SAME REQUEST, DIFFERENT ANSWER: HOW CONVERSATION HISTORY SHAPES THE DECISION TO REFUSE OR COMPLY

Anonymous authors
Paper under double-blind review

ABSTRACT

The same harmful request may be refused when asked directly, yet answered after several rounds of conversation. We combine research on text-based multi-turn jailbreaks with mechanistic interpretability, holding the final request fixed on Llama-3.1-8B-Instruct and Crescendo attack histories and using history controls and internal-state interventions to trace how prior exchanges change refusal or compliance. We find that the actual exchanges between both participants and the assistant's most recent response affect attack outcomes, while retaining only length, topic, or one participant's content has difficulty reproducing the effect of the complete history. This effect can partly transfer with the response-opening state, namely the residual states at the template positions before the answer and the first few generated positions: transferring single-turn states can restore refusal, while transferring states from the attack history can induce compliance or recover part of the effect missing from failed histories. The effective components change with depth; the refusal direction can strongly change answers, but transferring it alone has difficulty restoring the effect of the complete history. Attention-source interventions further identify a path that reads explicit refusal wording in the preceding assistant reply, while timing comparisons show that the main effective stage for restoring refusal is at the opening of the answer, with later interventions changing answer content more often. These results suggest that the model's safety behavior is jointly regulated by multiple components that are sensitive to history and change across layers, and establish a mechanistic analysis workflow connecting history text, the response-opening state, and final behavior.

1 INTRODUCTION

Jailbreak attacks on large language models (LLMs) use inputs or interactions to induce safety-trained models to generate harmful content. By input channel, attacks can be text-based (e.g., GCG and PAIR) or multimodal (e.g., visual adversarial examples; Qi et al., 2024); by access to model information, they can be black-box (e.g., PAIR's iterative prompt rewriting; Chao et al., 2025) or white-box (e.g., GCG's gradient-based suffix search; Zou et al., 2023b). Conversation history further distinguishes single-turn from multi-turn attacks, including human multi-turn jailbreaks (Li et al., 2024) and Crescendo, which we use here. Repeated queries during prompt generation do not imply accumulated dialogue history; we focus on text-based attacks in the latter setting.

A safety-trained model may refuse a harmful request asked directly yet answer the same request after several exchanges. With the final request unchanged, its wording alone cannot explain this shift. We ask which internal states reflect history's influence and how they change the decision to refuse or comply.

Tools for mechanistic interpretability and representation analysis include linear probes, sparse autoencoders (SAEs), transcoders, activation patching, and representation steering. Probes read information from intermediate representations, but readout does not imply use during generation (Alain & Bengio, 2017); SAEs decompose activations into sparse features, and transcoders approximate feedforward mappings (Huben et al., 2024; Dunefsky et al., 2024). Activation patching tests the

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

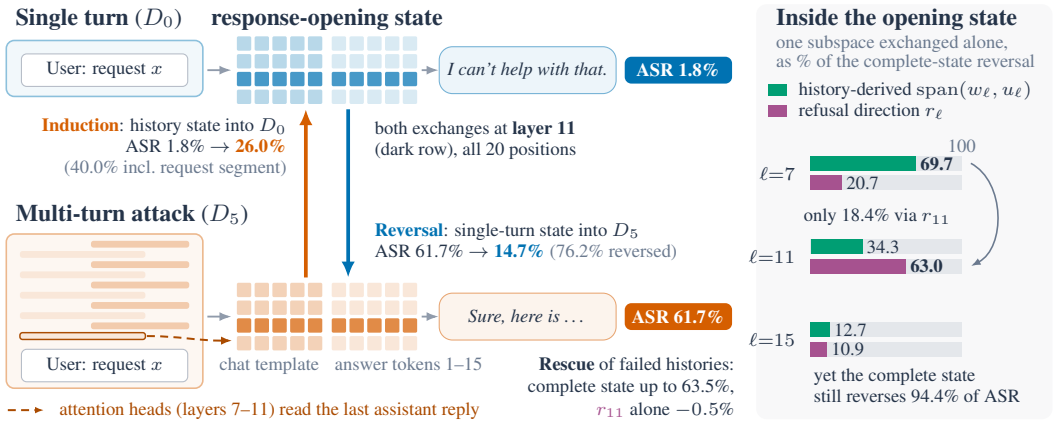


Figure 1: Same request, different answer. ASR for a byte-identical request x under single-turn D_0 and multi-turn D_5 conditions, based on 64 trajectories with 12 samples each. Vertical arrows denote layer-11 residual interchange; rescue is the share of lost ASR recovered in a user-only history (complete state from layer 19). Right: the six-seed cross-layer comparison in Figure 3; subspace effects as proportions of complete-state reversal are not additive. Pinning r at layer 11 blocks only 18.4% of the layer-7 effect of $\mathrm{span}(w_7, u_7)$.

effects of internal states by replacing activations (Zhang & Nanda, 2024), while representation engineering reads and manipulates representations to influence behavior (Zou et al., 2023a).

On Llama-3.1-8B-Instruct, we combine residual-state interchange, directional steering, and attention-head interventions, with linear probes and public dictionaries for auxiliary diagnostics. We fix a byte-identical, self-contained final request, vary only the history, and transfer internal states between paired contexts to test which history properties change behavior, which states carry the change, and whether restoring a component restores the history effect.

We connect text-based multi-turn jailbreaks with mechanistic interpretability through three contributions. First, we establish history controls with a fixed final request, separating request content from history and identifying history properties that affect refusal or compliance. Second, reversal, induction, and rescue interventions, combined with subspace comparisons and cross-layer pinning, locate transferable history effects and distinguish directions that change answers from components individually sufficient to restore history effects. Third, comparisons of attention sources and intervention timing connect history text, the response-opening state, and final behavior. This yields a mechanistic analysis workflow applicable to other text-based multi-turn dialogue sets and an exploratory account of safety behavior as sensitive to history, changing across layers, and involving multiple components jointly.

2 Related Work

Crescendo. Russinovich et al. (2025) propose Crescendo, a multi-turn jailbreak attack that starts with seemingly harmless questions and uses the model's prior replies to advance the conversation gradually, and evaluate its automated procedure across multiple models and tasks. It uses history accumulation and prior replies to influence subsequent answers, making it a suitable representative for our study of text-based multi-turn jailbreaks.

Activation patching and causal tracing. Meng et al. (2022) perturb inputs and restore selected internal activations to locate computational sites that causally affect factual predictions in language models.

Interchange interventions. Geiger et al. (2021) replace corresponding neural representations under different inputs to test whether they have the causal effects of the corresponding high-level variables.

Refusal directions and activation steering. Arditi et al. (2024) find that a single direction in the residual stream can strongly regulate refusal behavior in multiple chat models and analyze how adversarial suffixes suppress the propagation of this direction.

3 Preliminary: Experimental setup

Fixed request and history controls. The target model is Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Crescendo (Russinovich et al., 2025) generates five user–assistant exchanges H_i for each trajectory, and the final request x_i is a self-contained description of the benchmark target behavior. D_0 contains no history, and D_k retains the first k exchanges; all conditions use byte-identical x_i, the same chat template, and no added system instructions. History controls include length-matched padding H_{f}, unrelated benign dialogue H_{p}, same-topic dialogue without goal progression H_{r}, H_{u} and H_{a} retaining only the user's or assistant's substantive contributions, shuffled turns H_{o}, and H_{x} replacing only the preceding assistant reply with a fixed refusal. Construction details are in Appendix A.

Data. Target behaviors come from the goal set of Ren et al. (2025), HarmBench (Mazeika et al., 2024), JailbreakBench (Chao et al., 2024), and AdvBench (Zou et al., 2023b). We collected 314 successful trajectories confirmed by both judges, of which 64 (52 goals) are used for mechanistic analysis; selection rules and sample uses are in Appendix A.

Generation and evaluation. Answers are sampled at temperature 0.7, top-$p = 0.9$, and top-$k = 50$, with at most 1,536 new tokens. Attack success requires both the HarmBench classifier ($p_{\mathrm{yes}} > 0.5$) and the Qwen3-32B verifier to judge that the answer fulfills the request; unsuccessful answers are further classified as explicit refusal, unsuccessful compliance, non-substantive answers, or degeneration. ASR is averaged across seeds within each trajectory, then across trajectories. Interventions are paired with baselines on the same trajectories and seeds; batched generation for history controls and hooked generation for internal interventions each use their own baselines. Intervals use 2,000 target-clustered bootstrap resamples preserving pairing; seeds for each experiment are in Appendix A.

4 A mechanistic analysis framework from history controls to state transfer

We first define history effects, then use state interchange, subspace comparisons, cross-layer pinning, and source and timing interventions to test how history changes refusal or compliance.

4.1 Fixing the request: defining history effects and their properties

Write the history as $H_i = ((q_{i,1}, a_{i,1}), \ldots, (q_{i,5}, a_{i,5}))$. For a history transformation T, define

$$C_i(T) = \mathrm{chat}\big(T(H_i), x_i\big), \qquad D_{0,i} = C_i(T_\varnothing), \qquad D_{5,i} = C_i(T_{\mathrm{id}}), \tag{1}$$

where $T_\varnothing$ removes the history and T_{id} retains it; the final request and chat template are fixed as in Section 3.

Let $Y_{i,s}(C) \in \{0, 1\}$ indicate whether trajectory i passes both judges under seed s and condition C, and define

$$\begin{aligned} \mathrm{ASR}(C) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i} Y_{i,s}(C_i), \\ E_{\mathrm{hist}} &= \mathrm{ASR}(D_5) - \mathrm{ASR}(D_0), \\ L(T) &= \mathrm{ASR}(D_5) - \mathrm{ASR}(C(T)). \end{aligned} \tag{2}$$

E_{hist} measures the overall effect of the complete history, and $L(T)$ measures the effect lost through the transformation, both using paired seeds and matched generation implementations. D_k tests history accumulation, while the controls in Section 3 test the effects of length, topic, both participants' contributions, order, and the assistant's most recent reply; response types distinguish restored refusal from damage to generation quality.

4.2 PAIRED STATE INTERCHANGE: TESTING THE TRANSFER OF HISTORY EFFECTS

Denote the final request body positions by Q, and the five template positions from the final user message's end token through position p_0 that predicts the first answer token by G. Together, G and the first 15 generated positions form the response-opening window W. Comparing Q, W, and $Q + W$ tests the roles of request processing and the opening of the answer.

Let $h_{\ell,p}(C)$ be the residual stream at position p after decoder block ℓ in context C, and $\mathcal{Z}_{\ell,P}(C) = \{h_{\ell,p}(C) : p \in P\}$ the corresponding state. Fixing the request aligns Q and G token by token; generated positions are aligned through a shared answer prefix. Following interchange interventions (Geiger et al., 2021), we keep the target input unchanged and apply, for each $p \in P$,

$$h_{\ell,p}^{\text{target}} \leftarrow h_{\ell,p}^{\text{target}} + P_S\big(h_{\ell,p}^{\text{source}} - h_{\ell,p}^{\text{target}}\big). \tag{3}$$

P_S is the orthogonal projection onto subspace S; $P_S = I$ exchanges the complete residual, while other projections replace only that component and preserve its orthogonal complement. Source and target use separate KV caches, and the source advances with the target's generated prefix; implementation and numerical checks are in Appendix A.

Denote the intervention by $I_{A\to B}^{\ell,P,S}$. Reversal ($D_0 \to D_5$) tests whether an existing attack can be weakened, induction ($D_5 \to D_0$) tests whether compliance can be established without history, and rescue ($D_5 \to C(T)$) tests whether effects lost in a low-success-rate control history can be recovered. These distinguish behavioral regulation from restoration of history effects; sufficiency is limited to the specified target context, layer, and positions.

For a matched target baseline B, we report $\Delta(I) = \mathrm{ASR}(I) - \mathrm{ASR}(B)$ and define

$$\begin{aligned} U(I) &= \frac{\mathrm{ASR}(D_5) - \mathrm{ASR}(I)}{\mathrm{ASR}(D_5)}, && I : D_0 \to D_5, \\ J(I) &= \frac{\mathrm{ASR}(I)}{\mathrm{ASR}(D_5)}, && I : D_5 \to D_0, \\ R(I; C(T)) &= \frac{\mathrm{ASR}(I) - \mathrm{ASR}(C(T))}{\mathrm{ASR}(D_5) - \mathrm{ASR}(C(T))}, && I : D_5 \to C(T). \end{aligned} \tag{4}$$

For induction, we also report post-intervention ASR. $U = 1$ means eliminating all baseline attack success, $J = 1$ means reaching the complete history's ASR, and $R = 1$ means closing the gap between control and complete histories. Denominators must be positive; negative effects and effects above 1 are retained.

4.3 SUBSPACE COMPARISONS: DISTINGUISHING BEHAVIORAL REGULATION FROM HISTORY RESTORATION

We compute and fix history-derived directions on 34 definition trajectories $\mathcal{D}$:

$$w_\ell = \mathrm{unit}(\mathbb{E}_{i\in\mathcal{D}}[h_{\ell,p_0}(D_{0,i}) - h_{\ell,p_0}(D_{5,i})]), \tag{5}$$

$$u_\ell = \mathrm{unit}(\mathbb{E}_{i\in\mathcal{D}}[h_{\ell,p_0}(C_i(T_{\mathrm{x}})) - h_{\ell,p_0}(D_{5,i})]), \tag{6}$$

where $\mathrm{unit}(v) = v/\|v\|_2$, and T_{x} changes only the preceding assistant reply to a refusal; w_ℓ represents the difference between no history and complete history, and u_ℓ the difference between the refusal rewrite and complete history.

We construct the refusal direction r using the harmful–harmless activation contrast of Arditi et al. (2024); the main comparison uses the direction at the user message's end token in layer 11. Cross-layer comparisons construct r_ℓ, w_ℓ, u_ℓ in the same way; unindexed directions refer to layer 11. Controls include the five-dimensional refusal subspace R_5, the refusal contrast in multi-turn contexts r_{ctx}, the harmfulness contrast at the request's final token h, and two 64-dimensional subspaces orthogonal to $\mathrm{span}(w, u)$ (Appendix A).

We compare each subspace's share of the squared activation difference with its interchange effect. At the same layer, window, and paired baseline, define reversal capacity relative to the complete state as

$$m(S) = \frac{\Delta(I_{D_0\to D_5}^{\ell,W,S})}{\Delta(I_{D_0\to D_5}^{\ell,W,\mathrm{full}})}, \tag{7}$$

The denominator is the nonzero complete-window effect. Subspaces may overlap and interact, so effect ratios cannot be added. We also remove the r component from $\mathrm{span}(w, u)$ or jointly exchange r, w, u to test additional effects, and use induction and rescue to test the ability to establish compliance and restore history effects.

4.4 Cross-layer pinning: testing paths of influence

We perform interchange X at an earlier layer while using Equation 3 to restore a later-layer coordinate v at template positions G to the values of the unexchanged target run with the same seed. Define the proportion of reversal removed by pinning as

$$s(X \to v) = \frac{U(X) - U(X + \mathrm{pin}_v)}{U(X)} \tag{8}$$

where $U(X) > 0$. We combine layer-7 interchange with layer-11 pinning to test the dependence of earlier-layer effects on changes in later-layer coordinates.

4.5 Sources and timing: connecting history text to answers

Attention sources. We first replace the outputs of specified layers or heads over W with single-turn values. Then, using H_{x} as the source, we compare replacing complete head outputs, only keys and values at history positions, and only queries and keys and values from the final user message onward, to distinguish the effects of reading history from computation for the current request and answer. We also compare explicit-refusal rewrites with rewrites that withhold the requested content but contain no refusal wording. Head sets are fixed on the definition trajectories; alignment, head selection, and control-head construction are in Appendix A.

Intervention timing. We compare refusal-direction interchange over G, W, and throughout generation, and apply additive steering along w, u, r:

$$h_{\ell,p} \leftarrow h_{\ell,p} + \alpha d\, \mathbf{1}\{p \in P\}, \tag{9}$$

d is a unit direction, and one dose of α is the median paired difference between projections of D_0 and D_5 onto that direction on the construction trajectories. We fix the direction and dose and vary the position set P to compare onset and duration; different doses are compared separately. We count the proportion of explicit refusals among answers that change from successful to unsuccessful and use capability and benign-request evaluations to examine intervention costs.

5 Experimental results: from conversation history to refusal or compliance

Batched generation and state-interchange generation use different sampling implementations, with complete-history ASR of 0.534 and 0.617, respectively, on the mechanism set. Each effect below is calculated relative to the paired baseline of its corresponding generation implementation, following the sequence of history properties, state transfer, components and paths, and sources and timing.

5.1 History effects arise from exchanges between both participants and the assistant's most recent response

With the final request fixed, ASR on 200 single-turn-resistant trajectories rises from 5.8% after one exchange to 15.9% after four and remains unchanged after the fifth (Figure 2a). The answer to the same request thus changes with prior exchanges.

On the 64 mechanism trajectories, length-matched padding, unrelated dialogue, and same-topic dialogue without progression toward the goal all have difficulty reproducing the effect of the complete history; retaining only the user's or assistant's substantive contributions is likewise insufficient, while shuffling turns retains more of the effect but remains below the original history (Figure 2b). These controls show that actual exchanges between both participants are key to reproducing the complete effect, and the original ordering also contributes. Replacing only the preceding assistant

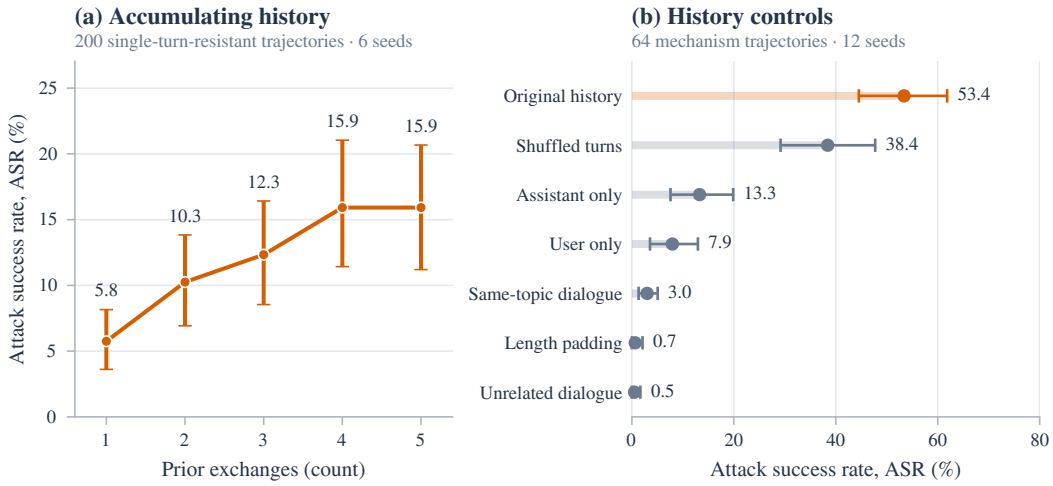


Figure 2: History effects under a fixed request. (a) History prefixes on 200 single-turn-resistant trajectories, six seeds each; (b) history controls under batched generation on 64 mechanism trajectories (Section 3), 12 seeds each. Points and labels show ASR (%); error bars are target-clustered 95% intervals. The panels differ in sample sets and axis ranges.

Table 1: Reversal, induction, and rescue of the layer-11 response-opening state (64 trajectories, 52 targets, 12 seeds per arm). As defined in Section 4.2, reversal U is the fraction of D_5 success removed, induction J is the induced ASR as a fraction of the matched D_5 ASR, and rescue R is the fraction of the D_5–H_u gap recovered. Brackets are target-clustered 95% intervals; — marks untested cells.

Transferred component	Reversal (%) $D_0 \to D_5$	Induction (%) $D_5 \to D_0$	Rescue (%) $D_5 \to H_u$
Complete residual in W	76.2 [65.4, 86.1]	42.2 [30.9, 53.9]	36.0 [25.2, 46.5]
Complete residual in Q	22.6 [14.5, 32.3]	11.2 [4.6, 19.7]	—
Complete residual in Q+W	—	64.8 [51.8, 77.7]	—
Refusal direction r in W	44.5 [36.5, 53.2]	21.5 [10.6, 34.1]	−0.5 [−4.2, 2.8]
History-derived $\mathrm{span}(w, u)$ in W	28.1 [19.9, 37.1]	12.9 [5.6, 21.5]	1.5 [−1.3, 5.6]
Target ASR without transfer (%)	61.7	1.8	9.2

reply with a refusal lowers ASR from 0.547 to 0.040 in the same run, further establishing the effect of the assistant's most recent response. Under each control condition, 82–99% of unsuccessful answers are explicit refusals, so subsequent analysis needs to explain how these history differences turn into a choice between refusal and compliance.

5.2 The response-opening state can transfer the behavioral effect of history

Table 1 compares reversal, induction, and rescue at layer 11. Replacing the response-opening window W in the complete attack history with single-turn states reverses 76.2% of attack success, substantially exceeding the 22.6% from exchanging only the request segment Q. Among paired answers that change from successful to unsuccessful, 92.5% become explicit refusals and none degenerates, indicating that this transfer changes outcomes mainly by restoring refusal.

Conversely, transferring W from D_5 into D_0 raises ASR from 0.018 to 0.260; jointly transferring $Q + W$ raises it to 0.400. The opening of the answer therefore carries history information that can induce compliance, with request-processing positions providing a complementary contribution. Transferring the complete W into failed control histories can also restore part of the lost effect (Table 7); specific component comparisons are given in Section 5.4. Together, the three types of transfer show that the response-opening state carries a substantial part of the history effect that can transfer between contexts.

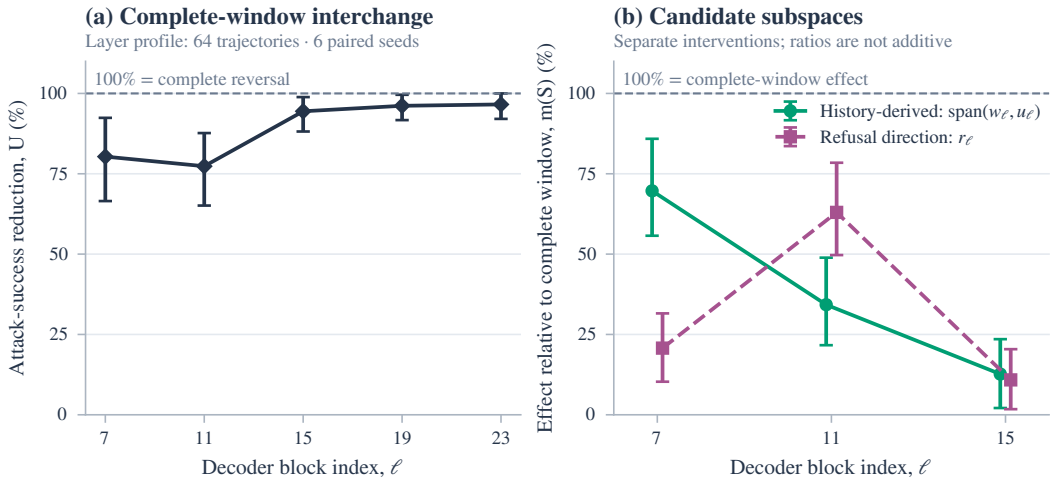


Figure 3: Effective subspaces change with depth. (a) Complete-window reversal, $100U$; (b) subspace effects relative to the complete window, $100m(S)$. 64 trajectories and six paired seeds (100–105); error bars are target-clustered 95% intervals. Layer 11 uses only the six corresponding seeds from Table 1, so values differ slightly; subspace points are slightly offset for readability.

5.3 Effective components change with depth, and the refusal direction participates in part of the paths

Under the same window and transfer operation, $\mathrm{span}(w_7, u_7)$ at layer 7 retains 69.7% of the complete-window effect, while the refusal direction r_7 retains 20.7%; at layer 11, the ordering reverses, with 63.0% for r_{11} and 34.3% for the history-derived subspace; by layer 15, each retains only 11–13%, while the complete window still reverses 94.4% of attack success (Figure 3). The effective components corresponding to history effects thus change with depth.

The size of a geometric difference cannot directly represent its behavioral effect: in the main 12-seed experiment at layer 11, r accounts for only 1.9% of the squared single-turn–history activation difference, yet reverses 44.5% of attack success; neither of the two 64-dimensional control subspaces accounting for 8–11% of the difference has a detectable effect (Table 5). Removing the r component from the history-derived subspace substantially reduces reversal, while adding w and u beyond r brings no gain (Table 6), indicating that the stronger effect of the history-derived directions at the same layer is related to their overlap with r.

Cross-layer pinning tests whether the history-derived subspace at layer 7 acts mainly through the refusal direction at layer 11. Fixing r at the layer-11 template positions to its unexchanged values removes only 18.4% of the reversal effect of $\mathrm{span}(w_7, u_7)$ at layer 7 (adjusted interval [6.8, 30.8]; Table 8). Thus, the refusal direction is a powerful intervention coordinate at an intermediate layer, but most of the behavioral effect of earlier-layer interventions can remain when this coordinate is fixed.

5.4 Rescue tests distinguish strong intervention directions from states that restore the effect of history

Rescue tests further examine whether directions that can restore refusal are sufficient to recover the compliance missing from failed histories. The user-only history already has an r value at the end token of the user message close to that of the complete history, yet still rarely succeeds. Transferring r or $\mathrm{span}(w, u)$ from D_5 gives rescue shares of only -0.5% and 1.5%, compared with 36.0% for the complete window at the same layer (Table 1). Complete-window rescue further reaches 63.5% [50.4, 76.2] at layer 19; it also exceeds transfer of r alone on padding and same-topic histories (Table 7).

These results distinguish strong intervention directions from the state information needed to restore the effect of history: the refusal direction can strongly change answers under existing attacks, but recovering the effect of failed histories requires information beyond the candidate directions. The rescue advantage of the complete state provides a basis for further identification of these components; we next examine how history text enters this state.

5.5 Attention-source interventions connect history text to the opening of the answer

Locating history-reading paths. At the response-opening window, replacing attention-head outputs in layers 0–7 or 8–15 with single-turn values reverses 81.2% and 91.8% of attack success, respectively, both higher than the two later-layer groups; each group also restores more than half of the natural difference along r at layer 11 (Table 11). The main attention carriers of history effects thus lie in layers 0–15.

Using the refusal-rewrite contrast u for the preceding assistant reply, we select 15 heads in layers 7–11 on the definition set. Restoring their single-turn outputs reverses 29.3% of attack success and restores about one-quarter of the natural difference along r, while matched control heads reverse only 3.0% (Table 11). With H_{x} as the source, replacing only the history keys and values of these heads on 219 evaluation trajectories reverses about 52%, close to complete-head replacement; replacing only queries and keys and values from the final user message onward gives about 14% (Table 3). This locates the effect of the preceding reply in these heads' reading of history text.

Identifying the textual features being read. On the same generation path, allowing these heads to read explicit-refusal text lowers ASR by 28.9 percentage points; reading implicit-refusal text that likewise withholds the requested content but contains no refusal wording changes ASR by only -1.5 percentage points [-4.8, 1.9]. We therefore call them refusal-reading heads, identifying a path from "explicit refusal wording in the preceding reply" through "attention reading" to "the response-opening state and behavior." When the entire reply is replaced, implicit refusal still lowers ASR by 27.2 percentage points [19.8, 34.8] in the corresponding batched run, indicating that the assistant reply also has features that act through other components.

Directions of influence of different reading components. When 16 counterbalancing heads in layers 6–11, selected for their contributions moving u toward compliance, are restored to single-turn outputs together with the refusal-reading heads, ASR instead rises by 11.2 percentage points [5.6, 17.2]. The complete history's combined change to these 31 heads is therefore protective, and the overall jailbreak effect mainly depends on other heads in layers 0–15. History reading involves components that act in different directions, and the refusal-wording path cannot account for all history effects.

5.6 Intervention timing distinguishes refusal restoration from changes in answer content

Exchanging r only at the five template positions retains 95.7% [83.9, 108.9] of the complete-window effect (Table 6), indicating that the main intervention effect of the refusal direction can be achieved before the first answer token is generated. Additive steering further shows that steering along w, u, or r from the 16th answer token onward gives weaker reversal than opening steering (Figure 4, Table 9). More crucially, among answers changed from successful to unsuccessful by opening steering, 56–80% become explicit refusals; with later steering, this share is at most 3.4%, and the remainder mainly remain compliant answers that do not meet the attack-success criterion. Thus, the main effective timing for restoring refusal is at the opening of the answer, while later ASR reductions more often reflect changes in answer content.

On 200 single-turn-resistant trajectories not used to define the directions, frozen w and u reverse 38.2% and 39.2% of attack success through short steering, respectively (Table 4), supporting the effectiveness of history-derived directions across samples.

Intervention timing also affects costs: in a duration control on 19 trajectories, continuous steering along w gives no stronger reversal than steering the first four positions, yet lowers GSM8K accuracy by 14.8 percentage points [10.3, 19.5]. Restricting one dose of w to the first 16 positions keeps the

point estimates of all costs within the preset tolerances; the GSM8K point estimate for one dose of u exceeds its tolerance, while two doses of r at template positions achieve a reversal effect of the same order but substantially increase over-refusal (Table 10). Among the tested directions and doses, short opening steering along w shows lower measured costs, indicating that mechanistic localization helps select intervention timing and directions.

5.7 Combined insight: history-sensitive regulation of safety behavior across layers

The evidence above supports an exploratory view: **in the model and multi-turn jailbreak setting studied, safety behavior exhibits internal regulation that is sensitive to history, changes across layers, and involves multiple components jointly.** Exchanges between both participants and the assistant's most recent response influence the response-opening state through earlier-layer attention, and different reading components can produce opposite effects; the refusal direction at an intermediate layer can strongly regulate answers, but restoring the effect of failed histories also requires information beyond the candidate directions. The opening of the answer is a key intervention stage at which these state changes influence refusal or compliance. The tests, from history controls and state transfer to components and paths, sources and timing, form a mechanistic analysis workflow connecting history text, internal states, and final behavior.

6 Discussion and limitations

Mechanistic implications. Joint tests of reversal, induction, and rescue distinguish the state information needed to change answers from that needed to restore the effect of history, supporting the exploratory explanation in Section 5.7: safety behavior is jointly regulated by multiple components that are sensitive to history and change across layers.

Extensions across models and scales. The complete quantitative analysis is limited to Llama-3.1-8B-Instruct and Crescendo. Similar phenomena of state transfer and subspace effects changing with depth have been preliminarily observed on Gemma3-4B and Qwen3-8B; ongoing analyses of the Gemma and Qwen families and larger models will compare transfer effect sizes, changes in effective components across layers, and the position of the refusal direction in the paths of influence, to test whether these characteristics remain consistent across model families and scales (Appendix C).

State structure still to be resolved. Subsequent work will seek components beyond the candidate directions by their ability to restore behavior and trace their history information through source interventions; it will also compare the effects of complete states at template positions and at the opening of generation to determine the range of positions needed to restore history effects. These tests will also be used for comparisons across models and scales.

7 Conclusion

In multi-turn conversation histories of large language models, exchanges between both participants and the assistant's most recent response influence refusal or compliance through the response-opening state, with effective components changing with depth. The refusal direction is a powerful intervention coordinate, but transferring it alone is insufficient to restore the effect of the complete history; the key intervention stage for restoring refusal is at the opening of the answer. These findings support an exploratory conclusion: in the setting studied, safety behavior is jointly regulated by multiple components that are sensitive to history and change across layers, and explaining its mechanism requires examining both strong intervention directions and the state information needed to restore the effect of history.

AI USE STATEMENT

In this research, we used generative AI tools for research ideation and execution, including refining hypotheses, designing and implementing experiments, interpreting results, and proving mathematical claims. We also used generative AI tools to search for and identify related work, draft parts of the paper, and polish the writing. In addition to these uses, language models were components of our experimental workflow, generating adversarial conversations and judging model responses (Section 3). We reviewed all AI-assisted work: the authors checked AI-assisted derivations and code for correctness, checked reported numbers individually against aggregate result files, and confirmed cited work against original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work examines failures of safety-trained models to support the analysis and mitigation of these failures. Its experiments involve harmful requests and model outputs, and the mechanistic findings may have dual-use implications. This study does not involve human subjects and did not undergo ethics review: attack conversations were generated between language models based on public benchmark goals, and all outcome labels came from automated judges. For ethical reasons, harmful text, namely attack histories and model answers, will not be released publicly; this manuscript reports only aggregate measurements and abstract experimental conditions.

REPRODUCIBILITY STATEMENT

Section 3 describes data sources and sizes, generation settings, judging criteria, and paired comparison methods; Section 4 defines history effects, state interchange, and cross-layer pinning and explains how source and timing tests connect history to answers. Appendix A records implementation checks, statistical procedures, and multiple-comparison families. All tables except Table 2, the numbers in Figure 1, and Figures 2–4 are regenerated by scripts from aggregate result files; the scripts copy estimates and intervals without re-estimating them, and another script checks the main estimates cited in the text against the same files. Appendix B.9 records the completion status of the main quantitative experiments, and Appendix C describes the ongoing cross-model extension and its preliminary observations. After further review of the ethical legitimacy of the experimental content, we will release reproducible code and materials. The planned release includes code for attack generation, judging, and condition construction; implementations of interchange, pinning, steering, and head-replacement interventions; history-derived directions, control subspaces, and head sets; configurations, seeds, and exclusion lists; and aggregate result files and scripts for regenerating all tables and figures and checking the numbers cited in the text. Code or data are not currently available.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations, Workshop Track*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 55005–55029, 2024. doi: 10.52202/079017-1745. URL https://proceedings.neurips.cc/paper_files/paper/2024/

hash/63092d79154adebd7305dfd498cbff70-Abstract-Datasets_and_Benchmarks_Track.html.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42, 2025. doi: 10.1109/SaTML64287.2025.00010. URL https://ieeexplore.ieee.org/document/10992337.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515–11542, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24375–24410, 2024. doi: 10.52202/079017-0768. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2b8f4db0464cc5b6e9d5e6bea4b9f308-Abstract-Conference.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. Circuit-Tracer: A new library for finding feature circuits. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 239–249, 2025. doi: 10.18653/v1/2025.blackboxnlp-1.14. URL https://aclanthology.org/2025.blackboxnlp-1.14/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, pp. 7827–7845, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, et al. OpenAssistant conversations – democratizing large language model alignment. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, pp. 47669–47681, 2023. doi: 10.52202/075280-2064. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. LLM defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*, 2024. URL https://arxiv.org/abs/2408.15221.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21527–21536, 2024. doi: 10.1609/aaai.v38i19.30150. URL https://ojs.aaai.org/index.php/AAAI/article/view/30150.

Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. LLMs know their vulnerabilities: Uncover safety gaps through natural distribution shifts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24763–24785, 2025. doi: 10.18653/v1/2025.acl-long.1207. URL https://aclanthology.org/2025.acl-long.1207/.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The Crescendo multi-turn LLM jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2421–2440, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/russinovich.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. Verifying chain-of-thought reasoning via its computational graph. In *International Conference on Learning Representations*, pp. 122812–122839, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c7b5a35ea98b62512a869c19ea7b03cb-Abstract-Conference.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b. URL https://arxiv.org/abs/2307.15043.

APPENDIX

A PROTOCOL AND REPRODUCIBILITY DETAILS

Table 2: Final sample counts and uses. Counts refer to trajectories unless the number of goals is stated. The analysis subsets below overlap and their counts cannot be added: both the 64 mechanism trajectories and the 200 history-length / frozen-direction replication trajectories come from the 219 evaluation trajectories.

Sample use	Count	Description
Successful attack trajectories confirmed by both judges	314	Each contains five complete exchanges and a final request
Trajectories included in the final analysis	313	One additional trajectory excluded for sharing a goal with evaluation trajectories
Trajectories for evaluating history effects	219	172 goals; final requests satisfy the single-turn-resistance screening rule
Mechanistic analysis trajectories	64	52 goals; behavior changes with history during selection
History-length and frozen-direction replication trajectories	200	Not used to define w or u
Trajectories for constructing intervention directions and determining doses	94	64 goals; no goals shared with the 219 evaluation trajectories
Trajectories for defining w, u, and selecting attention heads	34	Fixed before the corresponding evaluations
Trajectories used in the duration control experiment	19	Used to compare continuous and opening steering

A.1 CONDITION CONSTRUCTION AND SAMPLE SELECTION

Each condition ends with the same self-contained benchmark request, and the suffix alignment of every paired context is checked before interchange. All retained histories contain five complete exchanges, no turns are cropped, and at most two successful trajectories are retained per goal. A "successful attack trajectory" means that, when the trajectory was generated, both judges classified its final answer as fulfilling the harmful goal; this does not require success in every resampling. "Single-turn resistance" means that the final request has no successful answer among 12 single-turn samples; it is a sample-selection rule, not a guarantee of refusal under all sampling.

Final sample and its uses. Of the 314 successful attack trajectories, 219 are used to evaluate history effects and 94 to construct intervention directions and determine doses; one additional trajectory is excluded because it shares a goal with evaluation trajectories, leaving 313 trajectories in the final analysis. The 94 trajectories used to construct interventions and the 219 used for evaluation cover 64 and 172 goals, respectively, with no shared goals. History directions w, u, and attention-head sets are determined on 34 trajectories and then kept fixed; the 200 trajectories used for the history-length curve and frozen-direction replication do not participate in defining these directions. Mechanistic analysis uses 64 of the 219 evaluation trajectories whose behavior changed with history during selection, covering 52 goals. The duration control on 19 trajectories compares continuous and opening steering; subsequent directions and control subspaces are determined using only the designated construction samples. The mechanism set is used for multiple analyses, and its results constitute exploratory evidence; the sensitivity check for goal overlap is given in Appendix B.9.

A.2 ATTACK HISTORIES, JUDGES, AND DICTIONARY TOOLS

Histories were generated offline with an implementation of the Crescendo attacker (Russinovich et al., 2025) that uses Qwen2.5-7B-Instruct as the attacker model. Each attack runs for up to six user turns, and every retained trajectory contains five complete exchanges before the final request; attacker messages are limited to 300 tokens and instructed to stay short, intermediate target replies to 300 tokens, and the final answer to 1,536 tokens. Each goal receives up to three attempts, and at most two successful trajectories are kept per goal. All sampling of the target model uses temperature 0.7,

top-p 0.9, top-k 50, and no repetition penalty. Attack success requires both the HarmBench classifier (Mazeika et al., 2024) ($p_{\text{yes}} > 0.5$) and the Qwen3-32B verifier. The classifier receives the target behavior and the answer in its standard prompt, and p_{yes} is its probability of answering yes rather than no. The verifier receives the target behavior, a category-specific criterion of what counts as a violation, and the answer; it decodes greedily and returns a JSON verdict whose binary success field gives its label. An unparseable verdict is regenerated once, and a verdict that remains unparseable counts as unsuccessful. The dictionary diagnostics use circuit-tracer, an open-source library for finding feature circuits (Hanna et al., 2025), together with publicly released transcoders (Zhao et al., 2026) and sparse autoencoders[1] for Llama-3.1-8B-Instruct. We rely on these widely used public tools rather than training dictionaries for this study, so that the diagnostics can be repeated with released weights.

A.3 Interchange and pinning implementation

Source and target contexts keep separate KV caches, and at each generated step the source is advanced with the target's generated prefix. For a basis matrix B with orthonormal rows, the projector is $P_S = B^\top B$. Numerical checks cover empty-space identity, self-interchange identity, restoration of the source projection, and preservation of orthogonal components, with a maximum projection error of 0.007 at the tested precision. From the comparison of explicit and implicit refusal rewrites onward, an existing baseline is reused on a new generation path only after 16 trajectories with two seeds are regenerated and match it; every such reuse in the later interchange, rescue, pinning, and head experiments matched (32/32). The two layer-11 arms shared by two implementations reproduce 768/768 generations.

Pinning applies the interchange operator at layer 11 over the five template positions, with the unexchanged D_5 run of the same seed as source, during prefill; the natural single-turn–history difference along r at generated positions is close to zero. Pinning hooks run before the logging hooks, and the logged displacement of the pinned coordinate is below 2×10^{-4}.

The R_5 baseline is constructed by singular-value decomposition of harmful activations centered by the harmless mean; it is a project-defined low-rank baseline rather than a reproduction of an optimized refusal-cone method. The two 64-dimensional control subspaces are orthogonal to $\text{span}(w, u)$ and follow the preregistered fallback rule on differences at p_0 on trajectories used to construct interventions: control A spans the 64 leading principal directions of the D_0–D_5 difference, and control B is a random subspace of the leading 128. Neither reaches its target of 80% of the share on the construction samples of $\text{span}(w, u)$ (0.40 and 0.30, against 0.59).

A.4 Classification and statistics

A stratified check of 653 responses gave Cohen's $\kappa = 0.669$ between a rule-based classifier and the verifier, below the prespecified 0.8; short refusals were often mislabeled as degeneration by the rules, so unsuccessful responses are classified by the verifier. These labels reflect model agreement, not human validation.

Intervals are target-clustered bootstrap intervals with 2,000 replicates and seed 0, except that the benchmark costs in Table 10 resample items (GSM8K paired by problem) and the replication in Table 4 resamples trajectories with 10,000 replicates. Multiplicity-adjusted comparison intervals use Bonferroni correction within each preregistered family: eight comparisons for the layer-11 interchange, six for the layer profile, five for the overlap and position arms, three or four for rescue, three for pinning, three for the late-versus-opening window comparisons, and four for the head-block and head-group comparisons. Six-seed arms are always compared with baselines restricted to the same six seeds.

A.5 Outcome types and utility

Among unsuccessful outcomes, the tables distinguish explicit refusal, compliance judged unsuccessful, non-substantive responses, and degeneration. A reduction in attack success through degeneration is not interpreted as restored refusal. The prespecified cleanliness rule for the main

[1] https://huggingface.co/andyrdt/saes-llama-3.1-8b-instruct

mechanism rows requires at least 60% explicit refusals and at most 10% degeneration among paired reversals; complete-window reversal, r, and $\mathrm{span}(w, u)$ meet it (92.5%, 78.4%, and 60.5% refusals, no degeneration), whereas the request-segment reversal in Table 1 does not (52.2% refusals, 38.2% compliance judged unsuccessful).

A benign multi-turn check appends a matched harmless request to 49 attack histories and uses six seeds. Its generation cap was lowered from the preregistered 1,536 to 96 tokens after two slow restarts; under a fixed seed the opening does not depend on the cap, so the check reads refusal openings but not complete-answer quality. No refusal opening was observed in any arm; a broader refusal criterion flags 0.3% of the answers in the complete-window arm.

The capability benchmarks use fixed items: the first 400 GSM8K test problems (Cobbe et al., 2021), 500 MMLU test questions (Hendrycks et al., 2021) at evenly spaced indices of the subject-ordered test split, all 450 XSTest prompts (250 safe and 200 unsafe contrast prompts; Röttger et al., 2024), 300 prompts drawn from OR-Bench-hard-1k (Cui et al., 2025) with seed 0, and 200 English OASST1 threads (Köpf et al., 2023) with at least two user turns, drawn with seed 0 and cut after a user turn.

Random seed settings. Each experiment uses fixed seeds for paired comparisons on the same trajectories.

- Attack generation uses seed a for the a-th attempt on each goal (0, 1, 2), and goals from each source are shuffled with seed 0.
- Single-turn-resistance screening, history controls, frozen-direction replication, the main layer-11 subspace interchange, induction, subspace-overlap and intervention-position comparisons, rescue of user-only histories, attention-head output replacement, and history-source comparisons all use seeds 100–111, for a total of 12 seeds.
- The history-length curve, cross-layer profile, interchange of the two 64-dimensional control subspaces, rescue of padding and same-topic histories, cross-layer pinning, steering-timing comparisons, and the benign multi-turn refusal-opening check all use seeds 100–105, for a total of six seeds.
- Baseline-reuse checks use seeds 100–101 on 16 trajectories; answer generation for capability and over-refusal benchmarks uses seed 100, and item sampling uses seed 0.
- Target-clustered bootstrap resampling and item-level bootstrap resampling in Table 10 use seed 0.

No intermediate turn of any retained history reached its 300-token cap, and 1.6% of final answers in the labeling runs for sample selection reached the 1,536-token cap.

A.6 History controls and attention-source comparisons

History controls replace substantive contributions while preserving the condition-specific template structure. Padding and replacement texts are length-matched under the model's tokenizer, and the final request remains byte-identical. Replacement strings, matching tolerances, and shuffle seeds will accompany the reproduction materials described in the reproducibility statement.

In source-restricted head interventions, each replaced head is recomputed with its query, its keys and values at history positions, and its keys and values from the final user message onward each taken from either the target or the source context; all other heads are unchanged. The part of each context from the final user message onward (the request, the template positions, and the generated tokens) is token-identical across D_0, D_5, and H_{x}, so two contexts are aligned by offset within that part and differ only in history length. Keys keep the rotary phase of their own context; a query, rotated at its own position, is rotated further by the difference between the two history lengths before it meets the keys of the other context, so every query–key pair keeps the relative distance it has within a single context. The composed rotations deviate from direct rotation by at most 1.6×10^{-5}, and replacing complete head outputs along this path agrees with direct replacement to within 0.002 in the readout on all 314 trajectories. The source history H_{x} replaces the last assistant reply with a refusal. Table 3 reports the resulting source-localization effects with intervals; on the mechanism set, the first 32 tokens of the explicit refusal rewrite reproduce 70.7% [61.6, 80.5] of its effect

(55 trajectories, 46 targets). Head sets were fixed on the 34 definition trajectories before the head experiments. Candidate heads were all heads in layers 0–14; for each head, its output at the five template positions was replaced during prefill with its value in the source context (H_x for u, D_0 for w), and the shift of the corresponding readout at p_0 was recorded as a fraction of the source–D_5 gap. Heads were retained when a two-sided Wilcoxon signed-rank test passed Benjamini–Hochberg control at $q = 0.05$ and at least 80% of trajectories shifted in the same direction, ranked by mean shift, and kept if they were selected in at least 60% of 200 bootstrap resamples of the definition set; this yields 15 rather than 16 refusal-reading heads. The counterbalancing heads are all heads that pass the same test with shifts toward compliance, and each control head is a non-significant head with the smallest shift in the layer of a refusal-reading head (12 of 15 in the same layer, the others in adjacent layers). Likewise, each of the 16 neutral heads matched to the counterbalancing heads in Section 5.5 is the non-significant head with the smallest shift in the layer of one counterbalancing head (11 of 16 in the same layer, the others in adjacent layers). Head identifiers (layer.head): refusal-reading heads (15): 7.20, 8.2, 8.7, 9.16, 9.19, 9.25, 9.27, 9.29, 10.5, 10.9, 10.25, 10.26, 11.8, 11.15, 11.30; counterbalancing heads (16): 6.5, 7.23, 8.10, 8.23, 9.5, 9.14, 9.17, 9.18, 9.28, 10.0, 10.6, 10.10, 10.17, 11.7, 11.20, 11.23; matched control heads (15): 7.2, 8.16, 8.17, 8.25, 8.31, 9.10, 9.12, 9.21, 10.11, 10.21, 10.24, 11.1, 11.10, 11.13, 11.19; heads selected for w (16): 4.2, 4.16, 5.5, 6.22, 6.24, 7.4, 7.6, 7.12, 7.13, 7.14, 7.15, 7.20, 7.21, 7.22, 7.25, 7.29.

Table 3: Source-restricted replacement of the 15 refusal-reading heads with values from H_x, whose last assistant reply is a refusal (219 evaluation trajectories, 172 targets, 12 seeds; baseline ASR 0.177). U is the relative reversal of attack success; the last column is the change in ASR in percentage points with a target-clustered 95% interval. The lower rows are paired differences between arms.

Replaced components	ASR	U (%)	ΔASR (points)
Complete head outputs	0.089	49.7	−8.8 [−11.3, −6.3]
History-side keys and values only	0.085	51.8	−9.2 [−11.6, −6.7]
Queries and final-message keys and values only	0.152	14.0	−2.5 [−3.8, −1.1]
History-side only minus complete outputs			−0.4 [−1.0, 0.3]
Final-message side only minus history-side only			6.7 [4.7, 8.8]

A.7 Evidence provenance and regeneration

The aggregate results used in the paper are archived with the hashes of their source files. All tables except Table 2, and the numbers in Figure 1, are generated from these aggregates by `scripts/make_tables.py`, and Figures 2–4 by `scripts/make_figures.py`; both scripts copy estimates and intervals without re-estimation and record input hashes. The history curves, additive replication, and interchange experiments have distinct analysis sets and generation paths.

B Additional intervention results

B.1 Replication and complete layer-11 comparison

The directions in Table 4 were fixed on the original definition set and applied to the 200 single-turn-resistant trajectories used for history-length analysis and frozen-direction replication; additive steering acts at p_0 and the next 15 generated positions. Table 5 reports the paired interchange comparison, whose window W also includes the four template positions before p_0. The geometric share g is the squared norm of the projected $D_0 - D_5$ difference divided by the squared norm of the complete difference, aggregated over the window. The refusal direction holds $g = 0.019$ but reaches $m = 0.584$ [0.467, 0.723], whereas the two 64-dimensional control subspaces (Appendix A), with g of 0.106 and 0.084, have no effect. Three preregistered predictions failed here: that r would reach at most a quarter of the complete-window effect, that $\mathrm{span}(w, u)$ would reverse more than r, and that continuous exchange of r would stay below $\mathrm{span}(w, u)$. The five-dimensional refusal subspace R_5 does not exceed r, as predicted, but it exceeds $\mathrm{span}(w, u)$ by 0.129 [0.020, 0.233], an outcome the preregistration had listed as falsifying.

Table 4: Additive steering with frozen directions on the 200 single-turn-resistant trajectories, 12 seeds per arm, at layer 11 for $k = 16$ opening positions. Intervals are paired 95% bootstrap intervals for the absolute change in ASR that resample trajectories (10,000 replicates). This operator differs from the interchange in Table 5.

Arm	ASR	U	Δ [95% interval]
Zero-dose baseline	0.168	—	—
w, one dose	0.104	0.382	-0.064 [-0.086, -0.044]
u, one dose	0.102	0.392	-0.066 [-0.088, -0.047]
r, one dose	0.139	0.174	-0.029 [-0.042, -0.017]

Table 5: Layer-11 interchange from D_0 into D_5 (64 trajectories, 52 targets). The matched 12-seed baseline ASR is 0.617; the six-seed control subspaces, 64-dimensional and orthogonal to $\mathrm{span}(w, u)$ (A: leading principal directions of the construction-sample difference at p_0; B: random within the leading 128; Appendix A), use their six-seed baseline (0.609). g is the fraction of the squared $D_0 - D_5$ activation difference in the exchanged subspace over the window; m is the effect relative to complete-window interchange. Intervals are unadjusted target-clustered 95% bootstrap intervals.

Subspace and positions	Seeds	g	ASR	U [95% interval]	m
Complete residual, W	12	1.000	0.147	0.762 [0.654, 0.861]	1.000
r, W	12	0.019	0.342	0.445 [0.365, 0.532]	0.584
r, throughout generation	12	0.019	0.346	0.439 [0.352, 0.532]	0.576
R_5, W	12	0.031	0.365	0.409 [0.327, 0.506]	0.537
$\mathrm{span}(w, u)$, W	12	0.160	0.444	0.281 [0.199, 0.371]	0.368
Control subspace A, W	6	0.106	0.612	-0.004 [-0.129, 0.112]	-0.006
Control subspace B, W	6	0.084	0.602	0.013 [-0.092, 0.108]	0.017
Other contrasts					
r_{ctx}, W	12	0.004	0.604	0.021 [-0.030, 0.072]	0.028
Complete residual, Q	12	1.000	0.478	0.226 [0.145, 0.323]	0.296
h, request-final token	12	0.002	0.637	-0.032 [-0.084, 0.007]	-0.042

B.2 Subspace-overlap and intervention-position experiments

Table 6 reports a second implementation of layer-11 interchange. Removing the r component from the history-derived directions reduces their reversal from 0.278 to 0.078 [0.004, 0.155]; under the five-comparison correction, the interval for the corresponding absolute change, -0.048 [-0.108, 0.007], includes zero, and the r component accounts for 0.200 [0.125, 0.288] of the span's reversal. Exchanging $\mathrm{span}(r, w, u)$ gives no gain over r alone (-0.021 [-0.113, 0.066]). Both results miss their preregistered predictions: a residual reversal of at least 0.10 without the r component, and a gain over r alone. At layer 11, the history-derived directions and the refusal direction therefore do not act as independent channels.

Exchanging r only at the five template positions retains 0.957 [0.839, 1.089] of its window effect, and the end token of the user message alone retains 0.702 [0.581, 0.825] of the template-position effect. These position-control experiments locate where the exchange acts; they do not test an intervention that begins after the opening window.

B.3 Rescue across failed control histories

Table 7 extends the rescue comparison to padding and same-topic histories and to deeper layers for the user-only history. Complete-window transfer exceeds transfer along r for every history, with multiplicity-adjusted differences of 0.365 [0.228, 0.510] for the user-only history, 0.294 [0.143, 0.449] for padding, and 0.260 [0.130, 0.421] for same-topic dialogue. Transfer along r is small but nonzero for the same-topic history (0.091 [0.022, 0.173]), and transfer along $\mathrm{span}(w, u)$ rescues none of the padding gap, against a preregistered prediction of at least 0.30. Transferring the request segment instead of the window restores nothing. Among rescued histories, explicit refusal remains the most common unsuccessful outcome and no response degenerates.

Table 6: Subspace-overlap and intervention-position experiments in a second layer-11 implementation (64 trajectories, 52 targets, 12 seeds; matched baseline ASR 0.617). The two interchange interventions shared with Table 5 reproduce its generations byte for byte (768/768 pairs); independent judging changes their reversal by -0.008 [-0.020, 0.002] and -0.002 [-0.012, 0.007]. The last column gives explicit refusals among paired responses that change from successful to unsuccessful.

Exchanged component	ASR	U [95% interval]	Refusal flips
r, W	0.348	0.437 [0.355, 0.525]	79%
$\mathrm{span}(w, u)$, W	0.445	0.278 [0.197, 0.369]	60%
$\mathrm{span}(w, u)$ with the r component removed, W	0.569	0.078 [0.004, 0.155]	43%
$\mathrm{span}(r, w, u)$, W	0.361	0.416 [0.320, 0.517]	74%
r, five template positions G only	0.359	0.418 [0.323, 0.522]	81%
r, final user-message end token only	0.436	0.293 [0.211, 0.382]	74%

Readouts and transfers together show that what these histories lack is not the D_5 value along r. At the end token of the user message, the r projection is 3.39 for D_0, 2.25 for D_5, and 2.37 for the user-only history, whose ASR on the interchange path is nevertheless 0.092. Padding and same-topic histories keep r near its single-turn level (3.58 and 3.14), and transferring the D_5 value still restores at most 0.091 of the gap.

At layer 19, where complete-window rescue of the user-only history is largest, we removed the components along w_{19}, u_{19}, and r_{19} (11.5% of the squared D_5–user-only difference at p_0 on trajectories used to construct interventions) and took the leading directions of the remainder, which account for half of the difference. Transferring the leading three restores -0.052 [-0.141, 0.009] of the gap, and the reverse exchange in D_5 reverses 0.068 [0.015, 0.124]. The first of them separates the user-only and padding histories from all others; its condition-level Spearman correlation with a model-based score of previously supplied content has magnitude 0.22 (the sign of a singular vector is arbitrary). All three results miss their preregistered predictions: half of the complete-window rescue, a reversal of at least 0.20, and a correlation of at least 0.5. The deep-layer information that decides success is therefore not captured by the directions of largest variance.

Table 7: Rescue of failed control histories by transferring D_5 values (64 trajectories, 52 targets; D_5 ASR 0.617 over 12 seeds for the user-only history and 0.609 over the 6 seeds used for the others). Entries are normalized rescue R at layer 11 unless stated, with target-clustered 95% intervals; every ratio uses the paired baseline of its history.

Target history	Baseline ASR	Complete W	r in W	$\mathrm{span}(w, u)$ in W
User only	0.092	0.360 [0.252, 0.465]	-0.005 [-0.042, 0.028]	0.015 [-0.013, 0.056]
Length padding	0.016	0.346 [0.212, 0.475]	0.053 [-0.010, 0.129]	0.000 [0.000, 0.000]
Same-topic dialogue	0.039	0.352 [0.234, 0.468]	0.091 [0.022, 0.173]	0.023 [-0.016, 0.068]
User only, complete W at layer 15		0.591 [0.472, 0.695]		
User only, complete W at layer 19		0.635 [0.504, 0.762]		
User only, complete Q at layer 15		-0.020 [-0.054, 0.007]		

B.4 Cross-layer pinning

Table 8 reports the pinning test of Section 5.3. The layer-7 exchange of $\mathrm{span}(w_7, u_7)$ reproduces the generations of the layer-profile arm byte for byte, and independent judging changes 5 of their 384 labels (0.581 against 0.560). Pinning r or u at layer 11 removes the same 18.4% of this reversal, just below the preregistered range for partial routing (20–50%); the interval for r excludes a majority route. Pinning r removes 8.0% of the complete layer-7 effect, with an adjusted interval that includes zero. Pinning leaves the response types largely unchanged: explicit refusals make up 80.4% of paired reversals without pinning and 80.3% with r pinned.

Table 8: Cross-layer pinning (64 trajectories, 52 targets, six seeds; matched baseline ASR 0.609). A layer-7 exchange from D_0 into D_5 is repeated while a layer-11 coordinate is pinned, over the template positions, to its value in the unexchanged D_5 run (largest mean displacement of the pinned coordinate 1.1×10^{-4}). The routed share s is the fraction of the layer-7 reversal removed by pinning; its intervals are Bonferroni-adjusted for three comparisons. Other intervals are target-clustered 95% intervals.

Intervention	ASR	U (%)	Routed share s (%)
$\mathrm{span}(w_7, u_7)$ exchanged at layer 7	0.255	58.1 [45.7, 71.7]	—
with r pinned at layer 11	0.320	47.4 [34.9, 61.5]	18.4 [6.8, 30.8]
with u pinned at layer 11	0.320	47.4 [33.5, 61.9]	18.4 [6.1, 34.3]
Complete residual exchanged at layer 7	0.120	80.3 [66.5, 92.4]	—
with r pinned at layer 11	0.159	73.9 [58.1, 88.7]	8.0 [−0.7, 18.7]

B.5 Timing of additive steering

Table 9 reports the window scan of Section 5.6, with one dose of each direction at layer 11 unless stated. Late steering is weaker than steering the matching opening window, with multiplicity-adjusted differences in U of -23.2 [-37.4, -9.9] points for w, -17.3 [-31.5, -5.4] for u, and -13.9 [-27.7, -2.0] for r. Two preregistered predictions did not hold: late steering along w and u reverses more than the predicted 10% (11.4% and 16.9%), and late steering along r, predicted to keep at least half of its full-generation effect (26.6%), has no detectable effect (5.5% [-1.5, 12.2]). Answers changed by steering that starts at position 4 or later are almost never refusals; 73–89% of them are compliant answers judged unsuccessful, and none is degenerate. Across all answers, explicit refusals make up 18–19% under such steering, as without steering (19.0%), and 32–52% under opening steering; these response-type comparisons are descriptive rather than preregistered. With one dose added from p_0, the refusal direction is a weak steering target (19.4% over the first 16 positions), whereas two doses added at each of the five template positions reverse 43.9%. On the mechanism set, the natural difference along r is about 1.3 doses averaged over these positions and 2.3 doses at the end token of the user message, so this arm approximates the natural displacement. The weak effect of one dose of r in Table 4 (17.4%) thus reflects where and how strongly it is applied rather than a weak role of r.

Table 9: Timing of additive steering at layer 11 (64 trajectories, 52 targets, six seeds; matched zero-dose baseline ASR 0.617). Position 0 is the prefill step, where the direction is added at p_0 unless stated; positions $i \geq 1$ are generated tokens. The last column is the share of explicit refusals among paired answers that change from successful to unsuccessful. Intervals are target-clustered 95% intervals.

Direction	Steered positions	ASR	U (%)	Refusals (%)
w	positions 0–3	0.427	30.8 [20.3, 43.4]	76
w	positions 4–15	0.591	4.2 [−3.2, 11.2]	3
w	positions 0–15	0.404	34.6 [23.9, 47.1]	74
w	positions 16–63	0.542	12.2 [3.7, 22.0]	2
w	position 16 onward	0.547	11.4 [4.2, 18.3]	2
u	positions 0–15	0.406	34.2 [23.3, 46.4]	67
u	position 16 onward	0.513	16.9 [9.6, 24.9]	3
r	positions 0–15	0.497	19.4 [10.4, 29.6]	56
r	position 16 onward	0.583	5.5 [−1.5, 12.2]	0
r	throughout	0.453	26.6 [17.9, 35.6]	45
r, three doses	throughout	0.229	62.9 [51.7, 75.0]	65
r, two doses	template positions G only	0.346	43.9 [32.8, 56.3]	80

B.6 Costs of steering

Table 10 reports utility and over-refusal for the steering arms of Section 5.6 (GSM8K 400 problems, MMLU 500, XSTest 250 safe prompts, OR-Bench-hard 300, OASST1 200 dialogues). The prespecified tolerances for one dose of w or u were losses of at most 2 points on GSM8K and MMLU,

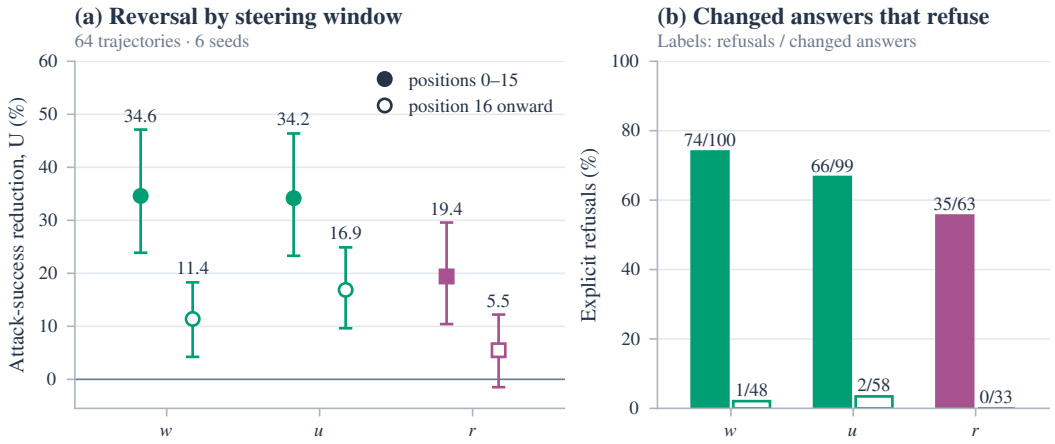


Figure 4: The decision is set at the onset of the answer. One dose along w, u, r at layer 11: the first 16 positions (filled) or from the 16th answer token onward (open). (a) Reversal, $100U$, with target-clustered 95% intervals; (b) the proportion of explicit refusals among answers that change from successful to unsuccessful.

increases in explicit refusal of at most 3 points on XSTest safe prompts and OR-Bench-hard and of at most 2 points on OASST1, and a win rate of at least 45%. Opening steering along w, u, or both keeps GSM8K between -2.5 and $+3.5$ points of the unsteered model, with intervals that include zero, MMLU within 2 points, and refusals in benign multi-turn dialogue within 1 point, while raising OR-Bench-hard refusals by up to 5.7 points when both directions are added. Continuous steering with three doses of r raises OR-Bench-hard refusals by 9.3 points, and two doses of r at the template positions raise them by 31.3 points and XSTest safe-prompt refusals by 4.0 points; as preregistered, this arm was not run on GSM8K, whose evaluation tool adds directions during prefill only at p_0. Against the prespecified tolerance of 2 points, the GSM8K estimate for one dose of u (-2.5 [-6.8, 2.0]) counts as a measured cost, so the preregistered cost prediction holds for w but not for u.

Table 10: Effect and costs of steering at layer 11 ($\times k$: k doses; 0–15: the prefill step and the first 15 answer positions). U is the relative reversal of attack success in the window scan (Table 9); the other columns are changes relative to the unsteered model in percentage points: GSM8K accuracy (400 problems, paired), MMLU accuracy, and the rate of explicit refusal on XSTest safe prompts, OR-Bench-hard, and benign multi-turn OASST1 dialogues, followed by the OASST1 pairwise win rate against the unsteered model (%). Brackets are 95% bootstrap intervals; — marks untested cells.

Steering	U	GSM8K	MMLU	XSTest safe	OR-Bench-hard	OASST1	Win rate
$w \times 1$, 0–15	34.6	0.0 [−4.3, 4.0]	−0.2	0.4 [−1.6, 2.4]	2.3 [−0.7, 5.3]	0.0 [−1.5, 1.5]	55
$w \times 2$, 0–15	—	0.5 [−3.5, 4.8]	−2.0	0.8 [−1.6, 3.6]	−4.7 [−9.7, 0.7]	−0.5 [−1.5, 0.0]	44
$u \times 1$, 0–15	34.2	−2.5 [−6.8, 2.0]	−0.2	−0.4 [−2.4, 1.2]	2.0 [0.0, 4.3]	0.5 [0.0, 1.5]	50
$u \times 2$, 0–15	—	−1.3 [−5.5, 3.0]	0.4	−0.4 [−2.4, 1.6]	1.7 [−1.0, 4.3]	0.5 [0.0, 1.5]	53
$(w+u) \times 1$, 0–15	—	3.5 [−1.0, 7.5]	0.8	1.6 [−0.8, 4.0]	5.7 [2.3, 9.3]	0.5 [0.0, 1.5]	50
$r \times 1$, 0–15	19.4	−0.5 [−4.0, 3.5]	−0.4	0.0 [−1.2, 1.2]	2.7 [0.7, 4.7]	0.5 [0.0, 1.5]	50
$r \times 3$, throughout	62.9	−1.8 [−5.8, 2.5]	−1.0	0.0 [−1.6, 1.6]	9.3 [6.0, 12.7]	1.0 [0.0, 2.5]	53
$r \times 2$, template only	43.9	—	−0.6	4.0 [1.6, 6.8]	31.3 [26.0, 36.3]	1.0 [0.0, 2.5]	51

B.7 Attention with single-turn sources

Table 11 reports head replacement with single-turn values over the opening window. Heads in layers 8–15 reverse more of the effect than heads in layers 16–23 or 24–31 (multiplicity-adjusted differences of 65.0 [51.4, 76.4] and 68.1 [55.7, 79.4] points), and not detectably more than heads in layers 0–7 (10.5 [−3.4, 26.0]). The refusal-reading heads reverse more than matched control heads (a difference of 16.3 [8.5, 24.8] points in ASR), whereas adding the counterbalancing heads changes the sign of the group's effect (a difference of 29.3 [17.3, 41.0] points), contrary to the

preregistered prediction that it would only weaken the reversal. The four preregistered comparisons use a Bonferroni factor of four; the comparison with layers 0–7 was added afterwards, and its interval includes zero even under this factor. Neither the layer-8–15 block nor the refusal-reading heads introduce refusals on the benign multi-turn check (0 of 294 answers each).

Table 11: Replacing attention-head outputs in D_5 with their single-turn values over the opening window (64 trajectories, 52 targets, 12 seeds; matched baseline ASR 0.617). The last column is the resulting shift of r at layer 11 over the template positions, as a percentage of its natural D_0–D_5 difference (zero by construction for heads above layer 11). Intervals are target-clustered 95% intervals.

Replaced heads	Heads	ASR	U (%)	r restored (%)
All heads, layers 0–7	256	0.116	81.2 [68.7, 92.5]	58.1
All heads, layers 8–15	256	0.051	91.8 [86.0, 96.4]	55.8
All heads, layers 16–23	256	0.452	26.8 [18.3, 36.7]	0.0
All heads, layers 24–31	256	0.471	23.6 [14.6, 33.7]	0.0
All heads	1024	0.033	94.7 [89.7, 98.2]	100.0
Refusal-reading heads (layers 7–11)	15	0.436	29.3 [20.0, 39.7]	26.0
Refusal-reading and counterbalancing heads	31	0.729	-18.1 [-30.5, -8.4]	-31.3
Matched control heads	15	0.599	3.0 [-4.8, 11.0]	-6.4
Heads selected for w (layers 4–7)	16	0.391	36.7 [24.9, 49.1]	25.4

B.8 Predictive analysis and dictionary diagnostics

Leave-one-condition-out models test whether candidate projections predict attack success under a held-out history manipulation. Held-out deviances are 12,852 for r, 11,403 for (w, u), and 11,048 for (w, u, r). Adding a model-based score of previously supplied content to (w, u) gives 10,873, an improvement that does not meet the prespecified significance threshold ($p = 0.129$). A separate target-grouped cross-validation fits linear probes to the opening residual; the selected layer-23 probe reaches condition-level Spearman correlation 0.976 and cell-level correlation of about 0.76, with 4.1% of its direction in $\mathrm{span}(w_{23}, u_{23}, r_{23})$.

Sparse-dictionary analyses serve as auxiliary diagnostics. In the tested layer-11 sparse autoencoder, about 86% of the relevant activation difference falls in reconstruction error, and transcoder attributions are spread over many features. These observations concern dictionary coverage and motivated the dense-subspace analyses of the main text.

B.9 Evidence status

All quantitative experiments on Llama-3.1-8B-Instruct reported in the main text and Appendices A and B were complete at the time of writing, and the manuscript uses their final results, including final response types and the preregistered analyses of the pinning, window-scan, cost, and head experiments. The ongoing cross-model extension and its preliminary qualitative observations are described separately in Appendix C. Two further preregistered predictions did not hold: that $m(r_\ell)$ would stay at or below 0.25 at every tested layer (0.63 at layer 11; Figure 3), and that $\mathrm{span}(w, u)$ would induce at least half as much as the complete window (0.25 of it; Table 1).

Sensitivity check for goal overlap. In one analysis, the trajectories used to construct interventions and evaluate effects shared a goal, whose evaluation trajectory belonged to the mechanism set, so the intervention settings may have been informed by that goal. Repeating the analysis on 63 mechanism trajectories covering 51 goals after removing that goal leaves the conclusions unchanged. The final sample excludes the overlapping trajectory used to construct interventions; this exclusion does not change the composition of the 64 mechanism trajectories.

C CROSS-MODEL EXTENSION AND PRELIMINARY OBSERVATIONS

C.1 SCOPE OF THE EXTENSION AND CURRENT PROGRESS

We are extending the mechanistic analysis workflow in this paper from Llama-3.1-8B-Instruct to the Gemma and Qwen families and further to models with larger parameter counts. Cross-family comparisons examine shared characteristics of history effects and their internal carriers across models; comparisons of scales within the same family examine whether the locations of effective states, the explanatory coverage of candidate subspaces, and history-reading paths change with scale.

At present, preliminary analyses of Gemma3-4B and Qwen3-8B have observed state-transfer and cross-layer subspace phenomena similar to those in Sections 5.2 and 5.3. This appendix records qualitative progress at this stage; systematic comparisons of larger models are ongoing, and complete quantitative results of the extension experiments will be reported after subsequent analyses are completed.

C.2 PRELIMINARY PHENOMENA CORRESPONDING TO SECTIONS 5.2 AND 5.3

The transferable effect of the response-opening state. Corresponding to Section 5.2, we also observe on Gemma3-4B and Qwen3-8B that, in paired contexts with a fixed final request and varying prior history, transferring the response-opening state can change the model's refusal or compliance behavior. This phenomenon suggests that the influence of history on answers can likewise be intervened on through internal states at this position. Subsequent work will summarize effects separately for reversal, induction, and rescue, comparing the state ranges needed to restore refusal, induce compliance, and compensate for failed histories.

Changes in candidate-subspace effects across layers. Corresponding to Section 5.3, preliminary analyses of both models also show that the coverage of history effects by candidate subspaces changes with depth. Cross-model comparisons will use the transfer effect of the complete state at the same layer as a reference, examine the relative effectiveness of history-derived directions and refusal directions separately, and analyze their relationships through cross-layer pinning. The comparison focuses on patterns of change in effective components; specific layers, effect shares, and proportions routed through the refusal direction will be determined by each model's quantitative results.

C.3 DESIGN OF SUBSEQUENT SYSTEMATIC COMPARISONS

Within-model pairing and cross-model alignment. The extension analysis will keep final requests byte-identical within each model, use each model's chat template, and define history controls and state-transfer operations as in Section 4. History-derived directions, refusal directions, and candidate head sets will be constructed separately within each model. Because models may differ in their numbers of layers and state dimensions, cross-model comparisons will report both actual layer indices and relative depth, aligning intervention ranges by the functional positions of the final request, chat template, and answer opening.

From transfer effects to paths of influence. The comparisons will first determine the complete state's capacity for reversal, induction, and rescue, then evaluate the effects retained by candidate subspaces and use cross-layer pinning to test the relationship between earlier-layer interventions and later-layer coordinates. On this basis, further attention-source and intervention-timing comparisons will test how history text reaches effective states and whether interventions change explicit refusal or the content of compliant answers.

Consistent reporting of effects and costs. Each model will separately report paired baselines, numbers of trajectories and goals, seed settings, effect estimates, and uncertainty, and undergo replication on samples not used to define directions and head sets. Intervention effects will be evaluated together with response types, general task capabilities, and over-refusal on benign requests to compare mechanistic characteristics and intervention costs across model families and scales.