3.1-8B-Instruct	1.225	0.903	0.658	0.359	0.151
R1-Distill-Qwen-7B	1.086	0.789	0.593	0.323	0.100
R1-Distill-Llama-8B	1.015	0.758	0.517	0.308	0.119

C.4 Experiment 2C: Reversal and hysteresis-like dynamics

Table 9: Reversal trajectories under opposing framing.

Opposing B exposures	Behavior: A-history	Behavior: Neutral-history	Representation: A-history	Representation: Neutral-history
0	0.628	0.019	0.912	0.019
1	0.345	-0.307	0.559	-0.393
2	0.046	-0.647	0.204	-0.813
3	-0.228	-0.990	-0.144	-1.236
4	-0.514	-1.315	-0.497	-1.665
5	-0.795	-1.651	-0.855	-2.083

Table 10: Per-model reversal thresholds.

Model	A-history reversal threshold	Neutral-history threshold
Qwen3-8B	3	1
Llama-3.1-8B-Instruct	3	1
R1-Distill-Qwen-7B	3	1
R1-Distill-Llama-8B	2	1

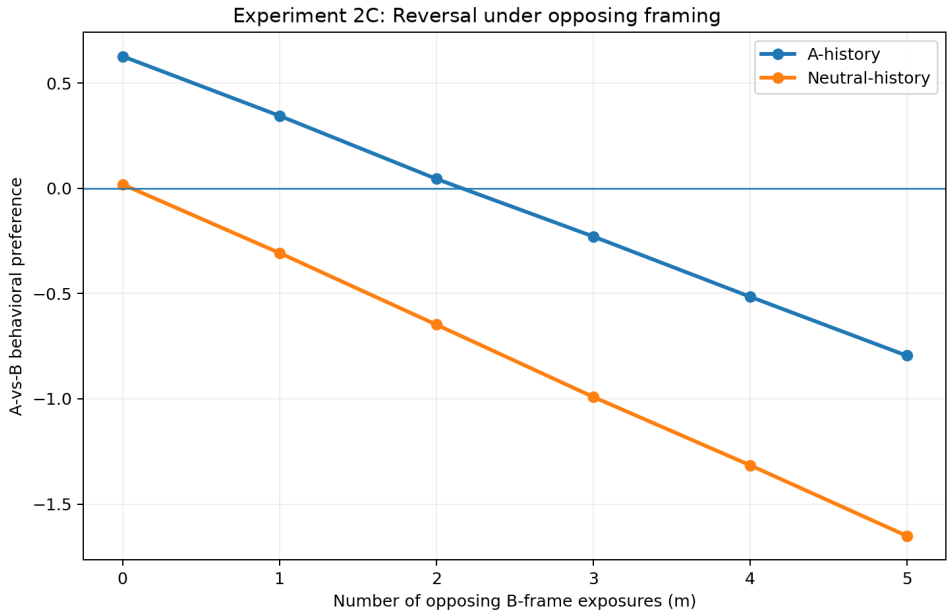


Figure 7: Reversal paths. Previously established A histories require more opposing B exposures to reverse than neutral histories.

C.5 Experiment 3A: Causal intervention

Table 11: Aggregate causal intervention effects.

Intervention	**Mean $\Delta\Delta B$**	**95% CI half-width**
Addition	+0.263	±0.008
Removal	-0.300	±0.007
Transfer	+0.317	±0.007
Random	+0.015	±0.006
Orthogonal	-0.003	±0.007

Table 12: Per-model causal intervention effects.

Model	**Addition**	**Removal**	**Transfer**	**Random**	**Orthogonal**
Qwen3-8B	0.312	-0.347	0.358	0.021	-0.005
Llama-3.1-8B-Instruct	0.287	-0.295	0.340	0.024	0.004
R1-Distill-Qwen-7B	0.230	-0.292	0.301	0.007	0.004
R1-Distill-Llama-8B	0.220	-0.265	0.268	0.007	-0.016

C.6 Experiment 3B: Intervention-strength dose response

Table 13: Intervention-strength dose response.

Strength	Addition	Removal	Random
0×	+0.006	+0.001	-0.003
0.5×	+0.136	-0.144	+0.005
1×	+0.256	-0.287	+0.008
2×	+0.444	-0.467	+0.009

Table 14: Per-model addition dose response.

Model	Addition 0.5×	Addition 1×	Addition 2×
Qwen3-8B	0.168	0.308	0.535
Llama-3.1-8B-Instruct	0.135	0.267	0.462
R1-Distill-Qwen-7B	0.126	0.230	0.403
R1-Distill-Llama-8B	0.115	0.218	0.374

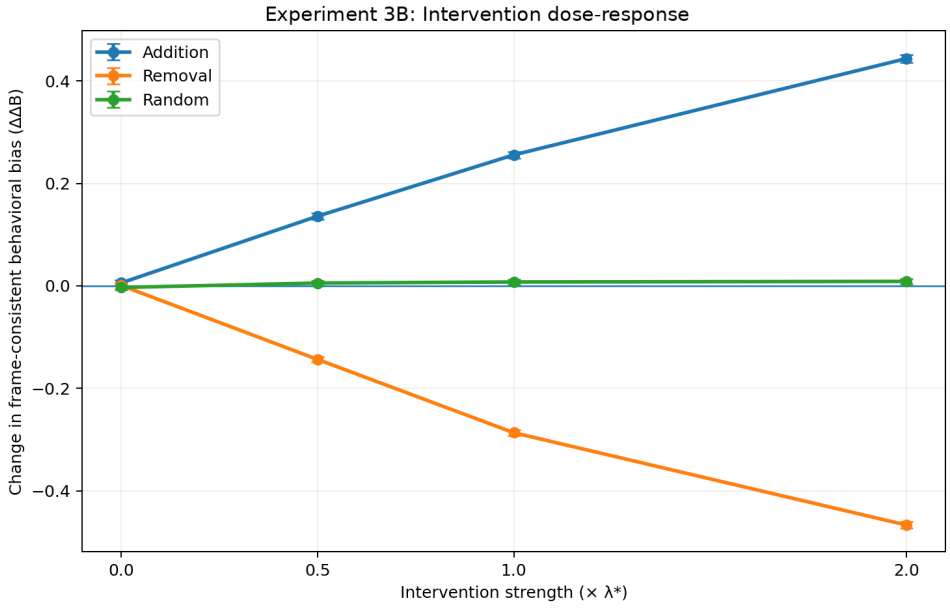


Figure 8: Intervention dose response.

C.7 Experiment 4A: Cross-frame generalization

Table 15: Dose-response generalization across framing axes.

Framing axis	E1	E3	E5	E5-E1
Gain/Loss	0.193	0.457	0.666	0.473
Risk/Benefit	0.176	0.388	0.586	0.410
Short-/Long-term	0.130	0.334	0.488	0.358

Table 16: Per-model E5 generalization across framing axes.

Model	Gain/Loss E5	Risk/Benefit E5	Short-/Long-term E5
Qwen3-8B	0.718	0.652	0.524
Llama-3.1-8B-Instruct	0.710	0.599	0.524
R1-Distill-Qwen-7B	0.622	0.563	0.479
R1-Distill-Llama-8B	0.615	0.530	0.427

Table 17: Cross-axis history dependence.

Axis	k1 retained	k2 retained	k4 retained	k8 retained
Gain/Loss	73.6%	51.6%	29.9%	12.1%
Risk/Benefit	72.8%	52.5%	28.7%	12.1%
Short/Long-term	73.2%	51.2%	26.4%	11.0%

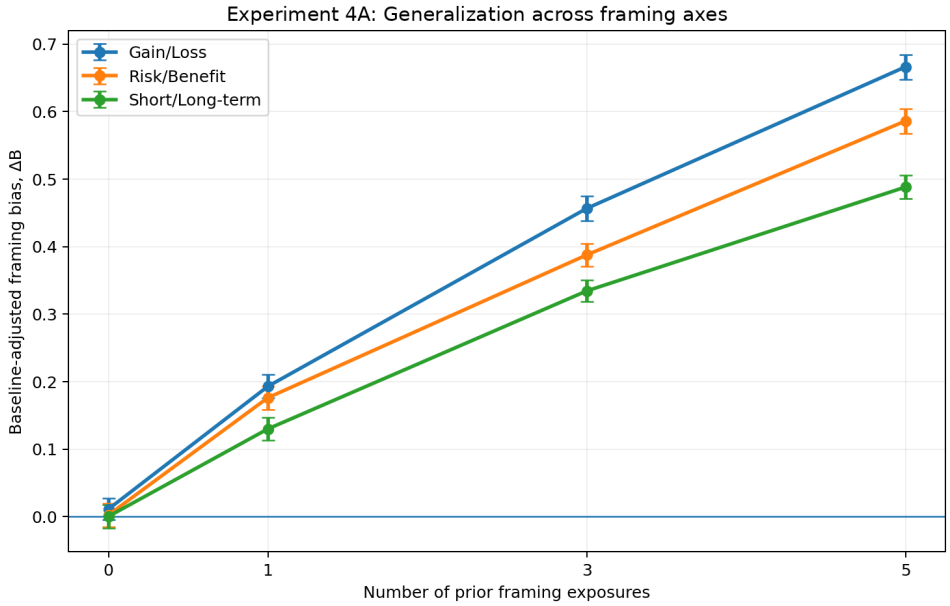


Figure 9: Dose-response generalization across three semantic framing axes.

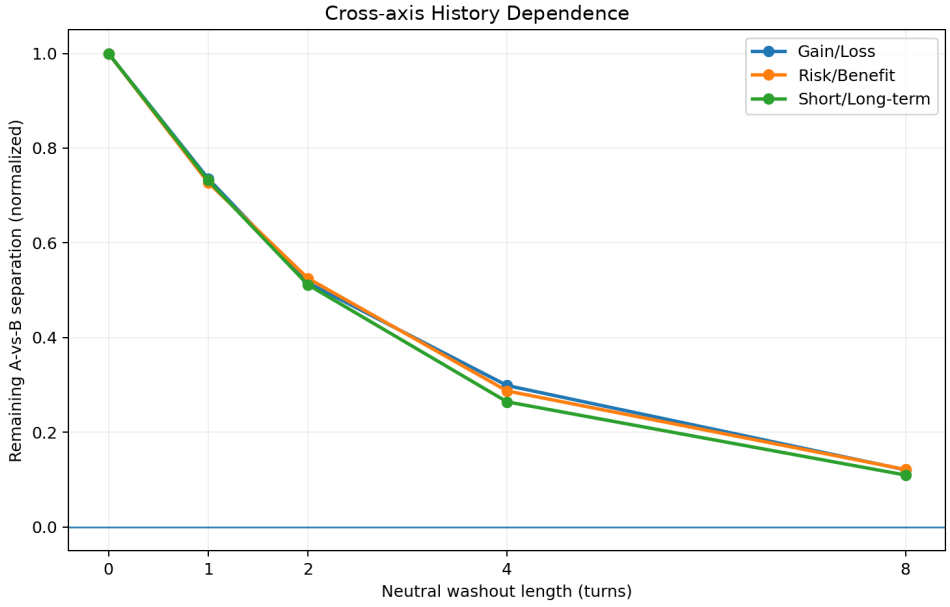


Figure 10: Same Present, Different Past persistence across framing axes.

C.8 Experiment 4B: Cross-model history-dependence generalization

Table 18: Cross-model history dependence after neutral washout.

Model	Type	k0	k4	Retain k4	k8	Retain k8
Qwen3-8B	General	1.309	0.412	31.5%	0.171	13.1%
Llama-3.1-8B-Instruct	General	1.228	0.351	28.6%	0.136	11.1%
Gemma-3-4B-it	General	1.115	0.307	27.5%	0.119	10.7%
Ministral-8B-Instruct-2410	General	1.183	0.338	28.5%	0.140	11.8%
R1-Distill-Qwen-7B	Reasoning	1.102	0.347	31.5%	0.106	9.7%
R1-Distill-Llama-8B	Reasoning	1.021	0.301	29.5%	0.086	8.4%
R1-0528-Qwen3-8B	Reasoning	1.151	0.335	29.1%	0.112	9.7%
Phi-4-mini-reasoning	Reasoning	0.945	0.240	25.4%	0.075	8.0%

Table 19: History dependence by model type.

Model type	Mean retain k4	Mean retain k8
General	29.0%	11.7%
Reasoning	28.9%	8.9%

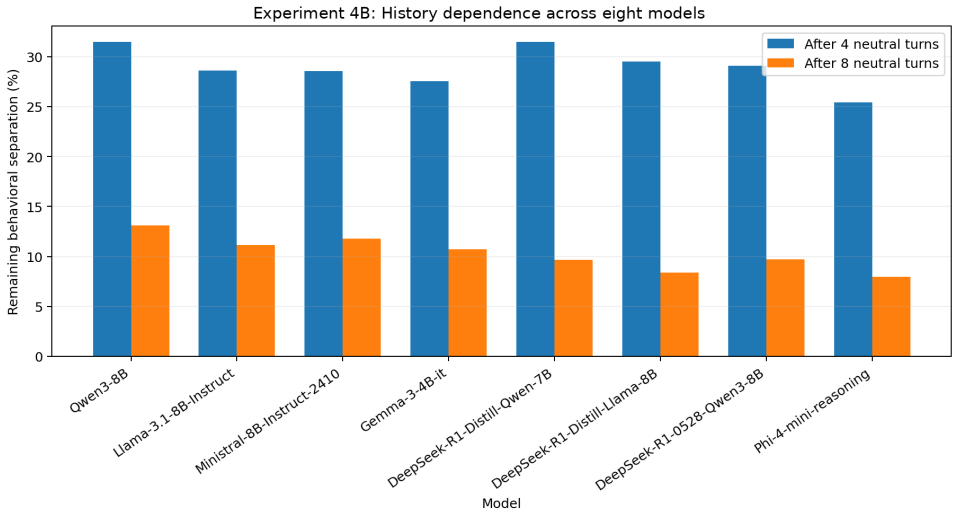


Figure 11: History dependence across all eight models.

C.9 Experiment 4C-1: Real-world observational validation

Table 20: Real-world observational analysis summary.

Quantity	**value**
Eligible conversation count	720
Correlation: history framing consistency → later response framing	r = 0.441
Simple regression slope	β = 0.347

Table 21: Later response framing by history group.

History group	**Later response framing score**	**n**	**95% CI half-width**
A-oriented history	+0.290	110	±0.067
Neutral history	-0.063	81	±0.086
B-oriented history	-0.310	113	±0.081

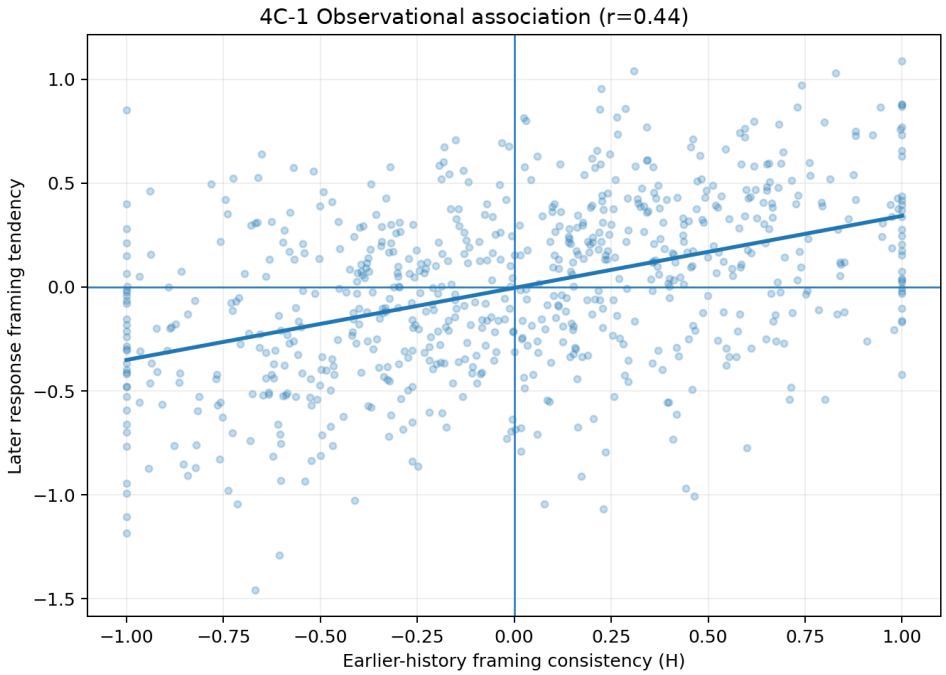


Figure 12: Observational association between earlier framing consistency and later response framing tendency.

C.10 Experiment 4C-2: Real-world counterfactual replay

Table 22: Per-model counterfactual replay.

Model	Type	B-history	Neutral	A-history	A-B gap
DeepSeek-R1-0528-Qwen3-8B	Reasoning	-0.220	+0.008	+0.240	0.461
DeepSeek-R1-Distill-Llama-8B	Reasoning	-0.187	+0.002	+0.206	0.393
DeepSeek-R1-Distill-Qwen-7B	Reasoning	-0.197	-0.006	+0.218	0.415
Gemma-3-4B-it	General	-0.212	+0.006	+0.234	0.446
Llama-3.1-8B-Instruct	General	-0.236	-0.006	+0.281	0.517
Ministral-8B-Instruct-2410	General	-0.222	-0.007	+0.262	0.484
Phi-4-mini-reasoning	Reasoning	-0.154	-0.010	+0.192	0.347
Qwen3-8B	General	-0.274	+0.003	+0.292	0.566

Table 23: Aggregate counterfactual replay contrasts.

Overall replay contrast	Mean difference
A-history - Neutral	+0.242
Neutral - B-history	+0.212
A-history - B-history	+0.454

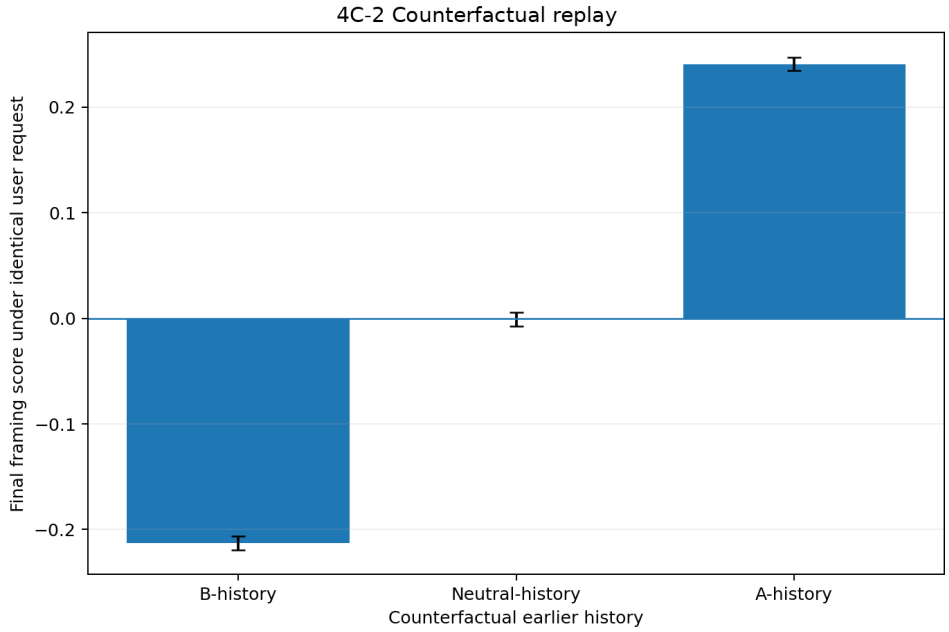


Figure 13: Counterfactual replay under identical final user request.

C.11 Auxiliary analyses and statistical diagnostics

Table 24: Linear-probe diagnostics.

Model	Test AUROC	Test accuracy
Qwen3-8B	0.91	0.813
Llama-3.1-8B-Instruct	0.89	0.816
R1-Distill-Qwen-7B	0.86	0.827
R1-Distill-Llama-8B	0.84	0.805

Table 25: Free-form generation validation diagnostic.

Diagnostic	value
Correlation with paired log-likelihood endpoint	r = 0.567

Table 26: Mixed-effects statistical summary.

Effect	Estimate	95% CI	z	Corrected p
Exposure: E1 vs E0	+0.161	[+0.142, +0.180]	+16.4	$< .001$
Exposure: E3 vs E0	+0.386	[+0.365, +0.407]	+35.7	$< .001$
Exposure: E5 vs E0	+0.602	[+0.580, +0.624]	+53.6	$< .001$
Monotonic exposure trend	+0.121	[+0.115, +0.127]	+39.9	$< .001$
Washout length	-0.112	[-0.120, -0.104]	-27.4	$< .001$
Representation strength $\rightarrow \Delta B$	+0.487	[+0.451, +0.523]	+26.5	$< .001$
Addition intervention	+0.263	[+0.247, +0.279]	+32.2	$< .001$
Removal intervention	-0.300	[-0.315, -0.285]	-39.2	$< .001$
Transfer intervention	+0.317	[+0.301, +0.333]	+38.8	$< .001$
Random-direction control	+0.015	[+0.002, +0.028]	+2.2	.041
Orthogonal-direction control	-0.003	[-0.016, +0.010]	-0.5	.642

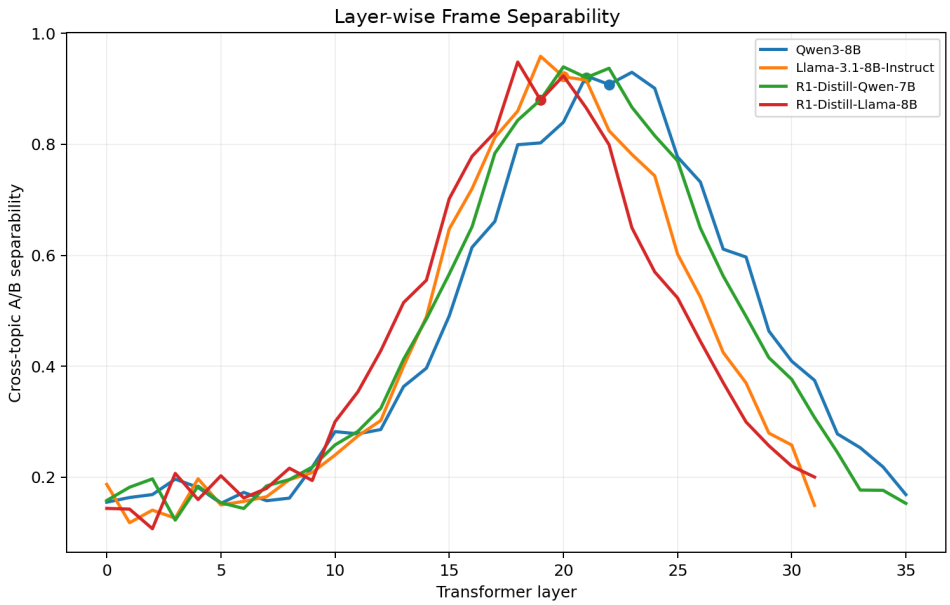


Figure 14: Layer-wise A/B separability; probe performance is treated as decodability evidence only.

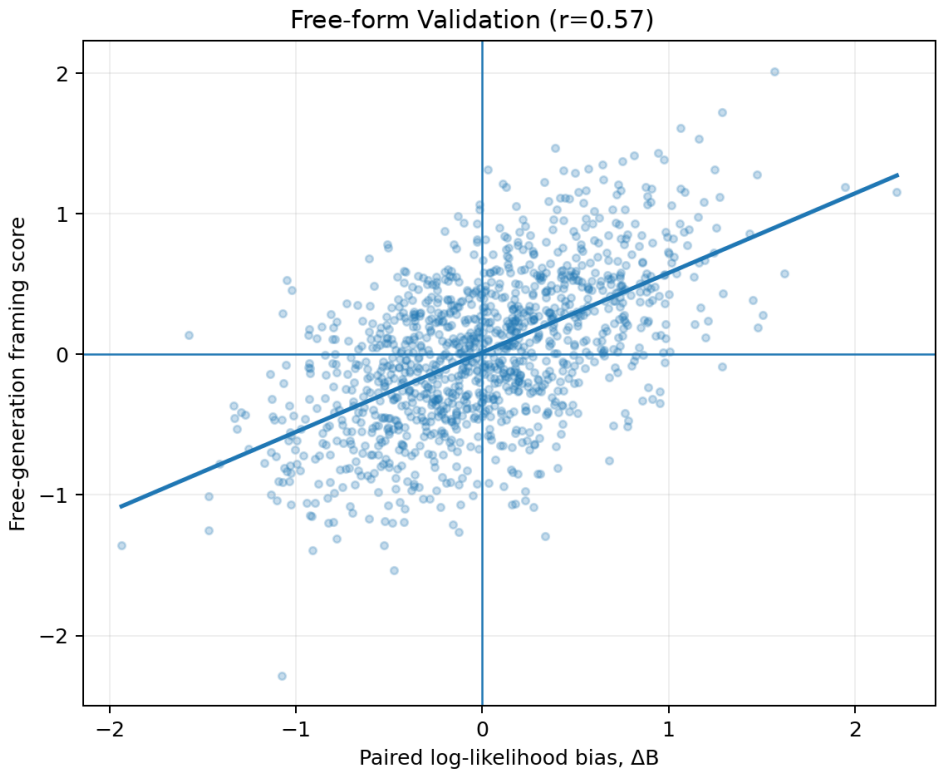


Figure 15: Association between paired log-likelihood bias and free-generation framing score.

D Extended Discussion

This interpretation also provides a possible internal account of behavioral path dependence previously observed in multi-turn interaction. Laban et al. (2026), for example, report substantial performance degradation when tasks are distributed over multi-turn conversations and find that models often become committed to assumptions made early in the interaction and fail to recover from them. Similarly, sustained persuasive dialogue has been shown to shift model beliefs and stances over multiple turns: Xu et al. (2024) demonstrate that factual beliefs can be manipulated by persuasive misinformation, while Tan et al. (2025) find that models can become highly susceptible to sustained misleading persuasion and emphasize the tension between resisting misinformation and remaining receptive to valid correction. Our results narrow this broader phenomenon by holding factual content constant. Under the Same Present, Different Past protocol, approximately 30.0% of the behavioral separation remained after four identical neutral turns and 11.6% remained after eight. Thus, in the factorized benchmark setting, multi-turn path dependence need not require accumulating new evidence or explicit persuasion: different histories of how the same facts were framed are sufficient to leave distinguishable downstream states.

The reversal experiments point to a further consequence: history can affect not only what a model currently prefers, but also how readily it updates. Neutral-history trajectories reversed after approximately one opposing exposure, whereas trajectories preceded by five exposures to the opposite frame generally required two to three. This resembles the stance-change dynamics studied in persuasive dialogue, but the factorized framing setup makes the implication somewhat different. The relevant failure mode is not simply that an LLM is "gullible" or "stubborn." Instead, its apparent receptiveness to new information may depend on the contextual state established before that information arrives. This may help reconcile why the same model can appear overly persuadable in one trajectory and resistant to correction in another: the response to the present signal is conditional on the path that produced the current internal state. Multi-turn sycophancy studies likewise find that stance changes unfold dynamically across conversational turns rather than being captured well by single-turn evaluation (Hong et al., 2025).

This view has concrete implications for long-horizon LLM applications. Consider a research assistant that repeatedly summarizes the same mixed evidence through a risk-oriented lens. Even if a later user asks for a neutral synthesis, our results suggest that removing explicit risk language from the final prompt may not fully restore a neutral starting point; the model may already arrive at that prompt in a history-shaped representational state. The analogous problem can arise in coding agents that repeatedly pursue one diagnosis for a bug, where later debugging steps may remain biased toward the early hypothesis; in tutoring systems, where repeated explanations through one conceptual framework may shape how subsequent ambiguous questions are interpreted; and in decision-support systems, where successive summaries emphasizing costs or benefits could influence later "neutral" comparisons. The engineering implication is that prompt neutrality at the current turn is not necessarily state neutrality.

The same issue may become stronger in agentic systems that recursively consume their own outputs. A long-running agent often searches for evidence, summarizes it, stores the summary, and then uses that summary to decide what to inspect next. If each cycle slightly favors one interpretation, representational accumulation could create a feedback loop: early framing $\rightarrow$ biased summary $\rightarrow$ biased next query $\rightarrow$ more frame-consistent evidence $\rightarrow$ stronger internal state. Our experiments do not directly test such closed-loop amplification, but the observed dose response and persistence identify a mechanism through which it could arise. The same reasoning applies to multi-agent systems when agents exchange compressed summaries: repeated shared framing could become part of the effective state of later collective reasoning even when no individual message contains an explicit directive. This suggests that evaluating agent systems only on final-answer correctness may miss the stability and recoverability of their internal trajectory.

The real-world validation makes this concern less dependent on the benchmark construction. Earlier framing consistency in naturally structured conversations was associated with later response framing ($r = 0.441$), while counterfactual replay of the same conversation and final request produced a 0.454 A-versus-B gap. The observational association alone would not establish a framing mechanism because naturally occurring conversations vary in topic, intent, and style. The counterfactual replay is more informative because it asks whether changing earlier framing while preserving the real conversational

skeleton changes the model's later response. The practical implication is that history-dependent framing is not merely an artifact of deliberately constructed five-turn sequences, but a property worth monitoring in real long-running interactions.

More broadly, these findings suggest that evaluation of long-context models should distinguish at least two notions of robustness. The first is input robustness: whether a model responds appropriately to the information currently presented. The second is trajectory robustness: whether it reaches comparable states when different histories eventually provide equivalent current information, and whether it can recover efficiently when earlier context has pushed it toward an unhelpful interpretation. Existing multi-turn work already shows that models can get lost after early mistakes (Laban et al., 2026); our results suggest that even subtler semantic organization of otherwise matched information may create a related form of path dependence. For future long-horizon assistants, robustness may therefore require not only better context windows, but explicit mechanisms for detecting accumulated state, re-grounding on source evidence, or periodically re-evaluating conclusions from a history-independent reference point.

E Additional Limitations and Future Directions

Second, our primary representation analysis assumes that framing can be usefully approximated by a linear direction in residual-stream space. Prior work provides substantial evidence that high-level properties can often be manipulated through activation directions, but more expressive representations may capture contextual state more faithfully. Recent methods have explored low-rank nonlinear activation interventions and sparse-autoencoder-based representation spaces, suggesting alternatives to a single linear axis (Jiang et al., 2025). At the same time, sparse features should not automatically be treated as unique mechanistic units: recent evidence shows that SAE features need not form a canonical or atomic decomposition of model computation (Leask et al., 2025). A useful next step is therefore to compare linear directions, low-rank subspaces, and sparse feature representations using the same history-dependence protocol, evaluating them not only by decodability but also by cross-topic generalization and causal intervention.

Third, the persistence studied here occurs entirely within a single model context. It should therefore be distinguished from persistent memory across sessions or explicit agent memory systems. This distinction becomes more relevant as deployed assistants increasingly use retrieval, summaries, or external memory stores to maintain information over much longer horizons. LongMemEval shows that current assistants still experience substantial degradation on information retained across sustained interactions and decomposes memory systems into indexing, retrieval, and reading stages, while MemBench further distinguishes factual from reflective forms of agent memory (Wu et al., 2025; H. Tan et al., 2025). Future work could examine whether semantic framing itself is preferentially written into, retrieved from, or amplified by these memory components. In particular, an agent that repeatedly summarizes its own prior outputs may create a closed feedback loop in which framing affects summaries, summaries affect later retrieval or queries, and the resulting evidence further reinforces the state. Testing this would require moving from fixed trajectories to interactive environments in which framing, memory writing, retrieval, and subsequent exploration are manipulated independently.

Fourth, the present experiments cover only a small set of factorized framing dimensions. Gain/loss, risk/benefit, and short-/long-term framing allow relatively strict factual matching, but real conversations contain richer forms of stance, persona, narrative emphasis, and social pressure. Multi-turn sycophancy work already demonstrates that stance changes unfold dynamically over conversation and that resistance to repeated pressure varies across tuning regimes (Hong et al., 2025). Extending the present paradigm to these more complex settings could reveal whether history-dependent state formation is a general property of semantic context or is strongest for particular classes of framing. Such studies should preserve the central identification principle of the current work—separating history from current input—while allowing more natural language variation.

Finally, our reversal experiments establish path-dependent updating but do not yet trace a complete hysteresis loop in the stronger dynamical-systems sense. Future experiments could repeatedly sweep framing strength from A to B and back again, measuring both behavioral and representational trajectories at each step. This would allow direct estimation of loop area, reversal thresholds, recovery time, and whether the system returns to its original state after symmetric exposure. More broadly, long-context benchmarks increasingly evaluate whether models can retrieve and reason over large

amounts of information (Bai et al., 2024), but relatively little evaluation asks whether models reach equivalent internal and behavioral states after informationally equivalent but historically different trajectories. We view this trajectory robustness—and the related question of recoverability from an accumulated contextual state—as a direction for evaluating long-horizon language models.