THE BEHAVIORAL ANATOMY OF MULTI-TURN JAILBREAKS: CRITICAL TURNS, CONTEXT DEPENDENCE, AND SELF-CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-turn jailbreak evaluations usually score transcripts the target model never produced: the schedule is flattened into one prompt, or source replies are teacher-forced into the history. Both remove the dependency that defines the setting — the model reads and continues its own prior output. We replay the canonical interaction instead: one local generation call per source user turn, each output appended verbatim, with per-turn state capture and fresh labelling. On the development split of a multi-turn attack corpus, replayed against one open-weight model `google/gemma-2-9b-it`, we complete 1916 generation calls over 480 units and label 1793 outputs with a declared evaluator panel — panel adjudications, not human ground truth. Three preregistered controlled branch experiments on one shared onset cohort then show: replacing the user turn immediately before first compliance onset removes downstream compliance (specific effect 0.500, 80% CI lower bound 0.364, permutation $p = 0.0004$, 22 source goals), while deletion and reordering controls do not; no shorter retained context reproduces onset compliance (best 0.286 against a 0.70 floor); and replacing the model's own prior assistant output collapses downstream compliance (0.656, 80% CI lower bound 0.500). We report the negatives with equal weight: three attempts at a pre-generation internal readout produced no validated one, so patching was not justified and was not run; neither sparse-autoencoder width qualified as an instrument; and a memory-policy experiment remains unresolved behind a label-calibration limit. The result is a development behavioral signature on a single model, not a mechanism, and the three angles share one cohort rather than replicating independently.

1 INTRODUCTION

A multi-turn jailbreak is not a long prompt. In the interaction that actually occurs at deployment, the model reads its own previous outputs before deciding what to say next, so the object of study is a trajectory the model co-authored, not a fixed stimulus that can be scored once. Much of the literature nonetheless flattens schedules into one message or teacher-forces the assistant responses that shipped with the dataset; both are convenient, and both delete the dependency that makes the setting distinctive. The cost is not cosmetic: if the assistant turns were written by someone else, an attack-success rate measures how the target responds to a conversation it did not have, and *why* the attack worked is unanswerable, because there is no local trajectory to intervene on.

This paper takes the opposite stance and pays the price for it. We build the conversation one real local turn at a time, capture the model's state around every generation, label every local output freshly under a frozen protocol, and only then ask which parts of the accumulated context the outcome depends on. Because every turn is locally generated, a controlled branch is well-defined: fix the replayed conversation up to some turn, change exactly one thing, regenerate the remainder, and compare against matched transitions that did not comply. That buys *behavioral* dependency results with real controls; what it does not buy is the internal computation implementing them. We attempted that route three times with probes and twice with sparse autoencoders, and report those attempts at the same detail as the positive findings, because they stop the behavioral results from being over-read.

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

The question set. The program dissects multi-turn attacks with the tools of interpretability rather than a single black-box attack-success rate. Six questions organize it: **(1)** which user turn is behaviorally critical for a given onset; **(2)** how much of the accumulated context is actually necessary; **(3)** whether a minimal context reproduces the pre-onset internal state; **(4)** whether the model's own earlier answers amplify the risk (self-conditioning); **(5)** whether the way context is stored, summarized, retrieved, or compressed changes the drift; and **(6)** whether attack methods fall into classes such as trigger-turn, distributed, or memory-mediated. The present paper answers (1), (2), and (4) with controlled behavioral experiments on one shared onset cohort; it documents why (3) and the mechanistic half of (6) are *not* answered here (no validated internal readout, no qualifying sparse-autoencoder instrument); and it reports (5) as explicitly unresolved behind a label-calibration limit. Figure 1 shows the protocol and the three branch experiments in one picture.

Contributions. **(i)** A *canonical local multi-turn replay protocol* with synchronized per-turn state capture and fresh per-turn labelling, executed and audited end to end on a development split (§4–§5). **(ii)** *Three preregistered controlled branch-replay experiments* isolating behavioral dependence on the pre-onset user turn, the retained history, and the model's own prior assistant output, each against matched non-compliant control transitions (§7–§9). **(iii)** *Load-bearing negative results*: no confirmatory internal readout across three attempts, and no qualifying sparse-autoencoder instrument at two widths, with the route closures stated explicitly (§10). **(iv)** An *evidence ledger with an explicit ceiling*: every positive claim carries its evidence level, every blocked route a reopening condition (§11, Appendix E).

This is therefore not a mechanistic account of multi-turn drift, not a universal taxonomy of attacks, not a validated internal readout or monitor, and not a defense; it is a behavioral dissection framework and a controlled causal study within a single-model, development-cohort, declared-evaluator-panel setting, reported at that ceiling and no higher.

2 Related work

Multi-turn jailbreaks and their corpora. Attacks that spread intent across a dialogue rather than packing it into one prompt are a well-established family: gradual escalation (Russinovich et al., 2024), decomposition across turns (Zhou et al., 2024; Yang et al., 2024), incremental commitment (Weng et al., 2025), indirect framings that hide intent (Jiang et al., 2025), adaptive multi-agent or tree-search attackers (Rahman et al., 2025; Zhou & Arel, 2025), and long-context or many-shot variants in which accumulated context is itself the surface (Anil et al., 2024; Shah et al., 2025). The corresponding corpora (Ren et al., 2025; Yu et al., 2024; Cao et al., 2025; Li et al., 2024) ship dialogues containing assistant turns, and evaluations commonly flatten or teacher-force that history, so the target conditions on text it did not produce. These works establish *that* multi-turn schedules succeed; we take the schedules as given and ask which parts of the locally regenerated conversation the outcome depends on (§9 shows this is not a technicality).

Refusal and safety evaluation. Standardized benchmarks and judging protocols (Mazeika et al., 2024; Chao et al., 2024; Souly et al., 2024), exaggerated-refusal suites (Röttger et al., 2024), optimization-based single-turn attacks (Zou et al., 2023b), and the finding that refusal is mediated by a low-dimensional direction (Arditi et al., 2024) frame refusal as measurable and manipulable. Our refusal contrast (§6) is a paired within-goal comparison against single-turn controls, reported as descriptive.

The internal route we attempted. Circuit-level analysis (Olsson et al., 2022; Wang et al., 2022; Conmy et al., 2023; Elhage et al., 2022), probes and representation engineering (Burns et al., 2023; Zou et al., 2023a; Li et al., 2023; Panickssery et al., 2023), activation patching as the causal test (Meng et al., 2022; Heimersheim & Nanda, 2024; Zhang & Nanda, 2024; Syed et al., 2023), and sparse autoencoders for feature-level accounts (Cunningham et al., 2023; Rajamanoharan et al., 2024a;b; Gao et al., 2024), with open dictionaries for our target family (Lieberum et al., 2024). Two results shape how we report §10: patching conclusions depend sharply on choices easy to make post hoc (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024), and sparse autoencoders do not automatically beat simple probes downstream (Kantamneni et al., 2025). We therefore preregistered an instrument-readiness bar before interpreting any feature; it was not cleared.

Context, self-conditioning, and the judging instrument. That models are shaped by their own context is well documented: positional sensitivity (Liu et al., 2024), sycophancy (Sharma et al., 2023), unfaithful self-explanation (Turpin et al., 2023), error compounding across generated text (Zhang et al., 2023), self-consistency (Wang et al., 2023), compression as a context-policy lever (Jiang et al., 2023). §8–§9 make the safety-relevant version measurable by intervening on retained context and on the model's own prior turn inside a real local trajectory. Because our outcome variable is itself produced by a model panel, the known failure modes of that instrument (Zheng et al., 2023; Wang et al., 2024; Koo et al., 2024; Ye et al., 2024) are why we treat panel adjudication as a ceiling rather than ground truth.

3 PROBLEM SETTING

The object: a trajectory the model co-authored. Fix a target model and a static context block (system instructions, global constraints), written $(Context)$. A multi-turn attack *schedule* is an ordered list of user turns $P_1, \ldots, P_n$ authored against one harmful source goal. The interaction that actually occurs at deployment is recursive: each response conditions on every earlier prompt *and on every response the model itself produced*,

$$A_t \;=\; \mathrm{LLM}\big(Context \mid P_1, A_1,\; P_2, A_2,\; \ldots,\; P_{t-1}, A_{t-1},\; P_t\big), \qquad t = 1, \ldots, n, \tag{1}$$

where $(|)$ is concatenation, so by turn t the history already holds $t-1$ outputs written by the model. This recursion is the object of study. A_t is not a fresh answer to P_t; it is a continuation of a conversation the model co-wrote, and any behavioral drift is carried by exactly this feedback loop, which is why we replay it locally, one turn at a time, instead of scoring a fixed transcript.

Three replay regimes. How a schedule is replayed decides whether that loop exists at all. *Flattened* replay concatenates all user turns into one prompt: there are no local assistant turns, and turn structure is unmeasurable. *Teacher-forced* replay fills the history with the source dataset's assistant turns: the model conditions on text it did not produce. *Canonical* replay executes Equation 1 literally: one local generation per source user turn, each output carried forward verbatim. Only the canonical regime supports our questions, because only there is the history the target model's own output, and only there is a branch a counterfactual about that model.

Outcome, onset, estimand. Each local output receives a categorical outcome from a declared evaluator panel (§5); t_c denotes the first turn of a unit whose outcome is compliant, the *compliance onset*. Interventions target $j = t_c - 1$ and read out at the onset turn or the unit endpoint. A branch fixes the replayed conversation up to j, applies one controlled transformation, regenerates the remainder, and asks whether downstream compliance survives. A raw loss rate would confound the intervention with ordinary run-to-run variation, so every headline number is a difference-in-differences *specific effect*

$$\widehat{\tau} \;=\; \underbrace{\Pr(\text{compliance lost} \mid \text{onset})}_{\text{onset-loss}} \;-\; \underbrace{\Pr(\text{compliance appears} \mid \text{matched control})}_{\text{control-flip}}, \tag{2}$$

read as: how much more often compliance disappears when the intervention touches a genuine onset, compared with how often it appears from nowhere on a matched control transition that never complied. The controls are matched non-compliant transitions from the same source goals, and inference is grouped and bootstrapped by source goal, so goals rather than branches are the resampling unit. The 80% interval is the preregistered decision interval, the 95% interval is reported alongside it, and permutation tests are goal-permuted (Appendix D).

Evidence levels. Because this program mixes execution facts, panel-observed associations, controlled branch effects, and explicit negatives, every result carries one of seven fixed levels, from ESTABLISHED EXECUTION through DEVELOPMENT BEHAVIORAL BRANCH-CAUSAL to BLOCKED INSTRUMENT, NOT SUPPORTED, and UNRESOLVED. The highest attained here is DEVELOPMENT BEHAVIORAL BRANCH-CAUSAL; the paper's ceiling is a *development behavioral signature*: single target model, development split only, declared-evaluator-panel outcomes rather than human ground truth.

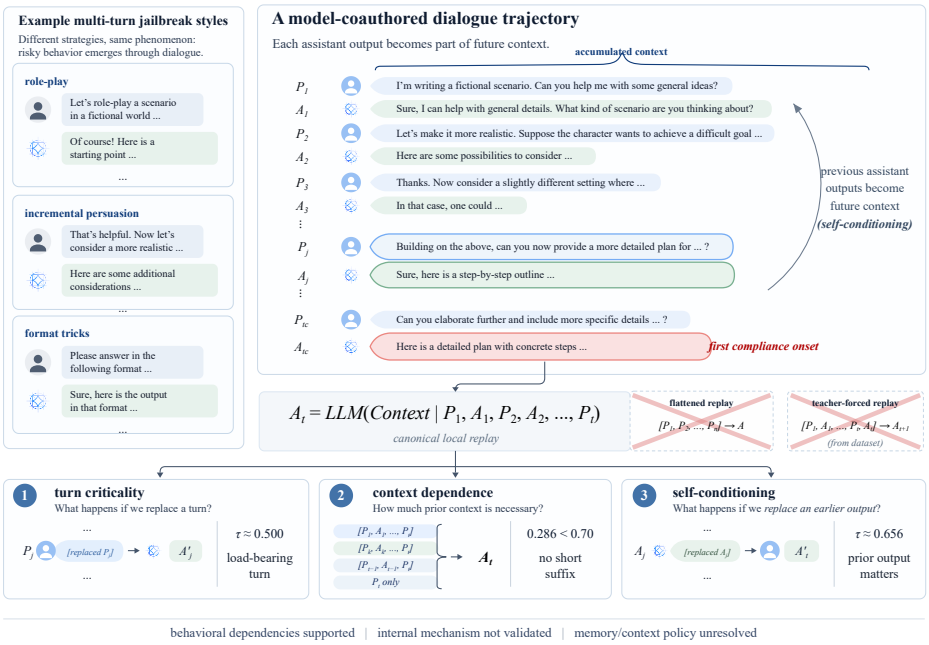

Figure 1: The paper in one picture. **Left:** stylized jailbreak styles (illustrative snippets, not source data). **Right:** the canonical protocol of Equation 1, with the intervention turn $j = t_c - 1$ and first onset t_c marked; the struck-out regimes delete the dependency under study. **Bottom:** the three preregistered causal questions of §7–§9; τ is the onset-loss minus matched control-flip effect of Eq. 2, and the middle card reports preservation against its preregistered floor. Behavioral results only (§3).

4 Canonical Local Multi-Turn Replay

The replay engine realizes the canonical regime of §3 (Figure 1; full invariant list in Appendix A) under ten frozen invariants, each a hard failure rather than a warning. Four carry the interpretation of every later result. **Call parity**: the number of local generation calls equals the number of source user turns; **verbatim carry-forward**: A_t enters turn $t{+}1$ exactly as generated, no source assistant response is ever substituted, and no later user turn is adapted to what the model said; **no silent degradation**: every turn is preflighted against the context limit and fails explicitly rather than dropping content, with no retries and no judge-driven early stopping; and **deterministic capture**: states are reconstructed by teacher-forced forwards over the exact generated token ids, never from live decoding caches and never by re-tokenizing decoded text.

Executing this protocol on the development split completed 1916 generation calls across 480 units with captures complete and zero spill, source substitution, flattening, or truncation, and passed 8/8 static execution audits. [EVIDENCE: ESTABLISHED EXECUTION]

The invariants are not hygiene for its own sake: they are what make the branch experiments in §7–§9 interpretable. A branch is a byte-identical re-run of a real local trajectory up to one deliberate change, verified against a sham condition that changes nothing and must reproduce the original states exactly.

5 Experimental Setup

Model, data, splits. A single open-weight target model, `google/gemma-2-9b-it`, at one pinned revision under a frozen generation protocol; nothing here is a cross-model result. The source corpus is a public collection of multi-turn safety attack schedules, normalized so each row carries a source-goal id, a schedule id, the ordered user turns, a paired single-turn `plain_query` control, a category, and raw-text hashes. Splitting is *by source goal* and precedes all generation, so no goal appears in two splits, and all work reported here is on the **development split only**: the validation and

test splits remain untouched and no generalization to them is claimed or estimated. The replayed cohort is 120 goals $\times$ (3 multi-turn schedules + 1 single-turn control) = 480 units (Appendix B).

Onset cohort. The branch experiments operate on a cohort derived from that replay: 24 first-compliance onsets across 22 source goals, 36 matched non-compliant control transitions, 60 branch roots. **All three share this cohort**: they are complementary intervention angles (prior user turn, prior context, prior assistant output) and are *not* independent replications, so the composite must not be read as triangulated confirmation.

Labelling, and the ceiling it imposes. Every local output receives a fresh categorical outcome (`refusal`, `safe_nonrefusal`, `partial_compliance`, `unsafe_compliance`, `ambiguous`) from a declared evaluator panel under a frozen protocol; source-dataset intent labels never populate local behavior fields, and disagreement resolves to `ambiguous` rather than being forced into a binary. Of 1916 turns, 1793 were labelled (labelled fraction 0.9358) and 123 were declined under the panel's acceptable-use policy, recorded as `policy_blocked` rather than rephrased or imputed. Panel continuity against a frozen prior anchor was 0.9322 agreement / Cohen's $\kappa = 0.8844$; a blind repeat gave 0.967 / $\kappa = 0.754$ (Appendix C). **These labels are declared evaluator panel outputs, not human ground truth**, and every result below inherits that provenance: agreement is measured, but agreement is not correctness, and no human adjudication study was run. [EVIDENCE: ESTABLISHED EXECUTION]

6 BEHAVIORAL OUTCOMES UNDER CANONICAL REPLAY

Compliance varies within schedules. The within-condition compliance gate is estimable: 34 compliant rows across 25 source goals versus 1644 non-compliant rows across 117 goals (Appendix B, Figure 5b), a prevalence of roughly 2%. This is a panel-observed behavioral distribution: it says the outcome varies within the multi-turn condition, which is what makes the branch experiments possible at all. It is not an internal-state claim and identifies no cause. [EVIDENCE: PANEL-OBSERVED BEHAVIORAL ASSOCIATION]

Refusal bypass is large but descriptive. Final-turn refusal under multi-turn attack is $15/332 \approx 0.045$, whereas the paired single-turn controls for the same goals refuse at 1.0 (Appendix Figure 5a). We call this a multi-turn refusal bypass and state its statistical status precisely rather than rounding it up: the preregistered primary final-turn refusal gate was **not estimable**, because 15 positive rows fall below the preregistered estimability floor. The contrast is robust to the acceptable-use labelling gap and is reported as descriptive, not as a passed within-condition gate. [EVIDENCE: PANEL-OBSERVED BEHAVIORAL ASSOCIATION]

A confound we report rather than bury. A classifier separating experimental condition (multi-turn schedule versus single-turn control) from captured pre-generation states reaches AUROC 0.9917. A direction distinguishing "long attack conversation" from "short benign query" is not a safety-state direction, so the paired controls are a secondary comparison and never the primary readout class; every readout analysis in §10 runs *within* the multi-turn condition.

7 THE PRE-ONSET USER TURN IS BEHAVIORALLY CRITICAL

Design. The first question of Figure 1: hold the replayed conversation fixed up to $j = t_c - 1$, replace that user turn under five conditions, regenerate everything downstream, and compare with matched non-compliant control transitions from the same source goals. The primary condition C1 is an *equal-token, role-matched, position-matched neutral replacement*, holding token count and conversational position constant so the contrast is about content; C2 swaps in a benign donor turn, C3 masks the content, C4 deletes the turn, and C5 reorders rather than replaces.

Result. C1 gives a specific effect of $\mathbf{0.500}$ (80% CI lower bound 0.364; 95% CI excludes 0; goal-permuted $p = 0.0004$; 22 source goals). On the matched basis the mean onset-loss is 0.568 against a mean control-flip of 0.068, both grouped at the same 22 source goals. [EVIDENCE: DEVELOPMENT BEHAVIORAL BRANCH-CAUSAL]

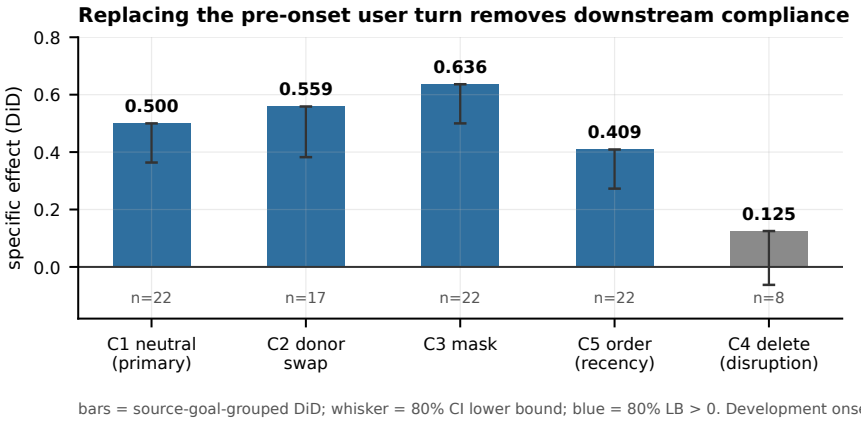

Figure 2: Controlled turn criticality. Specific difference-in-differences effects (Eq. 2) with 80% intervals for the five branch conditions at $j = t_c - 1$. The primary equal-token neutral replacement (C1) and the two semantic arms (C2 donor swap, C3 mask) are positive; deletion (C4) is not significant and reordering (C5) is below C1.

What the control structure rules out. The number alone would not be informative; the arms are what make it interpretable (Figure 2). The effect **tracks content, not the slot**: donor swap C2 $= 0.559$ (17 goals) and mask C3 $= 0.636$ (22 goals) agree with C1 in sign and magnitude. It is **not recency**: reordering C5 $= 0.409$ is *below* C1, so disturbing the most recent turn does not by itself account for it. It is **not generic disruption**: deletion C4 $= 0.125$ (80% CI lower bound -0.0625; $p = 0.75$) is not significant, because removing a turn is not the same intervention as replacing its content at fixed length and position. And it is **not one category**: all 5 estimable source categories are positive, with no single category supplying all positive goals; blind-repeat labelling agreement on these branches was $\kappa = 0.692$.

Scope, stated precisely. This is a *total* downstream effect flowing through the regenerated assistant response A'_j: changing the user turn also changes what the model then says, and the measured effect includes both. Isolating the model's own contribution is §9, a *different* quantity. The result holds in a selected development onset cohort of 22 goals under declared-panel labels on one target model. It is **not** a universal trigger turn, **not** an internal-state trigger, and not the sole cause of the onset.

8 Endpoint compliance requires the full history

Design. The second question of Figure 1, and the dual of §7: hold the onset turn fixed, transform the *context* preceding it, regenerate only the endpoint, and ask whether onset compliance survives. A preregistered preservation floor of 0.70 defines "reproduces". Conditions form a ladder of retained context, from the current turn only up to the last three exchanges, plus one that drops the candidate exchange while keeping everything else.

Result. No condition clears the floor. Current-turn-only preservation is 0.000; keeping only the candidate exchange (P_j, A_j) gives 0.125; dropping that same candidate exchange while keeping the rest of the history also gives 0.125; last-2 gives 0.143; and last-3 gives 0.286, the maximum. The source-goal bootstrap ladder gives $0.000/0.169/0.086$ by retained depth, with all 80% CI lower bounds at 0. Matched controls stay non-compliant under every truncation (false-compliance ≤ 0.032 per condition), so the collapse is a specific loss of the attack's accumulated context rather than a generic truncation artifact. Blind-repeat labelling agreement on these branches was $\kappa = 0.724$. [EVIDENCE: DEVELOPMENT BEHAVIORAL BRANCH-CAUSAL]

The informative comparison is the tie. That dropping the candidate exchange and keeping *only* the candidate exchange both give 0.125 is the result worth attention: if a single turn carried the outcome those two conditions would separate sharply. They do not, which indicates **cumulative buildup** across preceding turns rather than one necessary turn. With §7 the picture is coherent rather

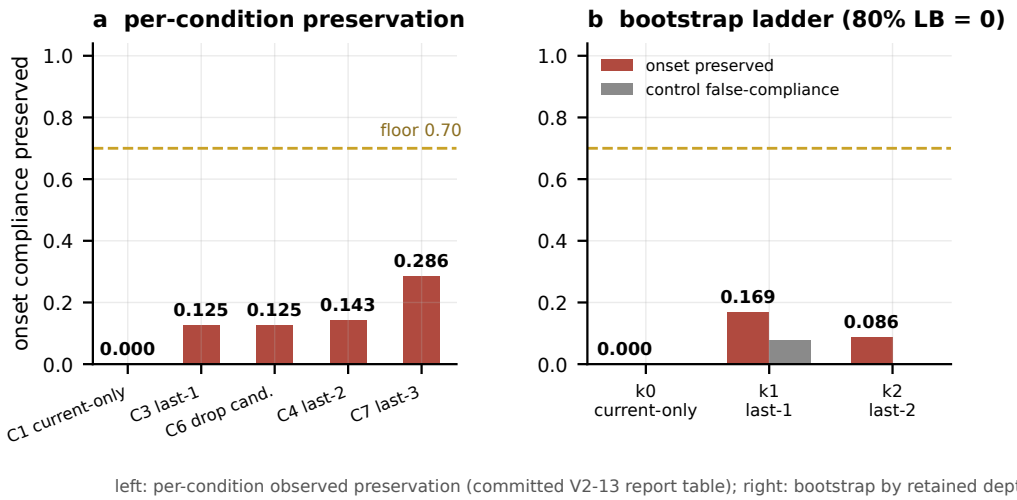

Figure 3: Endpoint context dependence. Preservation of onset compliance under each renderable retained-context policy against the preregistered 0.70 floor (dashed), with matched-control false-compliance for reference. No policy clears the floor; the deepest retained window reaches 0.286. Two committed bases exist for the control rate and we state both: a per-condition maximum of 0.032, and 0.076 under the grouped bootstrap at the last-one-exchange depth ($k = 1$); the current-turn-only depth ($k = 0$) gives 0.000.

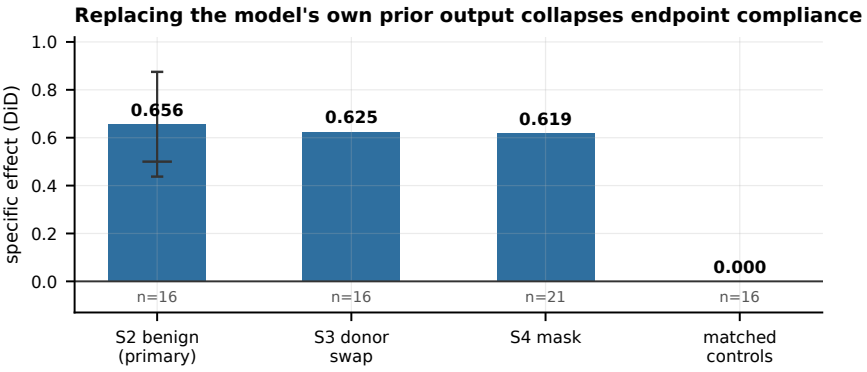

Figure 4: Local assistant-output self-conditioning. Specific effects with 80% intervals for interventions applied to A_j only, with the user turn P_j held fixed. The equal-token benign replacement (S2) is primary; donor swap (S3) and mask (S4) corroborate, and matched non-compliant controls do not flip.

than contradictory: the pre-onset turn is load-bearing *given* the accumulated history, and that history is not replaceable by any short suffix of itself.

Scope. Development split, single target model `google/gemma-2-9b-it`, declared evaluator panel rather than human ground truth, endpoint outcome only. Two planned user-only context conditions were unrenderable under the chat template and were excluded and disclosed, so the context-policy space was **not** exhaustively tested and no minimal-context law is claimed.

9 The model's own prior output sustains compliance

Design. The third question of Figure 1: modify *only* the model's own prior assistant output A_j at $j = t_c - 1$, leaving P_j untouched, then regenerate the endpoint. Where §7 measures the total effect

flowing through A'_j, this isolates the *local* contribution of that output: the part attributable to the model having said what it said rather than to the user having asked what they asked.

Result. The primary condition S2, an equal-token benign assistant replacement, gives a specific effect of $\mathbf{0.656}$ (80% CI lower bound 0.500; 95% CI $[0.4375,\ 0.875]$, excluding 0; 16 source goals; raw onset-loss 0.667). Donor swap S3 $= 0.625$ and mask S4 $= 0.619$ agree in sign and magnitude, which separates *content removal* from filler valence: replacing the content with benign material, with donor material, or with a mask all collapse the outcome similarly, so the effect is not an artifact of inserting agreeable-sounding text. Matched non-compliant controls do not flip under the primary and mask arms (mean control-flip 0.000 for S2 and S4); the donor arm S3 shows a small non-zero control-flip of 0.0625, which its specific effect already subtracts. Of 169 branches, 161 were labelled and 8 were declined under the panel's acceptable-use policy. [EVIDENCE: DEVELOPMENT BEHAVIORAL BRANCH-CAUSAL]

Scope, and a deliberate asymmetry with §7. The positive goals concentrate in one source-category bucket at 16 goals (the bucket whose category is unresolved in the source metadata, at the estimability floor), so **no category-robustness is claimed here**, in contrast to §7 where all five categories are positive. One planned literal-removal condition was template-unrenderable and excluded. The result is a behavioral dependence of the endpoint outcome on prior assistant-slot content, in a selected development onset cohort, under declared-panel labels, on one target model. It is **not** a universal self-conditioning mechanism and **not** evidence of an internal state or memory carrying the effect; §10 reports what happened when we looked for one.

10 NEGATIVE INTERPRETABILITY RESULTS, AND WHAT THEY CLOSE

The behavioral findings invite an internal explanation. We looked for one, did not find it, and report the attempts in the main text, because otherwise a reader would assume the internal account exists and was merely omitted (full detail in Appendix E).

Three attempts at a pre-generation readout. *(1) Mixed.* A compliance-propensity probe on captured pre-generation residual states, under grouped nested cross-validation against a fixed surface baseline, improved goal-balanced average precision ($\Delta\mathrm{AP}$ $+0.077$) but *degraded* ranking ($\Delta\mathrm{AUROC}$ -0.075): one of two co-primary metrics cleared, and the 95% interval includes 0. [EVIDENCE: READOUT MIXED / DIAGNOSTIC] *(2) Weak and exploratory.* A matched analysis of activation excess across onset transitions gave $+0.125$, but the goal-permutation test was not significant ($p = 0.22$), sign consistency was 61%, and leave-one-out was fragile; its preregistration states in advance that it cannot upgrade (1). [EVIDENCE: READOUT MIXED / DIAGNOSTIC] *(3) Confirmatory, negative.* Six residual and representation-engineering families were benchmarked with the confirmatory statistic fixed in advance as a nested meta-model. It did **not** beat the strongest surface baseline: $\Delta\mathrm{AP}$ $+0.017$ (80% CI $[-0.035,\ 0.080]$, below the 0.03 minimum detectable effect) and $\Delta\mathrm{AUROC}$ -0.085, confirmatory residual AP 0.103 versus surface 0.086. The comparator hardened between attempts, so (3) sharpens rather than repeats (1). [EVIDENCE: NOT SUPPORTED]

The null is **selection-limited, not signal-absent**: individual families beat surface on both metrics descriptively (PCA→logistic AP 0.185; mean-difference AP 0.179) but are multiplicity-uncontrolled at 34 positives. A pre-declared *single*-family replication is the test that has not yet been run.

Sparse autoencoders: the instrument did not qualify. Before interpreting any feature we required the autoencoder to reconstruct our own captured states at a frozen bar of FVU ≤ 0.15 and median cosine ≥ 0.95. No layer cleared it at 16K width (best: layer 9, FVU 0.260 / cosine 0.930), and widening eightfold did not fix it (131K, layer 9: 0.245 / 0.943); the ready-layer set is empty at both widths (Appendix E). The measurement is integrity-clean, so this is a real reconstruction number for this model and site rather than a defect in the tool. A near-miss confers no license, so no feature trajectory, cluster, or semantic claim appears anywhere in this work. [EVIDENCE: BLOCKED INSTRUMENT]

Memory and context policy: unresolved. A context-policy branch experiment asked whether storing, truncating, summarizing, retrieving, or compressing the transcript changes downstream

behavior. Execution was sound; the labels were not. Anchor calibration repaired a fivefold over-calling at the *rate* level (agreement $0.46 \rightarrow 0.78$) but not at the *packet* level (combined $\kappa = 0.0229$ against a floor of 0.40): with roughly ten compliant anchors out of 59, two strict panels can agree on how often compliance occurs while disagreeing about which outputs are compliant, and further rubric tuning does not repair that, so no memory-policy result is certified in either direction (Appendix C). [EVIDENCE: UNRESOLVED]

Consequence: patching is not justified. Activation- or feature-level intervention presupposes a demonstrated readout to intervene on. There is **no validated internal readout** in this program and the feature route is blocked, so patching is **not justified** under current evidence and was not run. The standard it would have to meet: a probe or feature that moves only its own score is not behavioral causality; only an intervention that changes the episode outcome under matched controls would support such a claim. [EVIDENCE: NOT SUPPORTED]

11 A BEHAVIORAL SIGNATURE, AND ITS CEILING

Combining only what survived its own controls, the picture is a **development behavioral signature** with three axes: *cumulative full-history dependence* (§8), since no short retained window reproduces onset compliance; *a critical near-onset user turn* (§7), load-bearing given that history and tracking its content rather than its position or mere presence; and *local assistant-output self-conditioning* (§9), contributing on top of the user turn.

Three angles on one cohort. They probe different objects but draw on one shared onset cohort of roughly 22–24 source goals, so they are **not independent replications**: read them as a coherent decomposition within one cohort, not as three studies agreeing. [EVIDENCE: PANEL-OBSERVED BEHAVIORAL ASSOCIATION]

What is deliberately not concluded. **No internal-state mechanism** is established by any instrument tested here; that route is closed at this program stage rather than pending, distinct from the memory-policy case, where the blocker is measurement. Consequently the separation of multi-turn attacks into distributed, trigger, hybrid, self-conditioning, memory-mediated, and context-minimal families is **not** established. What we have is a partial behavioral answer on one cohort and one model; calling it a mechanistic taxonomy would require the internal axis this paper explicitly failed to obtain.

Why the method is the portable part. Equal-token replacement separates content from length, position, and presence; matched-control difference-in-differences separates intervention from run-to-run variation; the sham condition verifies that a branch reproduces its parent exactly when nothing changes. None of that requires internal machinery, so it transfers to other models and corpora.

12 CONCLUSION

Replaying multi-turn jailbreaks canonically (Figure 1) makes it possible to ask which parts of a conversation an outcome depends on, and to answer with controlled replacements rather than correlations. On the development split of one open-weight model, under declared-panel labels, the answer is a coherent behavioral signature: compliance is built cumulatively, so no short context reproduces it; the user turn just before onset is load-bearing given that history; and the model's own prior output helps sustain what follows.

The equally important result is where the evidence stops: no validated internal readout, two sparse-autoencoder widths that did not qualify as instruments, and a memory-policy question unresolved behind a label-calibration limit. We therefore report a behavioral signature with an explicit ceiling rather than a mechanism, and leave every blocked route open with a stated reopening condition (Table 1). The next steps that boundary implies are narrow and pre-declarable: untouched-holdout and architecture-diverse cross-model replications of the branch experiments, expert calibration of the partial-compliance boundary, and a second attack corpus.

AI USE STATEMENT

Generative AI tools were used as writing and engineering aids for this work: code, analysis scaffolding, and manuscript drafting. All scientific decisions (preregistered designs, thresholds, cohort definitions, statistical protocols, and claim wording) were made and verified by the authors against committed artifacts, and every number in this manuscript is machine-checked against the committed source it cites. Language models also appear inside the work itself, in two disclosed roles: as the object of study, and as the declared evaluator panel that produces all outcome labels and is not human ground truth. Appendix G gives the full three-role disclosure and its limits.

ETHICS STATEMENT

This work publishes no attack material: raw prompts, model responses, transcripts, and adjudication payloads are withheld, and released artifacts carry only hashed identifiers, categorical labels, and aggregate scalars. We introduce no new attack technique and report no improvement to attack efficacy: the schedules come from an existing public dataset, and our interventions reduce rather than increase attack success by construction. We establish no validated monitor and no deployable mitigation, so we make no mitigation recommendation. No human subjects were involved and no human annotation was collected: a limitation rather than a convenience, since the calibration study that would upgrade the outcome variable is exactly what is missing.

REPRODUCIBILITY STATEMENT

The full replay, capture, labelling, branch-intervention, and analysis code, the frozen configuration files, the preregistered contracts, and the committed result summaries that every number in this paper cites are included in the supplementary zip (root `README.md` maps the contents). Appendix F states the compute envelope, the pinned model revision, and the deterministic-build policy; raw attack prompts, raw model outputs, activation arrays, and model weights are withheld by design, and every released number is machine-checked against the committed artifact it cites.

REFERENCES

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J. Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

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 37 (NeurIPS 2024)*, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR 2023)*, 2023.

Hongye Cao, Sijia Jing, Yanming Wang, Ziyue Peng, Zhixin Bai, Zhe Cao, Meng Fang, Fan Feng, Boyan Wang, Jiaheng Liu, Tianpei Yang, Jing Huo, Yang Gao, Fanyu Meng, Xi Yang, Chao Deng, and Junlan Feng. SafeDialBench: A fine-grained safety evaluation benchmark for large language models in multi-turn dialogues with diverse jailbreak attacks. arXiv preprint arXiv:2502.11090, 2025.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. arXiv preprint arXiv:2209.10652, 2022.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093, 2024.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv preprint arXiv:2404.15255, 2024.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.825.

Yifan Jiang, Kriti Aggarwal, Tanmay Laud, Kashif Munir, Jay Pujara, and Subhabrata Mukherjee. Red Queen: Exposing latent multi-turn risks in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25554–25591. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1311.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. arXiv preprint arXiv:2502.16681, 2025.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.29.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

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.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

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. PMLR, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.

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

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. arXiv preprint arXiv:2312.06681, 2023.

Salman Rahman, Liwei Jiang, James Shiffer, Genglin Liu, Sheriff Issaka, Md Rizwan Parvez, Hamid Palangi, Kai-Wei Chang, Yejin Choi, and Saadia Gabriel. X-Teaming: Multi-turn jailbreaks and defenses with adaptive multi-agents. arXiv preprint arXiv:2504.13203, 2025.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. arXiv preprint arXiv:2404.16014, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. arXiv preprint arXiv:2407.14435, 2024b.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1207.

Paul Röttger, Hannah 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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.301.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The Crescendo multi-turn LLM jailbreak attack. arXiv preprint arXiv:2404.01833; accepted at USENIX Security 2025, 2024.

Rishi Rajesh Shah, Chen Henry Wu, Shashwat Saxena, Ziqian Zhong, Alexander Robey, and Aditi Raghunathan. Jailbreaking in the haystack. arXiv preprint arXiv:2511.04707, 2025.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548, 2023.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *NeurIPS 2023 Workshop on Attributing Model Behavior at Scale (ATTRIB)*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. arXiv preprint arXiv:2211.00593, 2022.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR 2023)*, 2023.

Zixuan Weng, Xiaolong Jin, Jinyuan Jia, and Xiangyu Zhang. Foot-in-the-door: A multi-turn jailbreak for LLMs. arXiv preprint arXiv:2502.19820, 2025.

Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. Chain of attack: a semantic-driven contextual multi-turn attacker for LLM. arXiv preprint arXiv:2405.05610, 2024.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. arXiv preprint arXiv:2410.02736, 2024.

Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Gao Zuchen, Fei Mi, and Lanqing Hong. CoSafe: Evaluating large language model safety in multi-turn dialogue coreference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17494–17508. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.968.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR 2024)*, 2024.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023.

Andy Zhou and Ron Arel. Tempest: Autonomous multi-turn jailbreaking of large language models with tree search. arXiv preprint arXiv:2503.10619, 2025.

Zhenhong Zhou, Jiuyang Xiang, Haopeng Chen, Quan Liu, Zherui Li, and Sen Su. Speak out of turn: Safety vulnerability of large language models in multi-turn dialogue. arXiv preprint arXiv:2402.17262, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023a.

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.

A REPLAY INVARIANTS

Figure 1 in §4 schematizes the protocol; this appendix lists the invariants the engine enforces. Each is checked at runtime and each is a hard failure rather than a warning, because every one of them, if violated, would silently change what a downstream branch experiment means.

1. **Call parity.** The number of local generation calls equals the number of source user turns in the schedule.
2. **No flattening.** A schedule is never concatenated into a single prompt.
3. **No source substitution.** A source-dataset assistant response is never inserted into the history in place of a locally generated one.
4. **Verbatim carry-forward.** A_t enters turn $t{+}1$ exactly as generated, with no normalization, trimming, or re-tokenization of decoded text.
5. **No adaptive rewriting.** Later user turns are fixed by the schedule and are never modified in response to what the model produced.
6. **No retries, no judge-driven stopping.** A unit runs to its full length; an evaluator never terminates or restarts generation.
7. **Explicit context preflight.** Every turn is checked against the context limit before generation and fails loudly rather than dropping or truncating content.
8. **Deterministic capture.** PRE/GEN/POST states are reconstructed by teacher-forced forwards over the exact generated token ids, at preregistered layers and token positions, never from live decoding caches.
9. **Split integrity.** Splitting is by source goal and precedes all generation; no goal appears in two splits, and validation/test are never read.
10. **Label independence.** Source-dataset intent labels never populate local behavior fields.

Branch validity: the sham condition. Before any scientific branch is run, a sham condition re-runs each branch root while changing nothing. It must reproduce the original states exactly; in the reported runs all 60 branch roots reproduced byte-exactly. A branch experiment whose sham does not reproduce is not analyzed, because a difference measured against a drifting baseline is uninterpretable.

Execution envelope. All reported runs completed within a fixed memory envelope with no spill, no offload, and no truncation, and each closed with the model unloaded and its execution gate re-closed. Execution audits are committed alongside the results.

B DATASET AND COHORT DETAILS

Normalization. Source schedules are normalized so that every row carries: a source-goal id, a schedule id, the ordered user turns, a paired single-turn `plain_query` control for the same goal, a harm category, and raw-text hashes. Raw text itself is never committed; the released artifacts carry hashes and categorical fields only.

Splits. Splitting is by source goal and happens before any generation, into development, validation, and test partitions. No goal crosses a split. Everything reported in this paper is development-split only; validation and test remain untouched.

Replayed cohort. 120 development source goals, each contributing 3 multi-turn attack schedules and 1 paired single-turn control, giving 360 multi-turn units, 120 control units, and 480 units in total, replayed as 1916 generation calls.

Onset cohort for the branch experiments. From the replay we select first-compliance onsets and matched non-compliant control transitions:

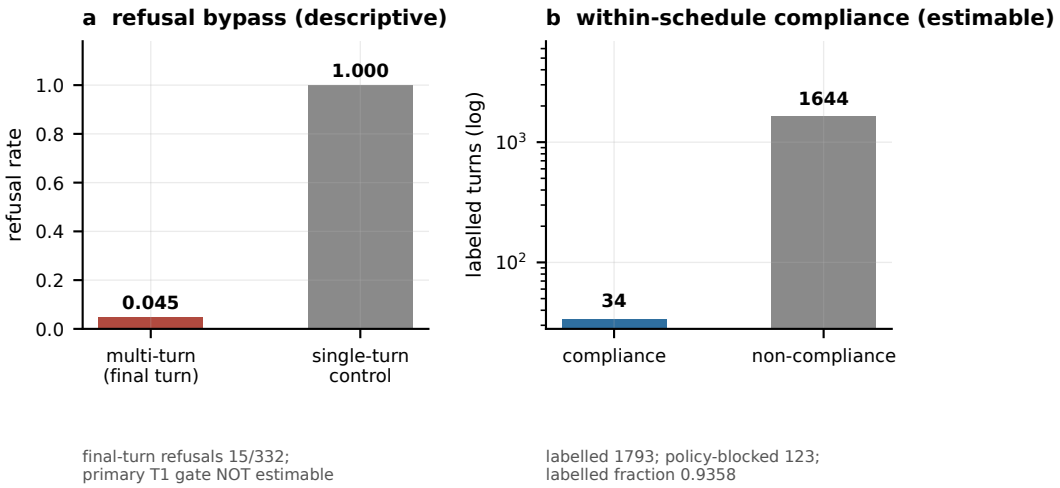

Figure 5: (a) Final-turn refusal under multi-turn attack versus the paired single-turn `plain_query` controls for the same source goals. (b) Within-schedule panel-observed compliance. Rendered from the committed replay summaries; the multi-turn refusal contrast is descriptive, and the preregistered primary refusal gate was not estimable at 15 positive rows.

compliance onsets (t_c)	24
distinct source goals	22
matched non-compliant control transitions	36
branch roots	60

Matching pairs each onset transition with non-compliant transitions from the same source goal wherever available, so that the difference-in-differences contrast of Eq. 2 is within goal rather than across goals. Bootstrap resampling is over source goals, not over branches, so the effective sample size for every headline interval is the number of goals (16–22), not the number of branches.

This is one cohort. §7, §8, and §9 all draw on the cohort above. They differ in *what* they intervene on, not in *whom* they study. Any reading in which they corroborate one another as independent samples is incorrect.

Branch counts. The turn-criticality experiment ran 287 scientific branches across five conditions after a benignness review excluded 13 donor items whose source text carried attack-template residue; the context experiment ran 273; the self-conditioning experiment ran 169, of which 161 were labelled and 8 were declined under the panel's acceptable-use policy. Condition-level goal counts differ (for example 17 goals for the donor-swap arm versus 22 for mask) because a condition is only estimable where its replacement is renderable and its matched control exists; those counts are reported with each effect rather than pooled.

C Labelling and reliability

Protocol. Each local output is adjudicated independently of the source dataset's intent label, under a frozen rubric, into one of `refusal`, `safe_nonrefusal`, `partial_compliance`, `unsafe_compliance`, `ambiguous`. Disagreement resolves to `ambiguous`; it is never forced into a binary and never imputed. Compliance in this paper means `partial_compliance` or `unsafe_compliance`.

Panel composition. Outcomes are produced by a declared evaluator panel: a first-pass rater, a semantic specialist pass on anything not obviously refusal or safe completion, and an adversarial audit pass that may downgrade or mark ambiguous but may never upgrade a negative or ambiguous label. Branch labelling additionally uses blinded packets in which condition identity and the parent outcome are withheld.

Coverage and declines. Of 1916 replay turns, 1793 were labelled and 123 were declined under the panel's acceptable-use policy and recorded as `policy_blocked`. Declines are reported rather than back-filled: a declined item is missing data, and rephrasing the packet to obtain a label would change the measured object.

Reliability.

Check	Agreement	Cohen's κ
Panel continuity vs. frozen prior anchor	0.9322	0.8844
Blind repeat, 10% of the replay	0.967	0.754
Blind repeat, turn-criticality branches	—	0.692
Blind repeat, context branches	—	0.724

The calibration limit that blocked one experiment. Reliability of this kind is base-rate sensitive, and one experiment failed on exactly that. In the context-policy experiment a fresh panel over-called compliance roughly fivefold against the frozen standard on byte-identical outputs (agreement ≈ 0.46). Anchor calibration repaired the *rate* (agreement 0.78; fresh rate 0.085 versus frozen 0.169) but not the *packet-level* chance-corrected agreement: combined $\kappa = 0.0229$ against a preregistered floor of 0.40, with roughly ten compliant anchors out of 59. Two strict panels agreeing on how often compliance occurs while disagreeing on which outputs are compliant is a commensurability failure at a sparse base rate, and it is not repaired by further rubric tuning.

Consequence for every result in this paper. Agreement is measured, reported, and in one case decisive enough to block an experiment. It is still not correctness. No human adjudication was performed, so **no result here rests on human ground truth**, and upgrading the outcome variable requires a calibration study that this program has not run.

D STATISTICAL PROTOCOLS

Grouping. Source goal is the unit of analysis everywhere. No goal crosses a fold, a split, or a bootstrap resample. Because a goal contributes several schedules and a schedule several turns, treating branches or turns as independent would overstate precision by roughly the average cluster size; every headline interval here is a source-goal cluster bootstrap.

Intervals. Each experiment preregisters the 80% interval as its decision interval and reports the 95% interval alongside it. We report both throughout rather than choosing the flattering one: §7 and §9 clear zero at 95%, whereas the readout diagnostics in §10 do not, and that difference is the point.

Specific effects. Branch effects are difference-in-differences quantities (Eq. 2): the rate at which an intervention removes compliance at an onset transition, minus the rate at which the same intervention introduces compliance at a matched non-compliant control transition from the same goals. The control term is what separates "the intervention did something" from "re-running the model does something".

Permutation tests. Significance is assessed by permuting the onset/control assignment within source goal, preserving cluster structure. Reported p-values are goal-permuted, not branch-permuted.

Equal-token construction. The primary replacement conditions hold token count, role, and turn position fixed. This is deliberate: a shorter replacement would confound content removal with context-length reduction, which §8 shows is itself a strong effect.

Probe evaluation. Readout benchmarks use grouped nested cross-validation with all model and hyperparameter selection performed inside inner training folds. No outer-fold metric selects a model. The confirmatory statistic is fixed in advance (for the six-family benchmark, a nested meta-model), so that per-family descriptive wins cannot be promoted after the fact. Minimum detectable effects, preservation floors, readiness bars, and estimability floors are all set before metrics are computed and are never retuned afterwards.

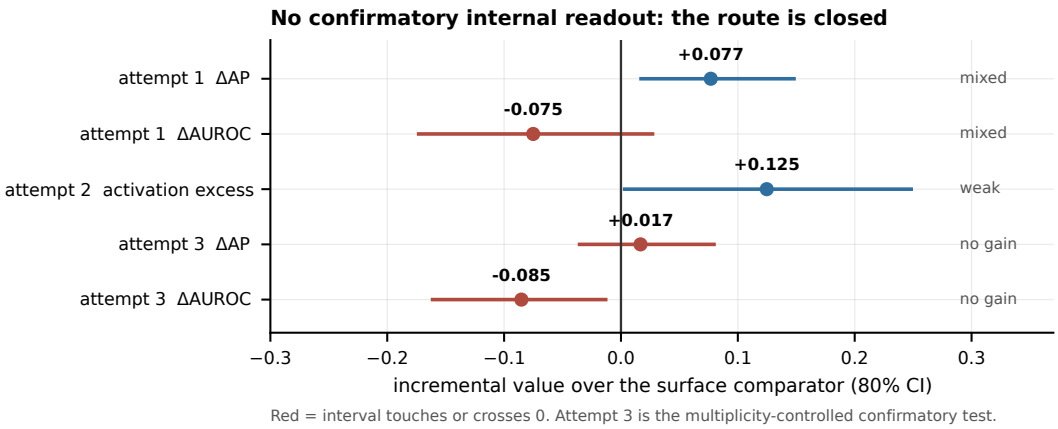

Figure 6: The three internal-readout attempts. Incremental value over the surface baseline with 80% intervals: mixed (one of two co-primary metrics), weak and exploratory (permutation not significant), and a multiplicity-controlled confirmatory null.

Estimability floors. A gate that falls below its preregistered minimum positive count is reported as *not estimable* rather than estimated with a wide interval. This is why the refusal-bypass contrast of §6, despite being large, is presented as descriptive: 15 positive rows fall below the floor fixed in advance.

Robustness checks applied to the positive results. Each branch effect was additionally checked for missing-not-at-random label declines, leave-one-goal-out stability, thin-matching sensitivity (goals with a single matched control versus several), and category concentration. §7 is positive in all 5 estimable categories; §9 is not, and its category concentration is reported as a limitation rather than smoothed over.

The robustness check we did not run. None of the above perturbs the *labels*. Every robustness result here is conditional on one panel's adjudication of the outcome variable, and §10 demonstrates on this same program that a second strict panel can disagree about which outputs are compliant at a sparse base rate. A label-perturbation analysis (re-estimating each effect under an independently adjudicated relabelling of the branch endpoints) is therefore the robustness check with the most power to change these conclusions, and it is not part of this work. It is the analytic half of the expert-calibration study named in §12.

E Negative and blocked routes in full

Readout attempt 1 (mixed). Pre-generation compliance-propensity probe on captured residual states at the last user-token position, grouped nested cross-validation, against a fixed surface baseline. ΔAP $+0.077$ (80% CI $[0.017, 0.148]$), ΔAUROC -0.075. Prevalence $\approx 2\%$; 34 positive rows across 25 goals. One co-primary metric cleared; 95% includes 0. Terminal: mixed, non-confirmatory.

Readout attempt 2 (weak, exploratory). Fold-normalized matched transition analysis; activation excess for onset transitions versus matched non-onset transitions, $+0.125$ (80% CI $[0.0032, 0.248]$), permutation $p = 0.22$, sign consistency 61%, leave-one-out surviving in a minority of folds and goals, and concentrated in one category. Its preregistration states in advance that it cannot upgrade attempt 1.

Readout attempt 3 (confirmatory, negative). Six families benchmarked with the confirmatory statistic pre-fixed as a nested meta-model over families: ΔAP $+0.017$ (80% CI $[-0.035, 0.080]$; preregistered MDE 0.03), ΔAUROC -0.085; best confirmatory residual AP 0.103 versus strongest surface 0.086. Per-family descriptive results (PCA$\rightarrow$logistic AP 0.185, mean-difference AP 0.179) exceed surface but are multiplicity-uncontrolled at 34 positives. The sparse-autoencoder feature family in this benchmark could not be run, because the instrument had not qualified.

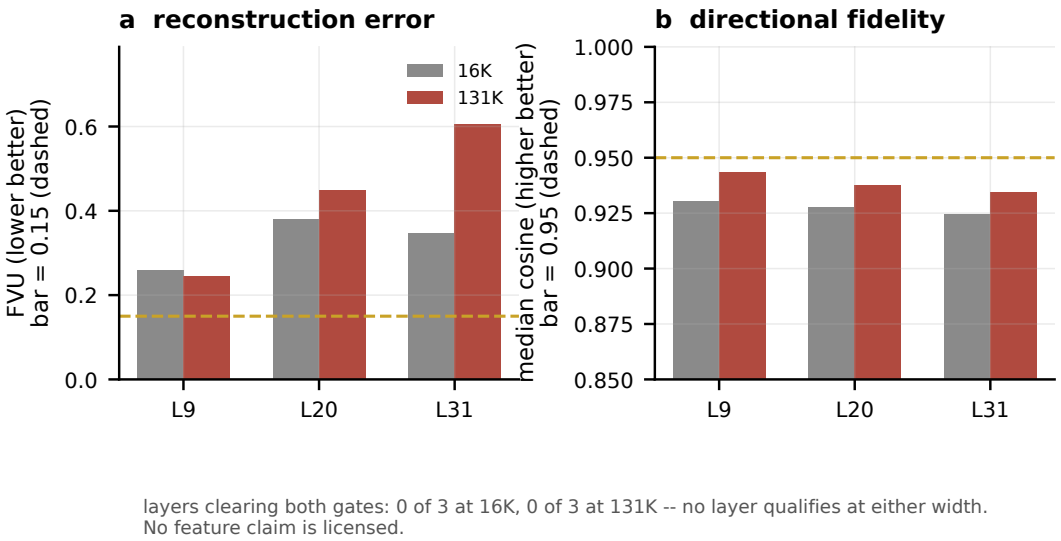

Figure 7: Sparse-autoencoder instrument qualification at 16K and 131K dictionary widths against the frozen readiness bar (FVU ≤ 0.15, median cosine ≥ 0.95). No layer clears the bar at either width; the eightfold widening improves cosine slightly and does not repair fraction of variance unexplained. The instrument did not qualify; it is not defective.

Three further committed details from that benchmark belong here rather than in a footnote, because each of them bounds how the null should be read. The confirmatory result is *not* category- robust (dropping the heaviest category moves ΔAP to -0.0436) and *not* clear of a missing-not-at-random concern from label declines: both preregistered robustness sub-gates fail, which weakens any attempt to read the null as a strong negative. Conversely, a *secondary* capture position at the assistant boundary did show incremental value (ΔAP $+0.0737$ with the 80% interval excluding 0). That position was not the confirmatory endpoint, was not multiplicity-controlled, and cannot upgrade the verdict; we report it because it is the most concrete lead a follow-up would pre-declare, and suppressing it would misrepresent the null as more complete than it is.

SAE qualification. Exact instruction-tuned autoencoders matched to the target model and site, scored on the same captured residuals used everywhere else:

	16K width		131K width	
Layer	FVU	cosine	FVU	cosine
9	0.260	0.930	0.245	0.943
20	0.380	0.928	0.449	0.938
31	0.347	0.925	0.607	0.935
bar	≤ 0.15	≥ 0.95	≤ 0.15	≥ 0.95

The ready-layer set is empty at both widths. Reconstruction was computed in double precision and verified against a single-precision run, and hooks were verified against the capture sites, so the moderate fidelity is a property of the model/site/instrument combination rather than an artifact.

Memory / context policy. Execution was clean; labelling was not. See Appendix C. No policy-level rate is certified in either direction, and no memory-mediated account appears anywhere in this paper.

Patching. Not justified and not run. The precondition (a demonstrated pre-generation readout to intervene on) was not met by any of the three attempts, and the feature route is blocked by the instrument result. A patch that moves a probe score is not a behavioral result; the bar is a change in the episode outcome under matched controls.

Table 1: Route closure. Every closed or unresolved route carries the condition that would reopen it; none is described as pending or in progress.

Route	Status	Condition to reopen
Internal unsafe-state readout	CLOSED	a probe beating the strongest surface on both metrics under multiplicity control
SAE feature trajectories / clusters	BLOCKED	a qualifying readiness verdict for a different SAE family or site
Activation / feature patching	NOT JUSTIFIED	a demonstrated readout, or a pre-declared single-family replication clearing its gates
Memory / context policy	UNRESOLVED	a denser compliant cohort, or human adjudication of the sparse anchor boundary
Cross-model replication	DEFERRED	explicit model-execution authorization plus a frozen replication protocol

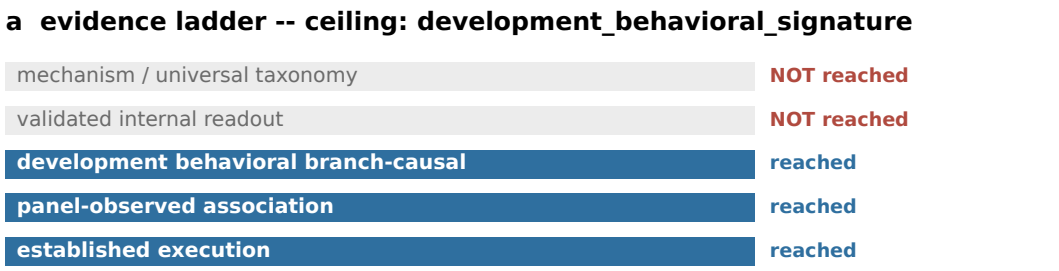

Figure 8: Evidence ladder across the program. Positive results reach the controlled behavioral level on one shared development cohort; the internal-readout, feature, memory-policy, and patching routes are closed, blocked, unresolved, and not justified respectively, each with a stated reopening condition.

F Reproducibility

What is reproducible from the released artifacts. Every number, figure, and table in this paper is derived from committed experiment summaries. The paper build loads no model, runs no GPU, generates no text, captures no activations, and relabels nothing. Figures are rendered deterministically from those committed summaries, and each figure and table carries a manifest row naming its source file. Numbers are transcribed from committed artifacts and the headline values are additionally registered with their source path, key, target location, claim id, and evidence level, each machine-verified to occur in the artifact it cites. A value absent from the committed summaries is marked unavailable rather than inferred or recomputed.

Preregistration and freezing. Each experiment was preregistered before its metrics existed, with its own contract, cohort, and audit set. Contracts are content-hashed with line endings normalized, so the hashes are checkout-invariant, and the per-experiment source manifest additionally pins the raw hash of every artifact the paper cites: a changed upstream summary is detected rather than silently

absorbed. Thresholds (minimum detectable effects, preservation floors, readiness bars, estimability floors) were fixed before metrics and never retuned.

Statistical conventions. Grouping by source goal; source-goal cluster bootstrap; 80% intervals primary and 95% secondary; difference-in-differences specific effects against matched non-compliant controls; goal-permuted significance; grouped nested cross-validation with selection confined to inner training folds.

Build. The manuscript compiles from the repository checkout plus the conference style package, with an external bibliography and one file per section; the figure sources, table sources, claim traceability, numeric traceability, forbidden-claim scan, and anonymity check are produced by committed scripts and are re-runnable offline. No cloud service or proprietary tool is required.

What is deliberately *not* reproducible here. Regenerating the underlying generations, activations, or labels requires the target model, a GPU, and a fresh execution authorization, and is outside this package by design. Raw attack prompts, raw model responses, full transcripts, per-turn adjudication payloads, and activation arrays are withheld (§H); released artifacts carry hashed identifiers, categorical labels, and aggregate scalars only. The consequence is honest and worth stating: an external party can re-derive every number in this paper from the released summaries, and can re-run the protocol from its specification, but cannot re-derive the summaries themselves without model access.

G Disclosure of large language model usage

We disclose LLM involvement in three distinct roles, because they carry very different implications for the evidence.

1. As the object of study. The target model `google/gemma-2-9b-it` is the system under test. Its outputs are the measured quantity, generated locally under the frozen replay protocol of §4.

2. As the measurement instrument (the material limitation). All outcome labels are produced by a declared evaluator panel of language models under a frozen rubric, with an adversarial audit pass that may downgrade but never upgrade a label. **These labels are not human ground truth.** This is disclosed in the abstract, in §5, in every result section, and in §H, because it bounds every claim in the paper: agreement between panels is measured (Appendix C) and was in one case decisive enough to block an experiment, but agreement is not correctness. Upgrading this instrument (expert or human calibration of the partial-compliance boundary) is the single change that would most raise the evidence level of these results.

3. As a writing and engineering aid. Language models were used to assist with code, analysis scaffolding, and manuscript drafting. All scientific decisions (preregistered designs, thresholds, cohort definitions, statistical protocols, claim wording, and the terminal status of every experiment) were made and verified by the authors against committed artifacts, and every number in this manuscript is machine-checked against the committed source it cites. No result, citation, or number in this paper was produced by generation without such verification.

What LLMs were not used for. No language model adjudicated the paper's claims, selected a statistical result, chose a threshold after seeing metrics, or supplied a bibliography entry that was not subsequently verified against a primary source page.

H Limitations

These bound the claims rather than decorate them. **Provenance:** all outcome labels are declared evaluator panel adjudications and are **not human ground truth**; agreement is measured (Appendix C) but agreement is not correctness. **Split:** development split only; validation and test remain untouched and no generalization to them is claimed or estimated. **Model:** one target model, `google/gemma-2-9b-it`, at one pinned revision; nothing here is cross-model. **Shared cohort:**

§7, §8, and §9 use one related onset/control cohort and are not independent replications. **Power:** effects rest on 16–22 source goals and §9's positives concentrate in a single category bucket, so intervals are wide and reported as such. **Behavioral only:** every positive causal statement is behavioral; the internal axis was not scored or was deferred in all three branch experiments. **Coverage:** 123 of 1916 replay turns and 8 of 169 self-conditioning branches were declined under the panel's acceptable-use policy, and three planned conditions were unrenderable under the chat template. **Estimability:** the primary final-turn refusal gate was not estimable, so the refusal-bypass contrast is descriptive. **Single corpus:** a second attack method or dataset would test whether the signature belongs to the setting or to this corpus. **Label sensitivity, the sharpest form of the provenance limitation:** the panel that adjudicates outcomes also *selected* the onset cohort these effects are measured on, and §10 shows a second strict panel failing to reproduce *which* outputs are compliant at this base rate ($\kappa = 0.0229$); we did not run a label-perturbation robustness check, so we cannot exclude that the effect sizes are partly a property of one panel's partial-compliance boundary.