# PROBING AND STEERING CONSTRAINT FOLLOWING IN MULTI-TURN DIALOGUE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Why do language models stop satisfying an initial constraint as a dialogue grows? We construct a multi-turn evaluation set from IFEval-inspired instructions and Dolly questions: 23 constraint types, 230 sessions and 20 turns per session. Outcomes use repository-specific checks, not official IFEval scoring; some checks approximate rather than fully validate the requested constraint. Across Qwen2.5 3B, 7B, and 14B and Llama3 8B, we extract 18,400 pre-response states, excluding each current answer. Session-disjoint linear probes identify the constraint on 48.3–72.2% of subsequently checker-failed turns; all four family-adjusted intervals exceed the 4.35% uniform-label reference. This establishes decodability, not functional use. We compare concept activation vector (CAV) and instruction-directed attention steering in 108 common-cap cells totaling 496,800 response-condition evaluations. At unit strength, combined steering improves checker pass rates by 21.7–29.8 percentage points over baseline, with subadditive gains. Across all 36 model–strength combinations in the complete joint grid, adding CAV to attention both gains and loses successes; losses exceed gains in 24. Modest CAV steering adds value to weak attention, but no tested addition to stronger attention has a positive pointwise 95% interval excluding zero. Individual rescues therefore do not guarantee net joint benefit. The findings concern this exploratory evaluation protocol, not a demonstrated natural failure mechanism or semantic answer quality.

## 1 INTRODUCTION

Consider an illustrative dialogue: a large language model (LLM) is told to *always respond in bullet points*. It complies at turn 2, turn 5, perhaps turn 8, then quietly stops. Multi-turn evaluations document unstable instruction adherence (Qin et al., 2024), but a behavioral failure alone cannot tell us whether the instruction is no longer represented or no longer controls the output.

Testing that distinction requires a fixed target, turn-level outcomes and aligned representation measurements. IFEval supplies verifiable constraints (Zhou et al., 2023), while FollowBench varies constraint complexity (Jiang et al., 2024); neither directly supplies this persistence experiment. We adapt 23 IFEval-inspired types into 230 sessions of 20 Dolly-derived questions, keeping the initial instruction in context without repeating it. The contribution is a controlled adaptation and evaluation protocol, not a new source corpus or the first multi-turn benchmark. Our custom checks are imperfect proxies: throughout, compliance, success and failure refer to their pass/fail labels, not independently verified instruction satisfaction (§4.1).

The instruction never leaves the context window, but its presence need not control the output. To test whether its identity is encoded before failure, we extract residual states from saved dialogues ending immediately before each answer. A session-disjoint 23-class linear probe recovers the constraint on 48.3–72.2% of subsequently non-compliant turns (Table 1); accuracy remains 40.2–69.2% in sessions that never comply. This **decodability–compliance gap** precedes the current answer. It establishes recoverable information at that boundary, not that the model functionally accesses or uses it.

We then ask whether targeting a constraint direction or instruction attention improves compliance, and whether combining them adds value. Retrospective **constraint projection drift** and pre-response **attention dilution** motivate the two controls. Both improve compliance at unit strength, but their joint gain is subadditive. Across the full strength grid, additional rescues coexist with lost successes.



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.

We separate this outcome accounting from two possible sources of interaction: coupling within the current computation and divergence of generated histories. Neither mechanism is established.

**Contributions.** (1) A controlled multi-turn evaluation set and protocol for tracking one fixed constraint target across a growing dialogue. (2) Response-excluded, session-disjoint evidence that constraint identity is decodable before non-compliant answers. (3) An exploratory 108-cell intervention study distinguishing individual efficacy, joint subadditivity and marginal benefit, with full-grid gain–loss accounting and bounded tests of interaction hypotheses.

## 2 Related Work

**Instruction-following evaluations.** IFEval supplies verifiable instructions (Zhou et al., 2023); FollowBench varies constraint complexity (Jiang et al., 2024). Multi-turn benchmarks address system-message adherence (SysBench; Qin et al., 2024), structured instruction flows (StructFlowBench; Li et al., 2025), evolving instruction challenges (EvolIF; Jia et al., 2025) and realistic constraints (SEQUOR; Canaverde et al., 2026). Multi-turn evaluation is therefore not itself our contribution. Our controlled adaptation holds one initial target fixed and aligns turn-level checks with response-excluded states and policy interventions. Its narrower scope enables these comparisons, but its custom checks and persistence wording limit validity. Multi-turn training (Sun et al., 2024), long-context alignment (Bai et al., 2024) and retrieval degradation (Liu et al., 2024) address related questions, not whether a failed constraint remains decodable before the answer.

**Probing and representation analysis.** Linear probes (Alain & Bengio, 2017) reveal information encoded in representations (Conneau et al., 2018; Belinkov, 2022; Hewitt & Manning, 2019; Tenney et al., 2019; Gurnee et al., 2023), and the linear representation hypothesis posits concepts as directions (Park et al., 2024; Marks & Tegmark, 2024; Burns et al., 2023). Heo et al. (2025) find single-turn instruction-following success linearly decodable. We ask instead what happens to the constraint representation across twenty turns once compliance has collapsed, using probing as dynamic trajectory measurement.

**Steering and attention intervention.** Concept activation vectors (Kim et al., 2018), representation engineering (Zou et al., 2023), activation addition (Turner et al., 2023; Rimsky et al., 2024), inference-time intervention (Li et al., 2023), function vectors (Todd et al., 2024) and causal abstraction (Geiger et al., 2024) intervene on internal representations; Stolfo et al. (2025) use instruction vectors for single-turn adherence. A subspace intervention can be effective without identifying the mechanism the model ordinarily uses (Makelov et al., 2024), so we compare CAV steering with random and wrong-constraint directions. Our attention arm is closest to PASTA (Zhang et al., 2024), which reweights attention toward user-designated spans. Our contribution is not either intervention but their joint behavior across turns: interaction, rescue overlap, dose, and history.

## 3 Theoretical Motivation

Under causal masking, instruction-position states cannot change when later turns are appended to an unchanged prefix. The generation-position residual, however, combines attention to instruction and non-instruction tokens (Elhage et al., 2021). We distinguish two motivations for intervention, not two established failure mechanisms (derivations in Appendix A).

**Representation drift.** Growing non-instruction context could change readout alignment (Elhage et al., 2022; nostalgebraist, 2020; Dar et al., 2023). Our historical diagnostic measures post-response orientation:

$$\mathrm{drift}(t) = \cos(\mu_t, \mu_1), \tag{1}$$

where $\mu_t$ is the unit-normalized per-constraint mean cached residual, distinct from the contrastive CAV used for steering. Neither this cosine nor decodability measures causal influence.

**Attention dilution.** Instruction attention is $A_{\mathcal{I}} = \sum_{i \in \mathcal{I}} e^{s_i} / \sum_j e^{s_j}$. More non-instruction keys can reduce this share, but changing queries and logits can counteract dilution. A bounded-relative-

logit assumption gives $A_{\mathcal{I}} = O(|\mathcal{I}|/((t-1)n))$ (Appendix A). This motivates a control, not a verified decay law.

# 4 Dataset and Experimental Setup

## 4.1 A Controlled Multi-Turn Evaluation Set

**Design requirements.** We need a fixed evaluation target, repeated content questions and reproducible turn-level outcomes.

**Construction.** We adapt 23 IFEval types (Zhou et al., 2023), excluding `repeat_prompt` and `response_language`, spanning formatting, keywords, length, case, punctuation, content markers and structure. Ten variants per type yield 230 sessions. A seeded sampler selects 20 distinct questions per session from a 119-question Dolly-derived pool (Conover et al., 2023). Turn 1 contains the instruction and a question; later turns contain questions only. Prior answers and the initial instruction stay in context. The protocol applies the initial target to every turn, but the prompts have no uniform all-turn persistence wrapper: some say *the response* rather than *every response*. Thus late failure can reflect scope ambiguity as well as persistence (Appendix B).

**Outcome validity.** The frozen repository checker returns a binary field named `strict`; it is *not* the official IFEval strict metric. Some rules only approximate the prompt: `constrained_response` checks at most 50 whitespace words, not membership in answer choices; JSON parsing permits any JSON value or a fenced block; several structural rules only check marker presence. Table 4 specifies all 23 rules. These limitations affect every aggregate and failed-turn subgroup; we retain the recorded outcomes without rescoring and make no semantic-quality claim.

## 4.2 Models and Measurements

**Phase 1** runs the behavioral sessions; **Phase 2** caches activations and measures probes, CAV and attention trajectories; **Phase 3** runs the interventions of §6 with freshly generated baselines. The response-excluded verification reuses saved Phase 1 dialogues, not newly generated answers.

**Models and decoding.** We use instruction-tuned Qwen2.5 3B, 7B and 14B (Qwen Team, 2024) (36/28/48 layers, $d$=2048/3584/5120) and Llama 3 8B (Grattafiori et al., 2024) (32 layers, $d$=4096), all with grouped-query attention. Decoding is greedy (`do_sample=False`) with eager attention and batch size 1, in float16 for Qwen2.5 3B and bfloat16 for the other three models (the registered configuration, in all phases). Phase 1 used response caps of 256 tokens (Qwen) and 200 tokens (Llama); these trajectories are used for the probe and diagnostic analyses of §5 and for the Phase 2 steering calibration (session targets, head ranking and attention-decay factors), so Llama's steering is calibrated on its 200-token histories. All Phase 3 cells, including each model's baseline, use one common 256-token cap and seed 42. Equal caps do not equalize generated lengths or truncation rates, which we report.

**Pre-response linear probe.** For each saved Phase 1 dialogue and turn $t$, a new forward pass includes only the first $2t-1$ alternating messages, ending at the user turn, and appends the original chat template's generation prompt. We read the final prompt token's residual at a fixed decoder-block output (zero-indexed layers 21/14/25/30 in Qwen 3B/7B/14B and Llama). All 18,400 states are complete; no new answers are generated. Five session-disjoint folds fit 256-component principal component analysis (PCA) and logistic regression ($C$=0.01, seed 42) on training sessions only, using turns 2–20 (4,370 rows/model); neither layer nor hyperparameters are searched. Primary failed-turn intervals use 2,000 whole-session bootstrap draws and a family-of-four adjustment, conditional on fitted out-of-fold predictions. Appendix W gives input hashes, checks and secondary populations. Historical per-layer probes and geometry include the current answer and remain separately labeled post-response diagnostics (Appendix O).

**CAV extraction.** For each constraint type we form 200 paired prompts (10 variants $\times$ 20 fillers), contrasting **with** (instruction + filler) against **without** (Wikipedia neutral text + filler) (Kim et al., 2018): $\mathrm{CAV}^{(\ell)} = \mathbf{d}^{(\ell)}/\|\mathbf{d}^{(\ell)}\|$ with $\mathbf{d}^{(\ell)} = \frac{1}{N}\sum_i(\mathbf{h}^{(\ell)}_{\mathrm{with},i} - \mathbf{h}^{(\ell)}_{\mathrm{without},i})$ (Appendix C).

| Metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| All turns (accuracy, %) | 61.53 | 76.41 | 72.33 | 56.86 |
| Failed turns (accuracy, %) | 58.68 | 72.22 | 67.60 | 48.32 |
| Family-of-4 interval | [53.4, 63.8] | [66.6, 77.9] | [62.6, 73.1] | [41.5, 55.1] |
| Failed turns ($n$) | 3550 | 3099 | 3346 | 2800 |
| Training-majority on failures (%) | 2.85 | 3.39 | 2.99 | 0.71 |

Table 1: Constraint identity decoded *before* the current answer, on saved Phase 1 histories (230 sessions/model; Qwen cap 256, Llama cap 200). Failed turns are selected by the subsequent saved answer. Intervals are Bonferroni-adjusted for four primary accuracies using 2,000 session-cluster draws. The uniform-label reference is 4.35%; the training-majority baseline uses training labels only.

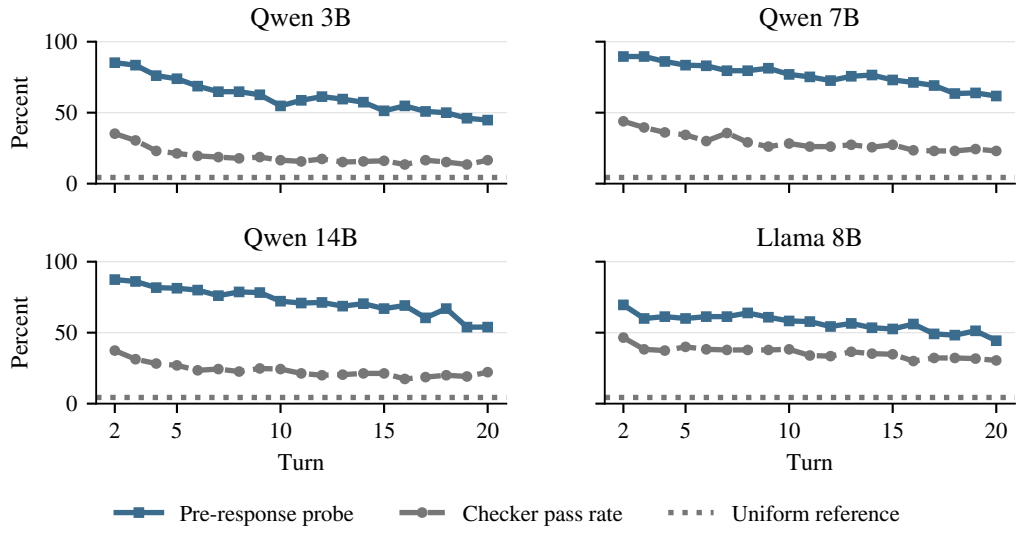


Figure 1: Current answers excluded: pre-response probe accuracy (blue) and saved-answer checker pass rate (gray) on the same held-out sessions. Lines are per-turn descriptive means, not paired inferential contrasts. The dotted line is the uniform-label reference.

## 5 Decodability Before Compliance Failure

**Behavioral landscape.** All models show steep decay: Qwen 2.5 7B drops from 66.5% compliance at turn 1 to 23.0% at turn 20. Pre-response probe accuracy at turn 20 remains 44.3–61.7% across models (Figure 1). Historical behavioral and post-response probe results remain in Table 41.

**Information is present before failed answers.** Pre-response accuracy on failed turns is 58.68/72.22/67.60/48.32% (Qwen 3B/7B/14B, Llama); every family-adjusted lower bound exceeds the uniform-label reference. Accuracy is lower on late failures (T10–20: 52.04/67.58/61.74/45.36%), but the signal persists. Failures with no earlier success yield 45.20/70.23/62.76/42.41%; never-compliant sessions yield 42.19/69.21/61.43/40.24% (secondary pointwise intervals and denominators: Appendix W).

**Scope.** Excluding the current answer addresses its direct leakage into the readout. The original instruction remains visible, and prior answers can carry constraint-related cues even when they fail the recorded checker. Decodability therefore does not show memory without the instruction, functional use, or mediation. Historical transcript and nonlinear-probe controls concern a different, post-response measurement (Appendices R and Q); they are not controls on the new pre-response states.

**Historical diagnostics motivate, but do not identify, intervention targets.** Post-response CAV projections correlate with outcomes in Qwen, but not clearly in Llama; pre-response instruction attention correlates in all four models (Appendix N). These mixed-timing associations do not identify independent failure pathways. Lagged, session-disjoint prediction adds little beyond turn, constraint, length and previous failure; no universal cosine threshold is established (Appendix O). Forecasting ordinary failures is distinct from explaining losses caused by combining policies.

# 6 Intervention Experiments

We test whether changing the constraint projection or instruction attention changes compliance. These interventions establish effects of the specified control policies; a positive gain does not establish that the patched quantity is the natural cause of decay.

## 6.1 Interventions and Policy Comparisons

Every condition starts from the same 230 initial sessions and generates its own 20-turn conversation, with steering at every turn. The estimand is a complete-policy difference; from turn 2 on, paired arms have different histories. Each model's reported 27 conditions comprise baseline; CAV and attention separately at $\{0.5, 1, 1.5, 2, 3\}$; all nine combined cells in $\{1, 2, 3\}^2$ plus $(0.5, 1), (1.5, 1), (1, 0.5), (1, 1.5)$; random and wrong-constraint directions at 1; and a reference policy at $(1, 1)$. These 108 cells contain 496,800 response–condition records. The final extension adds six joint cells per model to the original displayed matrix; provenance and the post-hoc presentation restriction are in Appendix U. Qwen's seven CAV layers and 192 heads were selected by strength-1 compliance sweeps on this bank under an earlier 512-token protocol. Llama's layers 8–14 are exploratory, with counts borrowed from Qwen7B (Appendix F). All sites remain frozen.

**CAV steering.** At each selected layer and decoding step, a unit contrastive direction $c$ and the session's Phase 2 target $p_1$ define $T_{\lambda_C}(h) = h + \lambda_C(p_1 - c^\top h)c$ on the last-token residual. The correction is state-adaptive. The historical post-response target is transferred to an active decoding hook, not measured at the same time boundary; "restoration" denotes this control rule, not restoration of a demonstrated pre-failure state.

**Attention steering.** For the selected query heads, a bias is added to instruction-span keys, $\text{score}'(q, k_i) = \text{score}(q, k_i) + \lambda_A \delta_s$ for $i \in \mathcal{I}$, inside the native eager attention after the causal mask and before softmax. The session-specific $\delta_s$ is calibrated once from the Phase 2 turn-2/turn-10 attention ratio and not recalibrated, so it does not guarantee exact attention restoration. CAV and attention strengths have different units.

**Controls.** Random and wrong-constraint controls retain the target-CAV coefficient $\lambda_C(p_1 - c^\top h)$ but inject a random unit vector or another constraint's CAV (fixed cross-family assignment). A *reference* policy tests the coefficient's dependence on attention-perturbed states (§6.4); it remains prefix-dependent.

**Uncertainty.** Intervention intervals use 10,000 paired whole-session bootstrap draws (seed 42), resampling 230 sessions jointly across arms. We report pointwise 95% intervals, reference-family-16 bounds and approximate grid-family-108 bounds (three contrasts per joint cell/model). Extreme adjusted tails have few bootstrap replicates. The bank and site choices were previously examined; all inference is exploratory, with no selected-winner inference or equivalence claim.

## 6.2 Do the Individual Interventions Improve Compliance?

**Both targets improve recorded outcomes.** At unit strength, all three policies improve checker pass rates, with pointwise intervals above zero (Table 2). Across models, CAV gains 13.17–22.24 pp, attention 13.37–19.37 pp, and combined steering 21.70–29.76 pp. Rescues accompany harm on baseline successes (Appendix I). Appendix Figure 4 shows all 23 types: gains are heterogeneous, not universal.

| Condition / contrast | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Baseline | 21.54 | 31.07 | 25.37 | 37.24 |
| CAV(1) | 43.78 | 46.93 | 42.65 | 50.41 |
| Attention(1) | 40.91 | 46.20 | 38.74 | 53.15 |
| Combined(1,1) | 51.30 | 54.96 | 50.13 | 58.93 |
| CAV − Baseline | +22.24 [+18.0, +26.6] | +15.87 [+12.0, +19.9] | +17.28 [+13.4, +21.4] | +13.17 [+10.1, +16.4] |
| Attention − Baseline | +19.37 [+15.8, +23.1] | +15.13 [+12.0, +18.4] | +13.37 [+10.8, +16.2] | +15.91 [+12.5, +19.4] |
| Combined − Baseline | +29.76 [+25.2, +34.5] | +23.89 [+19.7, +28.2] | +24.76 [+20.5, +29.2] | +21.70 [+17.8, +25.6] |
| Interaction $I$ | −11.85 [−16.1, −8.0] | −7.11 [−10.5, −3.6] | −5.89 [−9.1, −2.7] | −7.39 [−11.3, −3.5] |

Table 2: Core cap-256 results at strength 1. First four rows: checker pass rate (%); contrasts: pp with paired whole-session 95% intervals (10,000 draws, seed 42), pointwise and unadjusted. Interaction $I$ = Combined − CAV − Attention + Baseline. Direction controls: Table 3; complete integer grid: Appendix I.

| Contrast | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| CAV(1) − Random(1) | +22.35 [+18.1, +26.7] | +15.24 [+11.4, +19.2] | +16.93 [+13.0, +21.0] | +12.24 [+9.0, +15.5] |
| CAV(1) − Wrong(1) | +5.37 [+2.0, +9.0] | +7.09 [+3.8, +10.4] | +7.85 [+4.5, +11.3] | +2.78 [−1.0, +6.6] |

Table 3: Directional efficacy at strength 1: compliance differences (pp) with existing, unadjusted paired whole-session 95% intervals. CAV exceeds random in all four models; the wrong-direction contrast remains uncertain in Llama. This does not identify a unique constraint-specific causal pathway.

**Direction identity matters, but not exclusively.** Random directions change compliance by −0.11/+0.63/+0.35/+0.93 pp in the same model order, and CAV exceeds random in every model. Wrong-constraint CAVs retain much of the gain (+16.87/+8.78/+9.43/+10.39 pp): CAV minus wrong is +5.37 to +7.85 pp in Qwen with intervals above zero, but +2.78 pp [−1.00, +6.61] in Llama. We therefore claim relative directional efficacy, not constraint-specific causation (Appendix J).

### 6.3 When Does Combining Interventions Add Value?

**Subadditivity is not absence of benefit.** Let $B, C, A, J$ denote pass rates for baseline, CAV, attention and their combination at the corresponding strengths. The interaction $I = J - C - A + B$ asks whether improvements add; $\Delta_{C|A} = J - A$ asks whether CAV adds value to attention. Combined(1,1) exceeds both single arms, yet $I = -11.85/-7.11/-5.89/-7.39$pp (Qwen 3B/7B/14B, Llama), with all pointwise intervals below zero. Negative interaction alone does not imply that combination is unhelpful.

**Additional rescues must exceed lost successes.** Across all 4,600 paired indices, let $g$ count successes gained by adding CAV and $\ell$ count successes lost:

$$\Delta_{C|A} = 100(g - \ell)/4600 \quad \text{(pp)}. \tag{2}$$

Figure 2 reports both terms for all 36 model–strength combinations. Every comparison contains gains and losses; losses exceed gains in 24, gains exceed losses in 12. These are descriptive counts over the tested grid, not independent replications. They compare policy outcomes at matched indices, not before–after changes to an identical prompt: each arm generates its own history.

**Marginal benefit depends on strength.** The full contrast surface (Appendix Figure 6) shows +5.78 to +11.41 pp when CAV(1) or CAV(2) is added to Attention(1), with pointwise intervals above zero in all models and family-108 bounds above zero only in Qwen. At attention strengths 2 and 3, no tested CAV addition has a positive pointwise interval excluding zero; several additions cause substantial

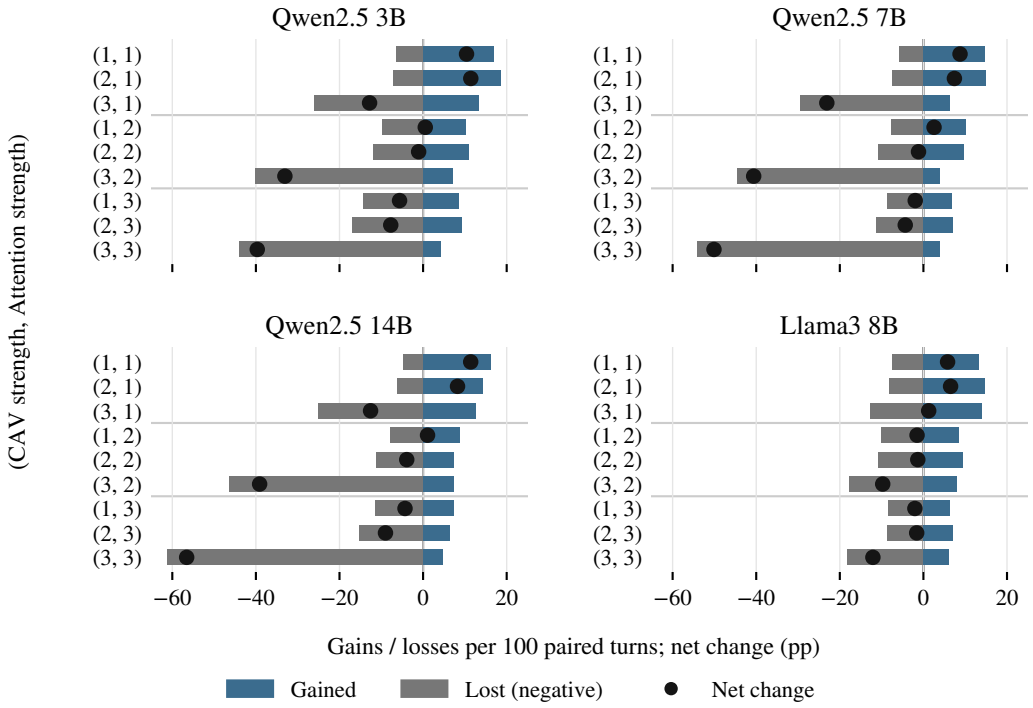


Figure 2: Full-grid accounting of adding CAV to attention. Each row compares Combined$(c, a)$ with Attention$(a)$; blue bars show gains and gray bars losses, each as a percentage of all 4,600 paired indices. Black dots are net differences (pp). All nine strength pairs and four models share one scale. These policy contrasts do not isolate an identical-input effect. Table 12 gives net intervals; Figure 6 marks family-adjusted exclusions of zero.

losses. This is not evidence of equivalence for uncertain cells. All grid means, counts and descriptive maxima are retained in Appendix I, without selected-winner inference.

**Complementary rescues are not reliably retained.** At unit strength, only 31–43% of the single-arm rescue union is common to both arms. Combined steering loses 16–31% of that union but rescues 12.5–17.5% of cases neither arm rescues. Complete behavioral redundancy is therefore not observed. Overlap, lost rescues and harms on baseline successes enter an exact accounting of interaction (Appendix I); these descriptive terms and the bounded outcome scale are not mechanisms.

### 6.4 Why Might the Interventions Interact?

We test two possible sources of interaction: coupling within the current computation and effects carried through different histories. They can coexist. Appendix Figure 8 displays the two tests; neither identifies mediation. These checks concern unit strength, not every grid cell.

**Within-response coupling: does the adaptive patch cancel attention's change?** At $\lambda_C = 1$, the ideal CAV rule sets $c^\top h$ to $p_1$, removing any incoming change along $c$. First-step measurements confirm this local cancellation, not its responsibility for behavioral losses. The reference policy computes coefficients in an attention-free CAV-only pass on the same prefix and applies them in the attention-steered pass. It improves on Combined(1,1) by +5.80/+1.46/+2.76/+3.41 pp (Qwen 3B/7B/14B, Llama), with family-16 bounds above zero for Qwen3B and Llama; interaction point estimates remain negative. Only Qwen3B exceeds all four scalar controls under that adjustment. The coefficient remains prefix-dependent, costs one extra forward per token and changes subsequent histories. This is a policy test, not identified mediation (Appendix K).

**Across-turn propagation: do different histories account for the interaction?** A fixed panel supplies all four arms with each model's baseline history: one outcome-independent session per type at turns 1/10/20, or 69 inputs/model. The fixed-minus-free interaction change has an interval including zero in every model (Qwen 3B: +15.94 pp [−0.04, +33.33]). This small panel establishes neither that history divergence explains subadditivity nor that history has no effect (Appendix L).

### 6.5 Practical Scope of the Compliance Gains

**Answer quality.** Interventions change length, stopping and repetition. Baseline cap-hit rates of 80.20/62.35/80.09/74.43% fall to 47.98/37.33/52.59/48.09% under Combined(1,1) (Qwen 3B/7B/14B, Llama). At CAV strength 3, Qwen 3B's duplicate-trigram rate reaches 67.91%, versus 2.86% at baseline. Appendix Figure 5 pairs every single-arm dose with its repetition proxy; Table 38 completes quality reporting for the added grid cells. These proxies delimit utility, not the cause of joint losses.

**Prompt-level control.** Repeating the instruction at every later user turn yields pass rates of 55.87–68.17% across models. Gains over Combined(1,1) range from 4.57 to 10.30 pp, with pointwise intervals above zero in every model. Appendix M reports each model's result. Re-insertion changes exposure and history; it is neither a mechanistic isolation nor a performance ceiling.

## 7 Discussion

**Decodability is not functional access.** The probe recovers constraint identity before checker-failed answers, including in never-passing sessions. Excluding the current answer resolves direct answer leakage, not whether generation uses the decoded information. The instruction and earlier transcript remain visible, and checker failure is not a complete validation of constraint violation.

**Complementarity is not composability.** Distinct single-arm rescues establish behavioral complementarity on this bank. Successful combination additionally requires retaining those successes and avoiding harm elsewhere. The full grid separates rescue availability from net utility across strength levels. The reference-policy and fixed-history checks constrain possible explanations but establish neither a shared mediator nor the natural mechanism of compliance decay.

**Implications.** Internal controls should be evaluated by net benefit across a declared strength range, not selected rescues alone. Our controlled dataset makes persistence and policy comparisons measurable, while its form-only checks leave semantic utility unresolved. Improving those scores and explaining the model's ordinary computation remain different objectives.

## 8 Limitations

All results concern imperfect repository checks, not official IFEval scores or verified semantic utility. Prompt persistence is not uniformly explicit, and corrected-checker sensitivity remains untested. The study reuses a previously examined 230-session bank, greedy decoding and seed 42. Adjustments cover only the reference, grid and four primary pre-response accuracies; maxima have no selected-winner inference. Qwen sites were selected on this bank under an earlier cap; Llama sites are exploratory. Equal caps do not equalize length or truncation. Type exclusions partly reflect observed behavior, and ten variants per type do not represent all constraints. Pre-response decoding is not functional use; historical geometry is post-response. Policy effects do not identify a mediator; histories differ and the fixed panel is small. Appendix V expands these limits.

## 9 Conclusion

On this constructed evaluation set, constraint identity is decodable before checker-failed answers, and individually useful interventions need not combine reliably. CAV and attention improve checker pass rates at unit strength, yet joint gains are subadditive. Across the complete joint grid, additional rescues coexist with lost successes, with losses dominating in 24 of 36 comparisons. Modest CAV steering adds value to weak attention; neither superiority over strong attention nor a mechanism of

joint failure is established. Faithful outcome validation, recoverable information, useful control and mechanistic explanation remain distinct requirements.

## ETHICS STATEMENT

This work studies representations and intervention policies in publicly available language models, not a validated deployment or safety-bypass method. Sessions adapt public IFEval and Dolly-15K resources and do not involve newly recruited human subjects. Model and source-data license terms apply separately. In particular, Qwen2.5-3B-Instruct uses the Qwen Research License, not Apache 2.0; it must not be grouped under the other Qwen checkpoints' licensing. IFEval is Apache 2.0 and Dolly-15K is CC BY-SA 3.0. Any release must preserve applicable attribution and redistribution requirements. Custom checker limitations are disclosed to avoid presenting proxy scores as verified instruction satisfaction.

## REPRODUCIBILITY STATEMENT

Section 4 and Appendices B–C specify sessions and CAV extraction; Appendix F lists frozen sites and precision. Appendix W documents the new response-excluded extraction, input hashes, fixed folds and native probe analysis. Appendices I and U account for all 108 reported intervention cells. Phase 3 uses batch size 1, greedy decoding, eager attention, seed 42 and cap 256, in float16 for Qwen3B and bfloat16 otherwise. Tables import completed native results without repeated fitting or bootstrap. Historical cache-prefix hashes remain absent (Appendix O).

## AI USE STATEMENT

In this work, the authors used generative AI tools to aid in polishing the writing and improving the readability of the manuscript, and to assist with writing and debugging the experimental code. The authors carefully reviewed, verified, and edited all AI-generated content and code, and take full responsibility for the content and correctness of this final work.

## REFERENCES


Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (Workshop)*, 2017.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1376–1395, 2024. doi: 10.18653/v1/2024.findings-emnlp.74. URL https://aclanthology.org/2024.findings-emnlp.74/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Beatriz Canaverde, Duarte M. Alves, José Pombal, Giuseppe Attanasio, and André F. T. Martins. SEQUOR: A multi-turn benchmark for realistic constraint following. *arXiv preprint arXiv:2605.06353*, 2026. URL https://arxiv.org/abs/2605.06353.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2126–2136, 2018.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned LLM. *Databricks Blog*, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 16124–16170, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

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. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, Proceedings of Machine Learning Research, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2305.01610.

Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley Ren, Udhay Nallasamy, Andy Miller, and Jaya Narain. Do LLMs "know" internally when they follow instructions? In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.14516.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4129–4138, 2019.

Qi Jia, Ye Shen, Xiujie Song, Kaiwei Zhang, Shibo Wang, Dun Pei, Xiangyang Zhu, and Guangtao Zhai. One battle after another: Probing LLMs' limits on multi-turn instruction following with a benchmark evolving framework. *arXiv preprint arXiv:2511.03508*, 2025. URL https://arxiv.org/abs/2511.03508.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688, 2024. doi: 10.18653/v1/2024.acl-long.257. URL https://aclanthology.org/2024.acl-long.257/.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2668–2677, 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Jinnan Li, Jinzhe Li, Yue Wang, Yi Chang, and Yuan Wu. StructFlowBench: A structured flow benchmark for multi-turn instruction following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9322–9341, 2025. doi: 10.18653/v1/2025.findings-acl.486. URL https://aclanthology.org/2025.findings-acl.486/.

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*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

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. URL https://aclanthology.org/2024.tacl-1.9/.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

nostalgebraist. interpreting GPT: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Yanzhao Qin, Tao Zhang, Tao Zhang, Yanjun Shen, Wenjing Luo, Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen, Zenan Zhou, Wentao Zhang, and Bin Cui. SysBench: Can large language models follow system messages? *arXiv preprint arXiv:2408.10943*, 2024. URL https://arxiv.org/abs/2408.10943.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=wozhdnRCtw.

Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. Parrot: Enhancing multi-turn instruction following for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9729–9750, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.525. URL https://aclanthology.org/2024.acl-long.525/.

Ian Tenney, Dipanjan Das, and Eleni Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.02262.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

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*, 2023. URL https://arxiv.org/abs/2310.01405.

# A Theoretical Derivations

**Residual stream drift.** At each layer $\ell$, the attention output at the generation position decomposes into instruction and non-instruction components:

$$\mathrm{Attn}_{\mathrm{gen}}^{(\ell)} = \sum_{i \in \mathcal{I}} \alpha_{\mathrm{gen},i} \mathbf{v}_i W_O + \sum_{j \notin \mathcal{I}} \alpha_{\mathrm{gen},j} \mathbf{v}_j W_O \tag{3}$$

As non-instruction tokens grow by roughly $(t-1)n$ in total, their contributions can change the readout direction. A smaller dot product with a particular unembedding vector lowers that token's logit, holding other factors fixed; constraint satisfaction is a sequence-level property, so this does not entail a decline in compliance.

**Attention dilution bound.** Let $m = |\mathcal{I}|$, $N_t$ be the number of non-instruction keys, and $r_t = \overline{e^s}_{\neg\mathcal{I}} / \overline{e^s}_{\mathcal{I}}$. Exactly,

$$A_{\mathcal{I}}(t) = \frac{m}{m + N_t r_t} \leq \frac{m}{m + cN_t} = O\left(\frac{m}{(t-1)n}\right), \tag{4}$$

where the inequality assumes $r_t \geq c > 0$ uniformly and the order bound assumes $N_t = \Theta((t-1)n)$. A ratio of means of exponentials cannot generally be replaced by the exponential of a difference of mean logits. Without the assumption, attention can stay flat or increase. A normalized $1/(t-1)$ reference falls 19-fold from turns 2 to 20; $1/t$ falls tenfold. Neither is entailed at finite turns by the order bound. Observed declines of 4.5-fold (Qwen7B) and 5.6-fold (Llama) do not establish or refute the assumption.

**Relationship between the two quantities.** Attention changes residual updates and residual changes can affect later queries, so coupling is plausible. Directional rotation and softmax normalization motivate different interventions, but neither their distinct mathematical descriptions nor behavioral overlap establishes distinct natural causes or a common mediator. Historical geometry and attention were also measured at different time boundaries.

# B Instruction Variant Construction and Scoring Contract

We adapt constraint types represented in the 541-prompt IFEval source. We exclude `repeat_prompt` because our turns use different content questions, and `response_language` based on an earlier observation of little decay; the latter is outcome-informed selection, not a universal claim about language constraints. The constructed file contains 230 distinct instruction strings. Fifteen types vary parameter dictionaries, sometimes alongside wording; eight use phrasing only. Ten wordings do not imply ten parameter settings: postscript has two and paragraph count four. Table 4 gives the exact counts.

The filler pool has 119 Dolly-derived questions; seed 42 samples 20 entries without replacement separately for each session, so content overlaps across sessions but sequences can differ. This is not a held-out-topic design. Instructions and questions are concatenated without a common persistence wrapper. The protocol scores the initial target on all turns, even when its wording is ambiguous about later answers. For constrained-response variants, the templates request choices without specifying options; the checker substitutes a word-length test. This known prompt–checker mismatch remains in the reported aggregates, not silently excluded after observing results.

Table 4: Dataset and implemented scoring contract for all 23 types. Each row has 10 instruction variants and 200 response opportunities per model/condition. $K$ counts distinct parameter dictionaries, not distinct wordings. These are frozen repository rules, not an assertion of official IFEval equivalence.

| Type | $K$ | Actual binary pass rule |
|---|---|---|
| Case: capital word frequency | 10 | Count whitespace tokens that are alphabetic and uppercase; apply the specified lower or upper threshold. |

Table 4 continued

| Type | $K$ | Actual binary pass rule |
|---|---|---|
| Case: English capital | 1 | At least one alphabetic character, and all alphabetic characters uppercase; does not test English language. |
| Case: English lowercase | 1 | At least one alphabetic character, and all alphabetic characters lowercase. |
| Combination: two responses | 1 | Contains the literal six-asterisk separator; does not verify two distinct answers. |
| Content: placeholders | 10 | At least the requested number of nonempty square-bracket spans. |
| Content: postscript | 2 | Contains the requested marker anywhere, not necessarily at the end. |
| Format: constrained response | 1 | At most 50 whitespace words. No choice-membership test; the instruction templates do not supply answer options. This is a proxy mismatch, not valid constrained-choice verification. |
| Format: JSON | 1 | Parses as any JSON value after stripping whitespace or extracting a fenced block; neither an object nor JSON-only surrounding text is required. |
| Format: multiple sections | 10 | At least the requested number of splitter-string occurrences, not exactly that many headed sections. |
| Format: bullet lists | 10 | Exactly the requested total of line-initial bullet or numbered-list matches. |
| Format: highlighted sections | 10 | At least the requested count of asterisk-delimited regex matches; does not verify semantic sections. |
| Format: title | 1 | A double-angle-bracket span occurs anywhere; not necessarily an initial title. |
| Keywords: existence | 10 | Every keyword occurs as a case-insensitive substring. |
| Keywords: forbidden words | 10 | No forbidden term occurs as a case-insensitive substring. |
| Keywords: frequency | 10 | Count case-insensitive, non-overlapping substring occurrences and apply the specified threshold; not word-boundary matching. |
| Keywords: letter frequency | 10 | Count case-insensitive occurrences and apply the specified threshold. |
| Length: nth paragraph first word | 10 | Exact blank-line paragraph count and a matching first word in the specified paragraph, after lowercasing and limited punctuation stripping. |
| Length: paragraphs | 4 | Exact count of nonempty pieces split on double newlines. |
| Length: sentences | 10 | Count nonempty pieces split on periods, exclamation or question marks; apply the specified threshold. |
| Length: words | 10 | Whitespace word count satisfies the specified threshold. |
| Punctuation: no comma | 1 | No ASCII comma occurs. |
| Start/end: ending | 10 | Whitespace-stripped response ends with the exact requested phrase. |
| Start/end: quotation | 1 | Whitespace-stripped response starts and ends with a double quotation mark. |

**Example session.** **User (Turn 1):** "Wrap the entire response in valid JSON using curly braces and key-value pairs. \n\n First day of baking with sourdough." Turns 2–20 present only filler questions unrelated to the constraint (e.g., "Explain the history of the printing press"). Compliance is measured at each turn by whether the response is valid JSON.

## C CAV Extraction Details

**Neutral text selection.** The **without** condition uses Wikipedia-derived expository text truncated to the instruction's token count before decoding; paired templated inputs are left-padded to equal length. This avoids using an explicit alternative instruction, but does not prove semantic neutrality or exact post-template token matching. Each type uses 200 pairs (10 variants and the same seeded set of 20 fillers). We did not establish a convergence threshold for this mean-difference estimator; the original TCAV method motivates directional analysis, not the choice of 200 pairs.

**Instruction-position stability.** Under causal attention, instruction-position states cannot depend on later tokens, so they are identical across turns for an unchanged prefix. Our historical drift diagnostic

instead uses last-token states extracted after including the current response. These show changing readout representations, not the state that caused that answer (Appendix O).

## D Instruction-Attending Head Identification

At turn 2's generation position, we compute the mean attention share each (layer, head) pair allocates to instruction-span tokens across all sessions. We rank all heads and select the top-$K$, where $K$ is determined by the $K$ sweep (Appendix F): $K = 192$ for each Qwen model, and Llama, which was not swept, borrows $K = 192$ from Qwen2.5 7B. Turn 2 is used because attention dilution is not yet substantial.

## E Per-Layer Probe Accuracy

Figure 3 shows historical post-response probe accuracy across layers and turns for all three Qwen scales. The selected contiguous bands are 21–27 (3B, reference 21), 13–19 (7B, reference 14) and 23–29 (14B, reference 25). This is exploratory localization on the same bank, not an independent layer-selection test or a pre-response measurement. The plotted signal persists late in the dialogue.

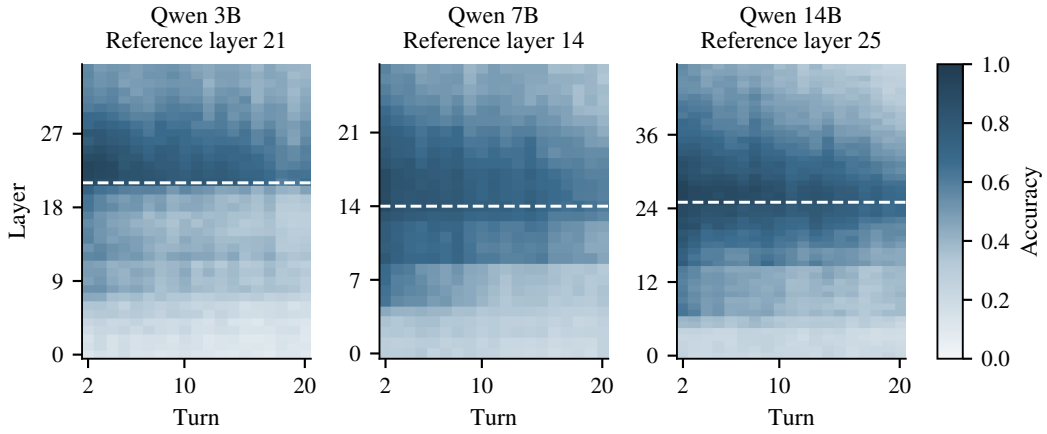


Figure 3: Historical post-response probe accuracy across the three Qwen scales; the current answer is included. White dashed lines mark the fixed reference layers. The same-bank layer patterns are exploratory, not independently validated localization or pre-response evidence.

## F Execution Sites, Lengths and Quality Proxies

Table 5 lists the frozen sites used by every Phase 3 cell.

**Site selection.** The sites were chosen by outcome-based sweeps, not by Phase 2 alone. For each Qwen model, the original sweeps generated all 230 sessions of this bank under the earlier protocol, with a 512-token response cap, and scored each setting by mean checker pass rate. Every layer was ranked by compliance under single-layer CAV at strength 1. The number of CAV layers $N$ was the compliance argmax over $\{1, 3, 5, 7\}$, applied to the top-ranked layers. The head count $K$ was the compliance argmax over $\{24, 48, 96, 192\}$ under attention steering at strength 1, with heads ranked by turn-2 instruction attention share (Appendix D). Each Qwen model selected $N = 7$ and $K = 192$. The selected layers coincide with the probe bands of Appendix E for 3B (21–27) and 7B (13–19) but not for 14B (25–29, 31 and 34, versus 23–29). Llama was not swept. Its layers 8–14 are an exploratory band of the seven layers nearest layer 11, placed from earlier correlational drift evidence rather than a compliance sweep and away from its probe reference layer (30), and $N = 7$ and $K = 192$ are borrowed from the Qwen2.5 7B sweeps. The CAV directions, session targets $p_1$,

head ranking and attention calibration $\delta_s$ come from Phase 2. Because the selection used the same bank, every Phase 3 cell is in-sample with respect to site choice, and the sites are not tuned for the 256-token cap.

**Implementation.** The attention bias uses the expanded query-head layout for grouped-query attention; the CAV patch acts on the last-token residual output at each decoding step. Llama's middle-layer band is exploratory.

| Model | Precision | Cap | CAV layers (zero-based) | Heads |
|---|---|---|---|---|
| Llama 8B | BF16 | 256 | 8, 9, 10, 11, 12, 13, 14 | 192 |
| Qwen 3B | FP16 | 256 | 21, 22, 23, 24, 25, 26, 27 | 192 |
| Qwen 7B | BF16 | 256 | 13, 14, 15, 16, 17, 18, 19 | 192 |
| Qwen 14B | BF16 | 256 | 25, 26, 27, 28, 29, 31, 34 | 192 |

Table 5: Frozen execution sites and precision, read from the run registration. Each model uses seven CAV layers (zero-based) and 192 attention heads.

**Quality proxies.** Original-matrix tables report stored descriptive proxies for the 84 displayed cells (Appendix U). *Tokens*: mean generated tokens. *Cap*: % reaching cap 256. *Rep-3g*: mean duplicate word-trigram fraction, $1 - \text{unique}/\text{total}$ over lower-cased whitespace tokens, 0 for fewer than three words. *Heavy*: % with Rep-3g at least 0.5. *No-word*: % with no alphanumeric word under that implementation. *Harm*/*Rescue*: baseline-successful pairs lost and baseline-failed pairs gained. The native integer-grid analysis separately uses Unicode word tokens, requiring at least ten words for its heavy-repeat flag; these different repetition definitions are not pooled. Neither proxy measures semantic quality, and repetition and length are not independent under the cap.

## G Per-Constraint Heterogeneity

Figure 4 and Table 6 report all 23 types at CAV(1) and Combined(1,1), for all four models. Each type contributes ten sessions and 200 turns per condition. Intervals resample the ten sessions within each type (whole-session paired bootstrap, 10,000 draws, seed 42) and are uncorrected across comparisons; ten-cluster percentile intervals may under-cover. Combined has intervals above zero for 16/14/15/16 types (Qwen 3B/7B/14B, Llama) and CAV for 14/11/11/10. Qwen 7B's Combined forbidden-words contrast has an interval below zero. Floor/ceiling effects and the scoring rules in Table 4 limit interpretation.

Table 6: Per-type compliance change (pp) versus the same model's baseline, cap 256. L8: Llama 3 8B; Q3/Q7/Q14: Qwen2.5 3B/7B/14B. **Bold**: 95% interval above 0; underlined: interval below 0. Exploratory and uncorrected.

| | CAV 1: Δ vs. baseline (pp) | | | | Combined 1/1: Δ vs. baseline (pp) | | | |
|---|---|---|---|---|---|---|---|---|
| Constraint type | Q3 | Q7 | Q14 | L8 | Q3 | Q7 | Q14 | L8 |
| Capital-word count | **+33.5** | +4.0 | **+29.0** | +10.5 | **+35.0** | **+23.5** | +23.0 | +9.5 |
| Uppercase letters | **+34.0** | **+35.5** | **+22.0** | +7.5 | **+30.5** | **+29.5** | **+26.0** | +5.0 |
| Lowercase letters | **+49.5** | **+51.0** | **+75.0** | **+30.5** | **+47.5** | **+68.5** | **+74.5** | **+32.5** |
| Two responses | **+49.0** | +10.5 | **+21.5** | **+16.0** | **+69.0** | **+34.0** | **+51.0** | **+22.5** |
| Placeholders | **+20.0** | **+3.0** | +0.5 | **+12.5** | **+44.0** | **+43.5** | **+4.5** | **+30.0** |
| Postscript | **+11.5** | **+38.0** | **+6.0** | **+24.0** | **+26.0** | **+60.0** | **+18.0** | **+45.0** |
| Constrained response | **+46.5** | **+19.0** | **+34.5** | +13.5 | **+50.0** | **+21.0** | **+37.5** | **+23.5** |
| JSON | **+54.0** | **+17.0** | **+34.5** | +21.0 | **+60.0** | **+17.5** | **+38.5** | **+27.5** |
| Multiple sections | **+43.0** | **+40.5** | **+28.5** | **+12.0** | **+69.0** | **+45.5** | **+49.5** | +11.5 |
| Bullet count | +0.0 | **+14.0** | **+25.0** | +9.5 | +11.0 | **+15.5** | **+56.5** | **+35.0** |
| Highlights | **+18.0** | **+14.5** | **+16.5** | **+16.0** | **+28.5** | +11.0 | **+20.0** | **+22.0** |
| Title | **+22.5** | +18.0 | +10.0 | −0.5 | **+31.0** | **+37.5** | +19.0 | +1.5 |
| Required keywords | −1.5 | +2.5 | −1.0 | +2.0 | +7.5 | **+3.5** | +0.5 | **+28.5** |
| Forbidden words | +0.0 | −1.5 | +0.5 | +2.0 | −0.5 | −4.0 | +0.5 | +2.5 |
| Keyword frequency | −2.5 | +1.0 | −1.0 | +8.0 | −1.0 | +1.5 | +0.5 | **+6.0** |

Table 6 continued

| | CAV 1: Δ vs. baseline (pp) | | | | Combined 1/1: Δ vs. baseline (pp) | | | |
|---|---|---|---|---|---|---|---|---|
| Constraint type | Q3 | Q7 | Q14 | L8 | Q3 | Q7 | Q14 | L8 |
| Letter frequency | −2.5 | +0.5 | −1.5 | +0.5 | +1.5 | −2.5 | +0.5 | **+12.5** |
| Nth paragraph first word | −1.0 | +0.0 | +0.0 | +0.5 | +0.0 | +0.0 | **+9.0** | **+14.0** |
| Paragraph count | −1.0 | −5.0 | +6.0 | **+11.5** | −4.0 | −3.0 | +2.5 | +12.5 |
| Sentence count | **+34.5** | +3.5 | +1.0 | +9.0 | **+38.5** | +3.0 | **+19.0** | **+36.0** |
| Word count | +10.5 | +5.0 | +2.0 | −2.0 | **+19.0** | +6.0 | **+5.0** | +13.0 |
| No comma | **+23.5** | +20.0 | +7.0 | **+22.0** | **+25.5** | +21.5 | +19.0 | **+27.0** |
| Ending phrase | +2.5 | **+11.5** | +5.0 | **+9.5** | **+28.0** | **+47.5** | **+17.5** | **+22.0** |
| Quotation | **+67.5** | **+62.5** | **+76.5** | **+67.5** | **+68.5** | **+69.0** | **+77.5** | **+59.5** |
| #types $\Delta > 0$ | 16 | 20 | 19 | 21 | 19 | 19 | 23 | 23 |
| #types CI $> 0$ | 14 | 11 | 11 | 10 | 16 | 14 | 15 | 16 |
| #types CI $< 0$ | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

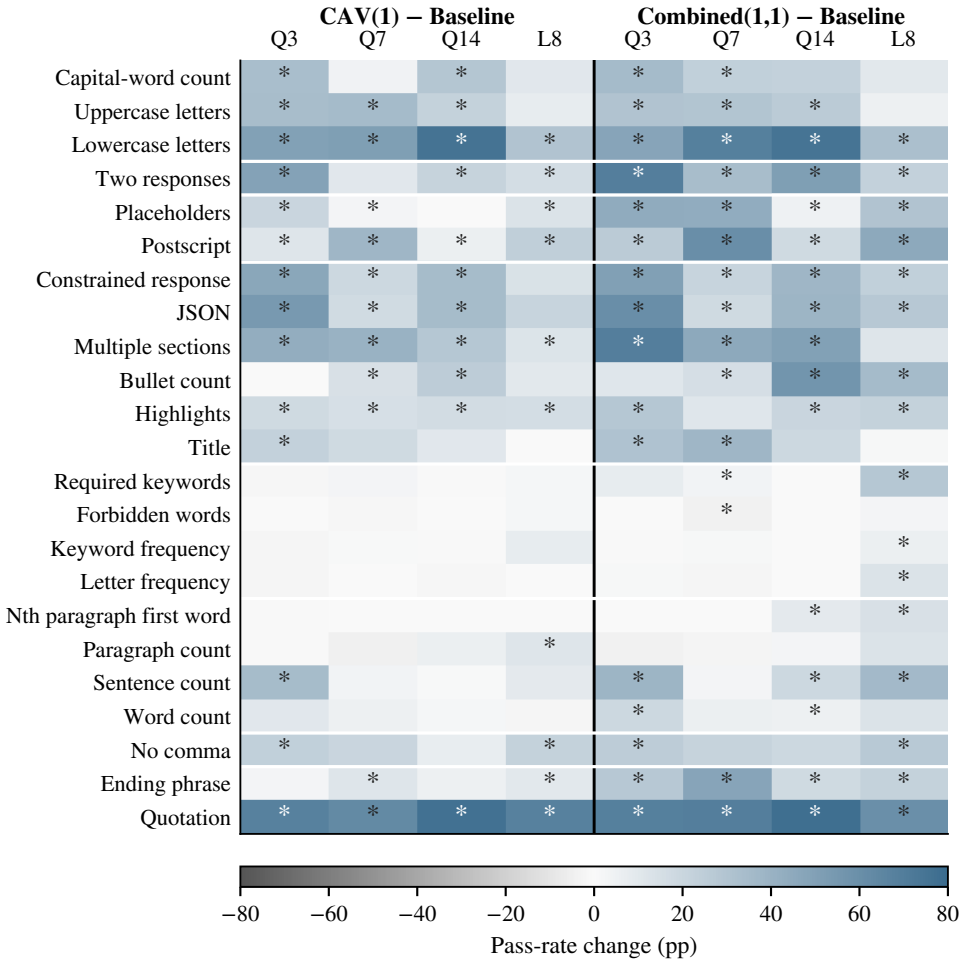


Figure 4: All 184 type–model–policy effects, with a common zero-centered scale. Colors encode pass-rate differences (pp); stars mark pointwise, unadjusted 95% intervals excluding zero. Every type has ten sessions. In particular, the constrained-response row measures the length proxy, not choice validity. Exact original entries remain in Table 6.

## H DOSE RESPONSE

**Descriptive dose and quality patterns.** The largest CAV means occur at $\lambda_C = 3$ in Llama (+15.46 pp), 2 in Qwen 3B/7B (+23.61/+17.09 pp) and 1.5 in 14B (+17.35 pp). At $\lambda_C = 3$, Qwen 3B's duplicate word-trigram rate is 67.91%, versus 2.86% at baseline. Attention gains grow up to $\lambda_A = 3$ in Qwen (+37.96/+39.80/+47.54 pp) and peak at $\lambda_A = 2$ in Llama (+25.76 pp). Attention(3) has trigram rates of 23.76% and 27.26% in Qwen 3B and Llama, versus 2.86% and 6.72% at baseline; increases in Qwen 7B/14B are smaller (+1.74/+2.83 pp). These are descriptive maxima and proxies, not selected-winner inference or measurements of semantic quality.

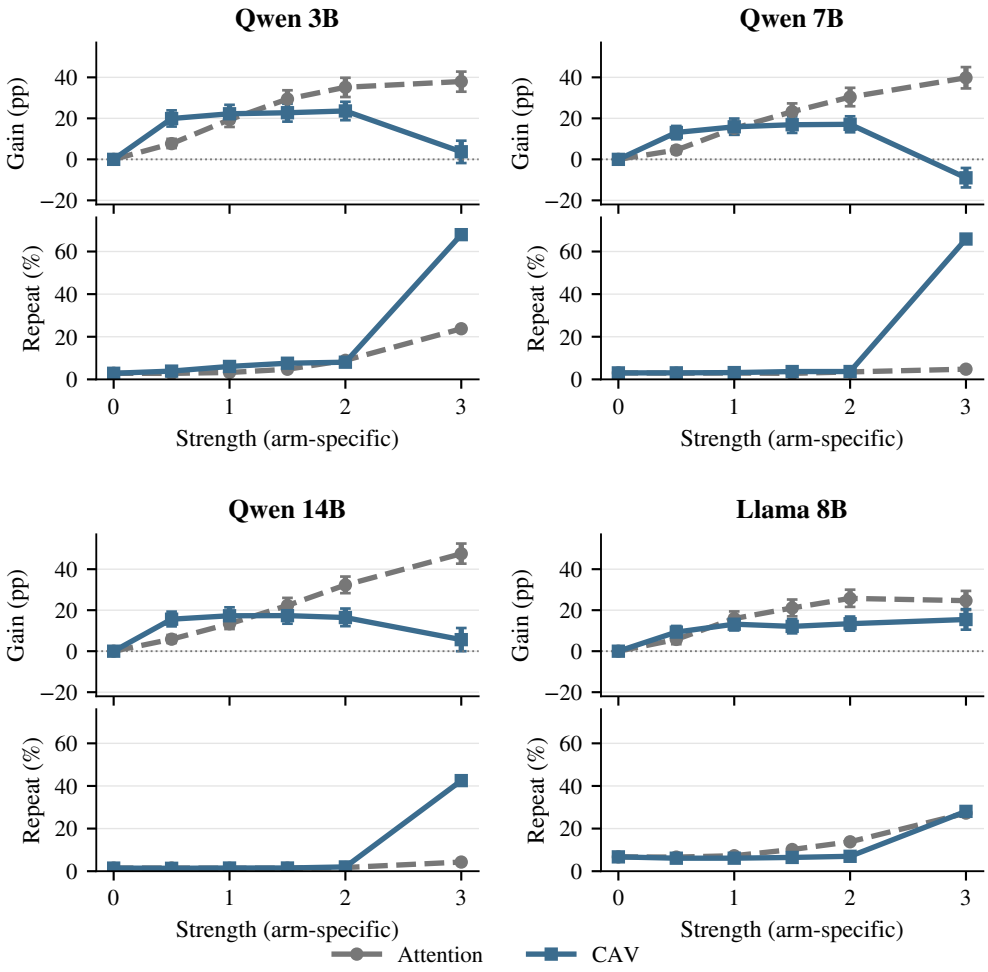


Figure 5: Effectiveness and repetition on the entire single-arm grid at frozen sites, cap 256. Within each model panel, the upper plot shows checker pass-rate changes with stored paired whole-session 95% intervals; the lower plot shows descriptive duplicate word-trigram rates under the original whitespace-token proxy; no quality intervals were estimated. Strength 0 is baseline. The strength units differ across arms, and repetition is not a semantic-quality score.

Tables 7 and 8 and Figure 5 use the common five-point grid $\{0.5, 1, 1.5, 2, 3\}$. This reporting-range restriction was made after results were available; original registrations and raw out-of-range records remain archived (Appendix U). All negative outcomes within the reported range are retained. The maxima are descriptive, not validated optima, and the strength axes are not matched physical doses.

Table 7: CAV dose response ($\lambda = 0$ is the baseline). $\Delta$ and intervals are relative to the baseline; Cap, Rep-3g and Heavy are defined in Appendix F.

| Model | $\lambda$ | CR (%) | $\Delta$pp [95% CI] | Cap (%) | Rep-3g (%) | Heavy (%) |
|---|---|---|---|---|---|---|
| Llama 8B | 0 | 37.24 | – | 74.43 | 6.72 | 0.13 |
| Llama 8B | 0.5 | 46.54 | +9.30 [+6.3, +12.3] | 61.63 | 6.10 | 0.04 |
| Llama 8B | 1.0 | 50.41 | +13.17 [+10.1, +16.4] | 55.57 | 6.11 | 0.15 |
| Llama 8B | 1.5 | 49.33 | +12.09 [+8.7, +15.6] | 53.35 | 6.46 | 0.15 |
| Llama 8B | 2.0 | 50.65 | +13.41 [+10.0, +16.9] | 51.41 | 6.99 | 0.74 |
| Llama 8B | 3.0 | 52.70 | +15.46 [+10.5, +20.5] | 64.07 | 28.08 | 25.87 |
| Qwen 3B | 0 | 21.54 | – | 80.20 | 2.86 | 0.30 |
| Qwen 3B | 0.5 | 41.46 | +19.91 [+16.0, +23.9] | 55.85 | 3.88 | 1.46 |
| Qwen 3B | 1.0 | 43.78 | +22.24 [+18.0, +26.6] | 53.09 | 6.07 | 3.35 |
| Qwen 3B | 1.5 | 44.28 | +22.74 [+18.4, +27.1] | 53.02 | 7.58 | 4.87 |
| Qwen 3B | 2.0 | 45.15 | +23.61 [+19.1, +28.1] | 51.76 | 8.06 | 5.11 |
| Qwen 3B | 3.0 | 25.24 | +3.70 [−1.8, +9.1] | 86.70 | 67.91 | 71.37 |
| Qwen 7B | 0 | 31.07 | – | 62.35 | 3.04 | 0.41 |
| Qwen 7B | 0.5 | 44.17 | +13.11 [+9.8, +16.3] | 48.09 | 3.06 | 0.50 |
| Qwen 7B | 1.0 | 46.93 | +15.87 [+12.0, +19.9] | 45.04 | 3.18 | 0.46 |
| Qwen 7B | 1.5 | 47.96 | +16.89 [+12.9, +20.9] | 44.74 | 3.67 | 0.98 |
| Qwen 7B | 2.0 | 48.15 | +17.09 [+13.2, +21.0] | 45.54 | 3.68 | 1.26 |
| Qwen 7B | 3.0 | 22.07 | −9.00 [−13.7, −4.2] | 89.35 | 65.86 | 71.28 |
| Qwen 14B | 0 | 25.37 | – | 80.09 | 1.46 | 0.00 |
| Qwen 14B | 0.5 | 40.98 | +15.61 [+12.1, +19.3] | 61.37 | 1.36 | 0.04 |
| Qwen 14B | 1.0 | 42.65 | +17.28 [+13.4, +21.4] | 59.37 | 1.49 | 0.04 |
| Qwen 14B | 1.5 | 42.72 | +17.35 [+13.4, +21.5] | 58.33 | 1.50 | 0.07 |
| Qwen 14B | 2.0 | 41.78 | +16.41 [+12.2, +20.8] | 58.48 | 2.01 | 0.72 |
| Qwen 14B | 3.0 | 31.00 | +5.63 [−0.02, +11.3] | 91.91 | 42.51 | 47.39 |

Table 8: Attention dose response ($\lambda = 0$ is the baseline). Columns as in Table 7.

| Model | $\lambda$ | CR (%) | $\Delta$pp [95% CI] | Cap (%) | Rep-3g (%) | Heavy (%) |
|---|---|---|---|---|---|---|
| Llama 8B | 0 | 37.24 | – | 74.43 | 6.72 | 0.13 |
| Llama 8B | 0.5 | 42.98 | +5.74 [+3.4, +8.1] | 68.33 | 6.62 | 0.13 |
| Llama 8B | 1.0 | 53.15 | +15.91 [+12.5, +19.4] | 62.91 | 7.26 | 0.70 |
| Llama 8B | 1.5 | 58.33 | +21.09 [+17.0, +25.2] | 54.83 | 10.18 | 4.93 |
| Llama 8B | 2.0 | 63.00 | +25.76 [+21.6, +30.0] | 49.91 | 13.80 | 9.41 |
| Llama 8B | 3.0 | 61.89 | +24.65 [+20.1, +29.4] | 52.89 | 27.26 | 25.96 |
| Qwen 3B | 0 | 21.54 | – | 80.20 | 2.86 | 0.30 |
| Qwen 3B | 0.5 | 29.22 | +7.67 [+5.5, +10.0] | 71.85 | 2.81 | 0.48 |
| Qwen 3B | 1.0 | 40.91 | +19.37 [+15.8, +23.1] | 63.30 | 3.24 | 1.17 |
| Qwen 3B | 1.5 | 50.96 | +29.41 [+25.2, +33.7] | 57.72 | 4.74 | 2.65 |
| Qwen 3B | 2.0 | 56.72 | +35.17 [+30.4, +39.8] | 54.80 | 8.93 | 7.04 |
| Qwen 3B | 3.0 | 59.50 | +37.96 [+33.0, +42.8] | 62.70 | 23.76 | 23.85 |
| Qwen 7B | 0 | 31.07 | – | 62.35 | 3.04 | 0.41 |
| Qwen 7B | 0.5 | 35.61 | +4.54 [+2.7, +6.4] | 58.00 | 2.91 | 0.33 |
| Qwen 7B | 1.0 | 46.20 | +15.13 [+12.0, +18.4] | 50.35 | 2.98 | 0.39 |
| Qwen 7B | 1.5 | 54.33 | +23.26 [+19.3, +27.3] | 44.02 | 2.81 | 0.52 |
| Qwen 7B | 2.0 | 61.46 | +30.39 [+25.9, +34.9] | 37.09 | 3.50 | 0.78 |
| Qwen 7B | 3.0 | 70.87 | +39.80 [+34.6, +45.0] | 27.26 | 4.78 | 1.54 |
| Qwen 14B | 0 | 25.37 | – | 80.09 | 1.46 | 0.00 |
| Qwen 14B | 0.5 | 31.22 | +5.85 [+4.0, +7.9] | 73.15 | 1.50 | 0.04 |
| Qwen 14B | 1.0 | 38.74 | +13.37 [+10.8, +16.2] | 63.30 | 1.41 | 0.04 |
| Qwen 14B | 1.5 | 47.67 | +22.30 [+18.8, +26.0] | 55.72 | 1.44 | 0.07 |
| Qwen 14B | 2.0 | 57.65 | +32.28 [+28.3, +36.4] | 47.98 | 1.73 | 0.07 |
| Qwen 14B | 3.0 | 72.91 | +47.54 [+42.7, +52.5] | 40.33 | 4.29 | 1.61 |

# I JOINT CELLS, INTERACTION AND RESCUE OVERLAP

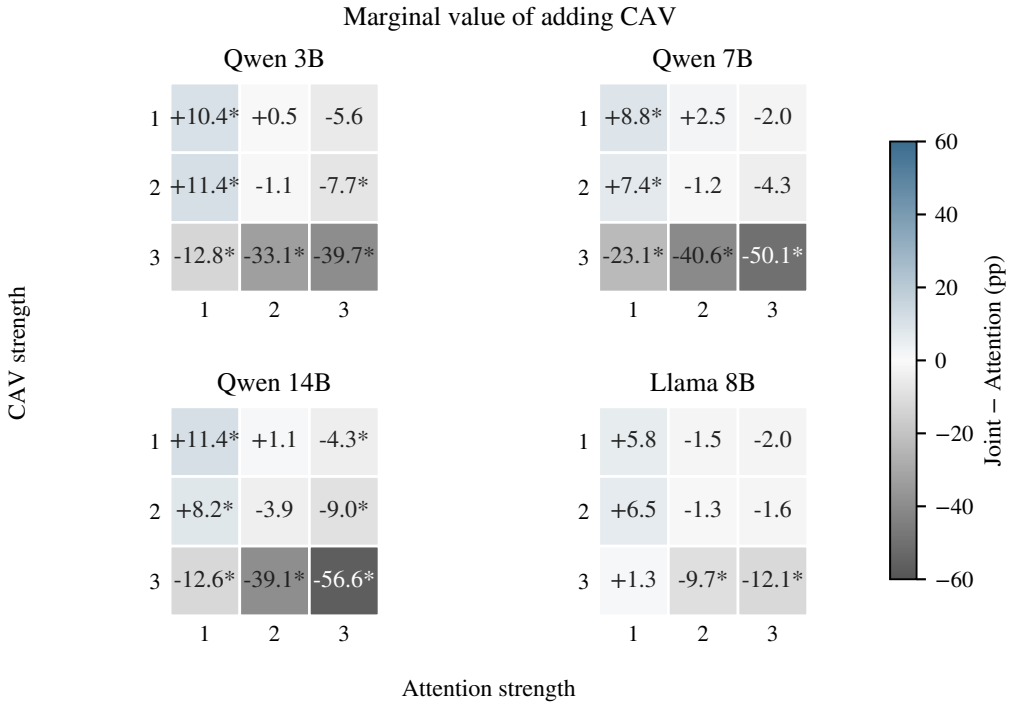


Figure 6: Marginal value of adding CAV to attention (pp), with a common zero-centered scale across models. All 36 joint cells are complete. A star marks an approximate family-of-108 interval excluding zero; all pointwise intervals are listed in Table 12. Each cell covers 230 sessions and 20 turns. Equal numerical strengths are not matched physical doses; this is not an independent confirmation set.

Table 9 preserves the original seven joint cells per model and their original intervals. Joint−Attn and Joint−CAV compare the single arm at the same strength; $I$ is the additive-scale interaction. The original grid was asymmetric. The completed extension now fills all nine integer pairs through 3 (Tables 10 and 12) and supplies native intervals for all three contrasts at every integer pair. Fractional-cell interactions outside this extension remain descriptive. No observed maximum establishes an attainable optimum.

Table 9: The original seven combined cells per model. Joint gain is relative to the baseline with a paired whole-session 95% interval. $I$ has an interval only at (1,1). The separate integer-grid extension is in Table 12.

| Model | $\lambda_C/\lambda_A$ | CR (%) | Joint gain [95% CI] | Joint−Attn | Joint−CAV | $I$ (pp) |
|---|---|---|---|---|---|---|
| Llama 8B | 0.5/1 | 59.46 | +22.22 [+18.4, +26.1] | +6.30 | +12.91 | −3.00 |
| Llama 8B | 1/1 | 58.93 | +21.70 [+17.8, +25.6] | +5.78 | +8.52 | −7.39 [−11.3, −3.5] |
| Llama 8B | 1.5/1 | 60.91 | +23.67 [+19.5, +27.8] | +7.76 | +11.59 | −4.33 |
| Llama 8B | 1/0.5 | 55.78 | +18.54 [+14.8, +22.4] | +12.80 | +5.37 | −0.37 |
| Llama 8B | 1/1.5 | 60.57 | +23.33 [+18.8, +27.9] | +2.24 | +10.15 | −10.93 |
| Llama 8B | 1/2 | 61.48 | +24.24 [+19.4, +29.1] | −1.52 | +11.07 | −14.70 |
| Llama 8B | 2/2 | 61.65 | +24.41 [+20.0, +28.9] | −1.35 | +11.00 | −14.76 |
| Qwen 3B | 0.5/1 | 50.70 | +29.15 [+24.7, +33.7] | +9.78 | +9.24 | −10.13 |
| Qwen 3B | 1/1 | 51.30 | +29.76 [+25.2, +34.5] | +10.39 | +7.52 | −11.85 [−16.1, −8.0] |
| Qwen 3B | 1.5/1 | 51.65 | +30.11 [+25.3, +34.9] | +10.74 | +7.37 | −12.00 |
| Qwen 3B | 1/0.5 | 47.35 | +25.80 [+21.3, +30.4] | +18.13 | +3.57 | −4.11 |
| Qwen 3B | 1/1.5 | 55.33 | +33.78 [+29.0, +38.5] | +4.37 | +11.54 | −17.87 |

Table 9 continued

| Model | $\lambda_C/\lambda_A$ | CR (%) | Joint gain [95% CI] | Joint−Attn | Joint−CAV | $I$ (pp) |
|---|---|---|---|---|---|---|
| Qwen 3B | 1/2 | 57.24 | +35.70 [+30.7, +40.5] | +0.52 | +13.46 | −21.72 |
| Qwen 3B | 2/2 | 55.65 | +34.11 [+29.2, +38.9] | −1.07 | +10.50 | −24.67 |
| Qwen 7B | 0.5/1 | 53.61 | +22.54 [+18.4, +26.7] | +7.41 | +9.43 | −5.70 |
| Qwen 7B | 1/1 | 54.96 | +23.89 [+19.7, +28.2] | +8.76 | +8.02 | −7.11 [−10.5, −3.6] |
| Qwen 7B | 1.5/1 | 54.00 | +22.93 [+18.7, +27.2] | +7.80 | +6.04 | −9.09 |
| Qwen 7B | 1/0.5 | 49.02 | +17.96 [+14.0, +22.0] | +13.41 | +2.09 | −2.46 |
| Qwen 7B | 1/1.5 | 58.20 | +27.13 [+22.8, +31.5] | +3.87 | +11.26 | −12.00 |
| Qwen 7B | 1/2 | 64.00 | +32.93 [+28.2, +37.7] | +2.54 | +17.07 | −13.33 |
| Qwen 7B | 2/2 | 60.28 | +29.22 [+24.5, +33.9] | −1.17 | +12.13 | −18.26 |
| Qwen 14B | 0.5/1 | 49.09 | +23.72 [+19.8, +27.9] | +10.35 | +8.11 | −5.26 |
| Qwen 14B | 1/1 | 50.13 | +24.76 [+20.5, +29.2] | +11.39 | +7.48 | −5.89 [−9.1, −2.7] |
| Qwen 14B | 1.5/1 | 48.52 | +23.15 [+18.9, +27.6] | +9.78 | +5.80 | −7.57 |
| Qwen 14B | 1/0.5 | 46.17 | +20.80 [+16.8, +25.1] | +14.96 | +3.52 | −2.33 |
| Qwen 14B | 1/1.5 | 54.15 | +28.78 [+24.6, +33.1] | +6.48 | +11.50 | −10.80 |
| Qwen 14B | 1/2 | 58.72 | +33.35 [+28.9, +37.9] | +1.07 | +16.07 | −16.22 |
| Qwen 14B | 2/2 | 53.74 | +28.37 [+24.0, +32.8] | −3.91 | +11.96 | −20.33 |

**Integer-grid extension and uncertainty.** The extension reuses the original single-arm strengths 1/2/3 and joint cells (1,1), (1,2), (2,2), adding (1,3), (2,1), (2,3), (3,1), (3,2), (3,3) per model. The complete integer subset contains 64 cells and 294,400 responses, including 24 added cells and 110,400 responses. Terminal grid analysis: 36/36 joint cells available. Only complete 230-session manifests enter the native terminal analysis. Intervals use 10,000 paired whole-session draws (seed 42), with pointwise 95% and approximate family-of-108 bounds for three contrasts per joint cell/model. Extreme adjusted tails contain few replicates. No bootstrap is repeated for these tables, and no best-cell inference is performed.

Table 10: Complete integer-grid compliance (%), with models as columns. Each entry is 230 sessions × 20 turns, cap 256. All means are descriptive; marginal-CAV intervals appear in Table 12.

| Condition / metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Baseline | 21.54 | 31.07 | 25.37 | 37.24 |
| CAV(1) | 43.78 | 46.93 | 42.65 | 50.41 |
| CAV(2) | 45.15 | 48.15 | 41.78 | 50.65 |
| CAV(3) | 25.24 | 22.07 | 31.00 | 52.70 |
| Attention(1) | 40.91 | 46.20 | 38.74 | 53.15 |
| Attention(2) | 56.72 | 61.46 | 57.65 | 63.00 |
| Attention(3) | 59.50 | 70.87 | 72.91 | 61.89 |
| Combined(1,1) | 51.30 | 54.96 | 50.13 | 58.93 |
| Combined(1,2) | 57.24 | 64.00 | 58.72 | 61.48 |
| Combined(1,3) | 53.87 | 68.91 | 68.59 | 59.85 |
| Combined(2,1) | 52.33 | 53.61 | 46.98 | 59.65 |
| Combined(2,2) | 55.65 | 60.28 | 53.74 | 61.65 |
| Combined(2,3) | 51.76 | 66.52 | 63.91 | 60.28 |
| Combined(3,1) | 28.11 | 23.09 | 26.17 | 54.41 |
| Combined(3,2) | 23.65 | 20.85 | 18.54 | 53.26 |
| Combined(3,3) | 19.83 | 20.76 | 16.35 | 49.80 |

| Arm | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| CAV | 45.15 (2) | 48.15 (2) | 42.72 (1.5) | 52.70 (3) |
| Attention | 59.50 (3) | 70.87 (3) | 72.91 (3) | 63.00 (2) |
| Combined | 57.24 (1,2) | 68.91 (1,3) | 68.59 (1,3) | 61.65 (2,2) |

Table 11: Highest observed compliance (%) within each displayed arm family, with strength(s) in parentheses. Includes the original fractional doses as well as the integer extension. These selected means carry no winner-adjusted inference and are not performance ceilings.

| $(\lambda_C, \lambda_A)$ | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| | +10.39 | +8.76 | +11.39 | +5.78 |
| (1,1) | [+6.7, +14.1] | [+5.8, +11.8] | [+8.0, +14.7] | [+2.0, +9.5] |
| | +0.52 | +2.54 | +1.07 | −1.52 |
| (1,2) | [−2.9, +3.9] | [−0.043, +5.1] | [−1.3, +3.5] | [−5.1, +1.9] |
| | −5.63 | −1.96 | −4.33 | −2.04 |
| (1,3) | [−9.6, −1.7] | [−4.1, +0.3] | [−7.0, −1.8] | [−5.1, +0.9] |
| | +11.41 | +7.41 | +8.24 | +6.50 |
| (2,1) | [+7.7, +15.3] | [+4.3, +10.6] | [+5.1, +11.5] | [+2.4, +10.6] |
| | −1.07 | −1.17 | −3.91 | −1.35 |
| (2,2) | [−4.9, +2.8] | [−4.2, +1.9] | [−6.7, −1.2] | [−5.1, +2.3] |
| | −7.74 | −4.35 | −9.00 | −1.61 |
| (2,3) | [−12.2, −3.4] | [−7.2, −1.6] | [−12.2, −6.0] | [−4.6, +1.3] |
| | −12.80 | −23.11 | −12.57 | +1.26 |
| (3,1) | [−19.0, −6.6] | [−28.0, −18.3] | [−18.2, −6.9] | [−3.4, +5.9] |
| | −33.07 | −40.61 | −39.11 | −9.74 |
| (3,2) | [−39.3, −27.0] | [−45.9, −35.3] | [−45.3, −32.9] | [−14.1, −5.4] |
| | −39.67 | −50.11 | −56.57 | −12.09 |
| (3,3) | [−45.6, −33.7] | [−56.0, −44.1] | [−62.6, −50.5] | [−16.3, −7.9] |

Table 12: All integer-grid Combined$(\lambda_C, \lambda_A)$ minus Attention$(\lambda_A)$ contrasts (pp), with native pointwise whole-session 95% intervals. All 36 comparisons are complete. Figure 6 marks approximate family-108 exclusions of zero using the same terminal source.

| Condition / metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Combined(1,1) | 777 / 299 | 676 / 273 | 746 / 222 | 615 / 349 |
| Combined(1,2) | 468 / 444 | 472 / 355 | 406 / 357 | 393 / 463 |
| Combined(1,3) | 394 / 653 | 309 / 399 | 331 / 530 | 297 / 391 |
| Combined(2,1) | 854 / 329 | 687 / 346 | 659 / 280 | 673 / 374 |
| Combined(2,2) | 496 / 545 | 450 / 504 | 338 / 518 | 440 / 502 |
| Combined(2,3) | 419 / 775 | 322 / 522 | 287 / 701 | 322 / 396 |
| Combined(3,1) | 608 / 1197 | 297 / 1360 | 576 / 1154 | 649 / 591 |
| Combined(3,2) | 325 / 1846 | 180 / 2048 | 336 / 2135 | 366 / 814 |
| Combined(3,3) | 193 / 2018 | 187 / 2492 | 212 / 2814 | 281 / 837 |

Table 13: Gained / lost successes when adding CAV to Attention at the same attention strength. Gain: $A = 0, J = 1$; loss: $A = 1, J = 0$, across all 4,600 paired indices, not only baseline failures. Net change is $(\text{gain} - \text{loss})/4600 \times 100$ pp. Counts compare policy trajectories with different histories, not identical inputs.

**All contrasts, not only marginal CAV value.** Tables 14–15 show the native pointwise and family-adjusted intervals for $J - A$, $J - C$ and $I$ at every integer pair and model. This makes the full 108-comparison family visible, including negative and inconclusive outcomes.

Table 14: All three contrasts in the complete integer grid (pp), with native pointwise 95% intervals. $J - A$: marginal CAV value; $J - C$: marginal attention value; $I = J - C - A + B$: additive interaction. All 108 estimates and intervals are shown without selection.

| Model | $(C, A)$ | $J - A$ [95% CI] | $J - C$ [95% CI] | $I$ [95% CI] |
|---|---|---|---|---|
| Qwen 3B | (1,1) | +10.39 [+6.74, +14.07] | +7.52 [+5.17, +9.96] | -11.85 [-16.09, -7.96] |
| Qwen 3B | (1,2) | +0.52 [-2.91, +3.93] | +13.46 [+9.52, +17.48] | -21.72 [-26.87, -16.63] |
| Qwen 3B | (1,3) | -5.63 [-9.61, -1.74] | +10.09 [+5.24, +14.93] | -27.87 [-33.22, -22.57] |
| Qwen 3B | (2,1) | +11.41 [+7.67, +15.28] | +7.17 [+4.52, +9.91] | -12.20 [-16.67, -7.91] |
| Qwen 3B | (2,2) | -1.07 [-4.91, +2.76] | +10.50 [+6.39, +14.65] | -24.67 [-30.17, -19.30] |
| Qwen 3B | (2,3) | -7.74 [-12.17, -3.41] | +6.61 [+1.63, +11.59] | -31.35 [-36.98, -25.59] |
| Qwen 3B | (3,1) | -12.80 [-19.04, -6.61] | +2.87 [-0.22, +5.98] | -16.50 [-21.52, -11.67] |
| Qwen 3B | (3,2) | -33.07 [-39.30, -26.96] | -1.59 [-5.02, +1.83] | -36.76 [-43.07, -30.50] |
| Qwen 3B | (3,3) | -39.67 [-45.61, -33.65] | -5.41 [-8.70, -2.39] | -43.37 [-49.54, -37.24] |
| Qwen 7B | (1,1) | +8.76 [+5.85, +11.78] | +8.02 [+5.48, +10.80] | -7.11 [-10.52, -3.61] |
| Qwen 7B | (1,2) | +2.54 [-0.04, +5.07] | +17.07 [+13.41, +20.83] | -13.33 [-17.46, -9.35] |
| Qwen 7B | (1,3) | -1.96 [-4.11, +0.28] | +21.98 [+17.35, +26.61] | -17.83 [-22.09, -13.61] |
| Qwen 7B | (2,1) | +7.41 [+4.33, +10.61] | +5.46 [+3.11, +7.87] | -9.67 [-13.43, -6.17] |
| Qwen 7B | (2,2) | -1.17 [-4.24, +1.89] | +12.13 [+8.37, +16.04] | -18.26 [-22.72, -13.98] |
| Qwen 7B | (2,3) | -4.35 [-7.17, -1.63] | +18.37 [+13.57, +23.33] | -21.43 [-26.11, -16.80] |
| Qwen 7B | (3,1) | -23.11 [-27.96, -18.33] | +1.02 [-1.44, +3.61] | -14.11 [-18.22, -10.04] |
| Qwen 7B | (3,2) | -40.61 [-45.91, -35.35] | -1.22 [-3.46, +0.96] | -31.61 [-36.78, -26.52] |
| Qwen 7B | (3,3) | -50.11 [-56.00, -44.11] | -1.30 [-4.00, +1.33] | -41.11 [-47.26, -34.89] |
| Qwen 14B | (1,1) | +11.39 [+7.98, +14.74] | +7.48 [+5.04, +9.96] | -5.89 [-9.13, -2.74] |
| Qwen 14B | (1,2) | +1.07 [-1.33, +3.48] | +16.07 [+12.78, +19.43] | -16.22 [-20.22, -12.43] |
| Qwen 14B | (1,3) | -4.33 [-7.00, -1.78] | +25.93 [+20.85, +31.07] | -21.61 [-25.96, -17.30] |
| Qwen 14B | (2,1) | +8.24 [+5.06, +11.46] | +5.20 [+2.80, +7.70] | -8.17 [-11.30, -5.17] |
| Qwen 14B | (2,2) | -3.91 [-6.74, -1.15] | +11.96 [+8.80, +15.22] | -20.33 [-24.33, -16.37] |
| Qwen 14B | (2,3) | -9.00 [-12.20, -5.98] | +22.13 [+17.11, +27.13] | -25.41 [-30.00, -20.76] |
| Qwen 14B | (3,1) | -12.57 [-18.17, -6.91] | -4.83 [-8.33, -1.57] | -18.20 [-22.65, -13.96] |
| Qwen 14B | (3,2) | -39.11 [-45.28, -32.93] | -12.46 [-17.35, -7.67] | -44.74 [-51.52, -38.30] |
| Qwen 14B | (3,3) | -56.57 [-62.59, -50.46] | -14.65 [-19.70, -9.74] | -62.20 [-69.48, -55.11] |
| Llama 8B | (1,1) | +5.78 [+1.96, +9.48] | +8.52 [+5.09, +11.98] | -7.39 [-11.30, -3.46] |
| Llama 8B | (1,2) | -1.52 [-5.07, +1.93] | +11.07 [+6.67, +15.50] | -14.70 [-18.83, -10.65] |
| Llama 8B | (1,3) | -2.04 [-5.11, +0.91] | +9.43 [+4.83, +13.91] | -15.22 [-19.41, -11.17] |
| Llama 8B | (2,1) | +6.50 [+2.41, +10.61] | +9.00 [+5.74, +12.33] | -6.91 [-10.87, -2.96] |
| Llama 8B | (2,2) | -1.35 [-5.13, +2.30] | +11.00 [+6.52, +15.52] | -14.76 [-19.46, -10.09] |
| Llama 8B | (2,3) | -1.61 [-4.57, +1.26] | +9.63 [+4.98, +14.24] | -15.02 [-19.44, -10.61] |
| Llama 8B | (3,1) | +1.26 [-3.37, +5.89] | +1.72 [-1.74, +5.22] | -14.20 [-18.91, -9.46] |
| Llama 8B | (3,2) | -9.74 [-14.11, -5.39] | +0.57 [-3.26, +4.33] | -25.20 [-30.54, -20.02] |
| Llama 8B | (3,3) | -12.09 [-16.35, -7.91] | -2.89 [-7.13, +1.28] | -27.54 [-33.80, -21.48] |

Table 15: Approximate family-of-108 bounds for the same contrasts as Table 14. Bonferroni percentile tails contain few of the 10,000 stored draws; these are exploratory, not winner-adjusted intervals.

| Model | $(C, A)$ | $J - A$ family bounds | $J - C$ family bounds | $I$ family bounds |
|---|---|---|---|---|
| Qwen 3B | (1,1) | [+3.43, +17.57] | [+3.40, +11.67] | [-19.89, -5.51] |
| Qwen 3B | (1,2) | [-6.18, +6.55] | [+6.44, +20.31] | [-31.46, -13.16] |
| Qwen 3B | (1,3) | [-12.06, +1.13] | [+1.59, +18.64] | [-37.92, -19.04] |
| Qwen 3B | (2,1) | [+4.86, +17.98] | [+2.66, +11.67] | [-20.32, -4.69] |
| Qwen 3B | (2,2) | [-8.18, +5.56] | [+3.64, +17.68] | [-34.42, -15.55] |
| Qwen 3B | (2,3) | [-15.81, -0.09] | [-2.41, +15.36] | [-41.09, -21.26] |
| Qwen 3B | (3,1) | [-23.26, -1.66] | [-2.51, +8.85] | [-25.12, -8.11] |
| Qwen 3B | (3,2) | [-44.17, -22.24] | [-7.58, +4.37] | [-47.52, -25.89] |
| Qwen 3B | (3,3) | [-49.96, -29.05] | [-11.12, -0.56] | [-54.84, -32.45] |
| Qwen 7B | (1,1) | [+3.48, +13.86] | [+3.77, +13.00] | [-12.98, -1.01] |

Table 15 continued

| Model | $(C, A)$ | $J - A$ family bounds | $J - C$ family bounds | $I$ family bounds |
|---|---|---|---|---|
| Qwen 7B | (1,2) | [-2.14, +6.91] | [+10.79, +23.72] | [-20.84, -6.27] |
| Qwen 7B | (1,3) | [-5.99, +2.00] | [+14.29, +30.96] | [-25.46, -10.32] |
| Qwen 7B | (2,1) | [+1.69, +12.76] | [+1.37, +9.92] | [-16.20, -3.36] |
| Qwen 7B | (2,2) | [-6.78, +4.26] | [+6.11, +19.70] | [-26.15, -11.17] |
| Qwen 7B | (2,3) | [-9.73, +1.02] | [+10.44, +27.16] | [-29.86, -13.45] |
| Qwen 7B | (3,1) | [-31.53, -14.09] | [-3.35, +5.83] | [-21.12, -6.91] |
| Qwen 7B | (3,2) | [-49.54, -31.23] | [-5.18, +2.59] | [-41.20, -23.07] |
| Qwen 7B | (3,3) | [-60.49, -39.92] | [-5.97, +3.55] | [-51.94, -30.32] |
| Qwen 14B | (1,1) | [+5.42, +17.04] | [+3.43, +12.19] | [-12.07, -0.24] |
| Qwen 14B | (1,2) | [-2.93, +5.17] | [+10.70, +22.36] | [-23.66, -9.62] |
| Qwen 14B | (1,3) | [-8.92, -0.07] | [+16.01, +36.30] | [-29.57, -13.88] |
| Qwen 14B | (2,1) | [+2.27, +13.36] | [+1.06, +9.48] | [-14.40, -3.07] |
| Qwen 14B | (2,2) | [-8.75, +0.89] | [+5.81, +17.64] | [-26.79, -13.34] |
| Qwen 14B | (2,3) | [-14.48, -3.62] | [+11.21, +30.88] | [-34.20, -17.33] |
| Qwen 14B | (3,1) | [-22.89, -3.11] | [-10.81, +0.76] | [-26.97, -10.93] |
| Qwen 14B | (3,2) | [-49.55, -28.24] | [-20.35, -4.57] | [-57.99, -32.63] |
| Qwen 14B | (3,3) | [-67.40, -45.95] | [-23.46, -6.16] | [-74.46, -49.77] |
| Llama 8B | (1,1) | [-0.50, +11.81] | [+2.76, +15.15] | [-14.28, -1.06] |
| Llama 8B | (1,2) | [-7.84, +4.57] | [+3.52, +18.57] | [-21.79, -8.05] |
| Llama 8B | (1,3) | [-7.26, +2.97] | [+1.14, +17.22] | [-22.65, -8.28] |
| Llama 8B | (2,1) | [-0.60, +13.37] | [+2.97, +14.96] | [-14.29, +0.20] |
| Llama 8B | (2,2) | [-8.11, +5.86] | [+2.89, +18.80] | [-22.82, -6.90] |
| Llama 8B | (2,3) | [-6.86, +3.75] | [+1.61, +17.50] | [-23.54, -7.15] |
| Llama 8B | (3,1) | [-6.71, +9.35] | [-4.17, +8.49] | [-23.23, -5.33] |
| Llama 8B | (3,2) | [-17.70, -2.36] | [-6.23, +6.92] | [-34.61, -16.04] |
| Llama 8B | (3,3) | [-19.87, -4.09] | [-10.05, +4.45] | [-39.38, -17.40] |

**Complementarity without net advantage.** All 36 comparisons contain gained and lost successes; losses exceed gains in 24 (Figure 2). Thus occasional repair of attention failures is compatible with negative marginal value. These counts do not identify the mechanism of either outcome. At Combined(3,3), Qwen's large compliance losses also coincide with heavy repetition under the native word-token proxy (82.8/78.5/51.9% for 3B/7B/14B; exact values retained in the terminal analysis). Quality proxies remain descriptive, not an explanation of causation.

**Recovery and retention.** For each paired initial session and turn index, recovery is the fraction of baseline failures made compliant and retention the fraction of baseline successes that stay compliant (Table 16). These compare policy trajectories, not identical late-turn prompts.

Table 16: Rescue and harm at strength 1 (4,600 paired turn indices per model; denominators are baseline failures and baseline successes).

| Model | Arm | Rescued | Harmed | Recovery (%) | Retention (%) |
|---|---|---|---|---|---|
| Llama 8B | CAV 1.0 | 759/2887 | 153/1713 | 26.29 | 91.07 |
| Llama 8B | Attention 1.0 | 886/2887 | 154/1713 | 30.69 | 91.01 |
| Llama 8B | Combined 1/1 | 1147/2887 | 149/1713 | 39.73 | 91.30 |
| Llama 8B | Reference 1/1 | 1270/2887 | 115/1713 | 43.99 | 93.29 |
| Llama 8B | Random 1.0 | 290/2887 | 247/1713 | 10.05 | 85.58 |
| Llama 8B | Wrong 1.0 | 760/2887 | 282/1713 | 26.32 | 83.54 |
| Qwen 3B | CAV 1.0 | 1171/3609 | 148/991 | 32.45 | 85.07 |
| Qwen 3B | Attention 1.0 | 1019/3609 | 128/991 | 28.23 | 87.08 |
| Qwen 3B | Combined 1/1 | 1508/3609 | 139/991 | 41.78 | 85.97 |
| Qwen 3B | Reference 1/1 | 1747/3609 | 111/991 | 48.41 | 88.80 |
| Qwen 3B | Random 1.0 | 258/3609 | 263/991 | 7.15 | 73.46 |
| Qwen 3B | Wrong 1.0 | 967/3609 | 191/991 | 26.79 | 80.73 |
| Qwen 7B | CAV 1.0 | 943/3171 | 213/1429 | 29.74 | 85.09 |

Table 16 continued

| Model | Arm | Rescued | Harmed | Recovery (%) | Retention (%) |
|---|---|---|---|---|---|
| Qwen 7B | Attention 1.0 | 884/3171 | 188/1429 | 27.88 | 86.84 |
| Qwen 7B | Combined 1/1 | 1273/3171 | 174/1429 | 40.15 | 87.82 |
| Qwen 7B | Reference 1/1 | 1355/3171 | 189/1429 | 42.73 | 86.77 |
| Qwen 7B | Random 1.0 | 285/3171 | 256/1429 | 8.99 | 82.09 |
| Qwen 7B | Wrong 1.0 | 709/3171 | 305/1429 | 22.36 | 78.66 |
| Qwen 14B | CAV 1.0 | 942/3433 | 147/1167 | 27.44 | 87.40 |
| Qwen 14B | Attention 1.0 | 730/3433 | 115/1167 | 21.26 | 90.15 |
| Qwen 14B | Combined 1/1 | 1295/3433 | 156/1167 | 37.72 | 86.63 |
| Qwen 14B | Reference 1/1 | 1389/3433 | 123/1167 | 40.46 | 89.46 |
| Qwen 14B | Random 1.0 | 206/3433 | 190/1167 | 6.00 | 83.72 |
| Qwen 14B | Wrong 1.0 | 618/3433 | 184/1167 | 18.00 | 84.23 |

**Rescue overlap (all four models).** Table 17 decomposes baseline-failed turns by the single arms that rescue them and records losses and novel rescues under the joint policy. Table 18 reports the existing post-hoc analysis's paired whole-session percentile intervals (10,000 draws, seed 42), resampling ratio numerators and denominators together. All displayed ratios have defined denominators in every draw. These intervals are unadjusted and exploratory; they do not license a search across significant subgroups.

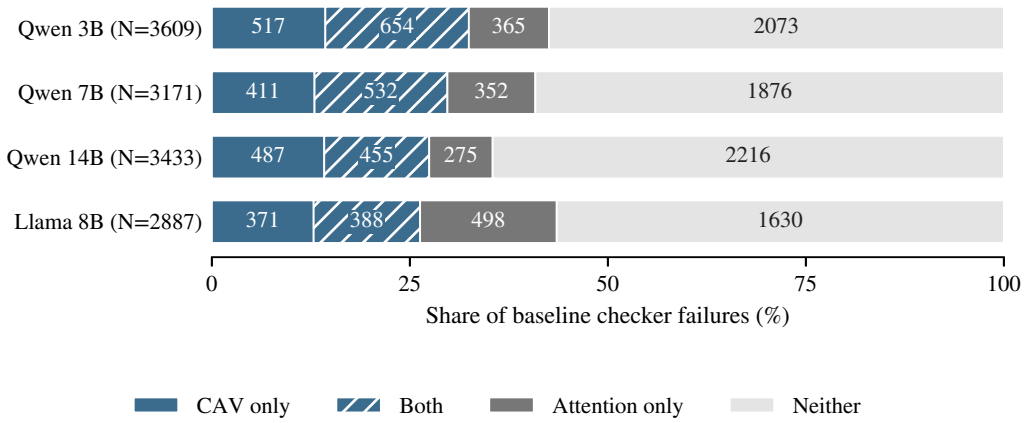


Figure 7: Unit-strength rescue sets among baseline checker failures. CAV(1) and Attention(1) rescue partly distinct sets: labels are raw counts, widths are percentages of each model's baseline failures, and $N$ is the denominator. Policies share initial sessions but generate different later histories. All-index gains and losses under the joint policy are reported separately in Figure 2.

| Metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Baseline failures | 3609 | 3171 | 3433 | 2887 |
| Both single arms rescue | 654 | 532 | 455 | 388 |
| Only CAV rescues | 517 | 411 | 487 | 371 |
| Only Attention rescues | 365 | 352 | 275 | 498 |
| Neither single arm rescues | 2073 | 1876 | 2216 | 1630 |
| Rescue union | 1536 | 1295 | 1217 | 1257 |
| Joint loses from union | 332 | 269 | 200 | 395 |
| Joint-only new rescues | 304 | 247 | 278 | 285 |
| CAV harms baseline successes | 148 | 213 | 147 | 153 |
| Attention harms baseline successes | 128 | 188 | 115 | 154 |
| Joint harms baseline successes | 139 | 174 | 156 | 149 |

Table 17: All-model rescue and harm counts at strength 1. Rescue groups are restricted to baseline failures; harm counts are restricted to baseline successes. Each model has 4,600 paired session–turn indices, but later histories differ across arms.

| Metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Common / union (%) | 42.6 | 41.1 | 37.4 | 30.9 |
| 95% interval | [36.7, 48.5] | [35.7, 46.4] | [31.8, 42.8] | [25.0, 37.0] |
| Joint loss / union (%) | 21.6 | 20.8 | 16.4 | 31.4 |
| 95% interval | [16.8, 26.9] | [16.5, 25.6] | [13.0, 20.4] | [24.8, 38.3] |
| Joint retains both-rescued (%) | 93.0 | 92.7 | 93.2 | 86.9 |
| 95% interval | [89.7, 95.6] | [89.9, 95.0] | [89.6, 96.1] | [78.8, 93.7] |
| Joint retains CAV-only (%) | 80.1 | 81.5 | 84.2 | 66.6 |
| 95% interval | [73.4, 85.7] | [75.5, 86.6] | [78.2, 89.0] | [54.8, 76.9] |
| Joint retains Attention-only (%) | 49.9 | 56.2 | 66.5 | 55.8 |
| 95% interval | [38.8, 61.4] | [46.0, 66.1] | [57.0, 74.6] | [44.7, 67.2] |
| Joint rescues neither-rescued (%) | 14.7 | 13.2 | 12.5 | 17.5 |
| 95% interval | [11.5, 18.3] | [10.2, 16.7] | [9.7, 15.7] | [13.6, 21.7] |

Table 18: Rescue overlap and joint retention (%) with existing, unadjusted whole-session 95% intervals. Only 31–43% of the single-arm rescue union is common to both, and the joint arm loses 16–31% of the union. Neither statistic identifies a shared mediator.

**Exact accounting, not mediation.** Let $O$ be the number rescued by both single arms, $U$ their rescue union, $L$ the cases in $U$ lost by the joint arm, $N$ joint-only rescues, and $H_C, H_A, H_J$ harms among baseline successes. The joint rescue count is $U - L + N$, and the interaction satisfies

$$I = 100\,\frac{-O - L + N + H_C + H_A - H_J}{4600} \quad \text{(pp)}. \tag{5}$$

Thus overlap is only one term; joint losses, novel rescues and changes in baseline-success harm also matter. This is an arithmetic decomposition of policy outcomes, not a causal mediation decomposition. Conditioning a descriptive independence benchmark on constraint type and turn gives expected common-rescue counts of 528.2/402.1/329.4/305.2 versus observed 654/532/455/388 (Qwen 3B/7B/14B, Llama). Each stratum has at most ten sessions; shared difficulty and temporal dependence remain, so the excess is not a mechanistic independence test.

## J Direction Controls

The wrong-constraint arm uses another cross-family CAV with the restoration amount computed from the target session's own CAV projection; it is a meaningful-direction control with a shared response cap, not a generic-noise control. Table 19 reports predefined checker-family subsets: format/case (130 sessions: change_case, punctuation, detectable_format, startend, combination), its complement (100 sessions: length_constraints, keywords, detectable_content) and case/punctuation (40 sessions, nested in format/case). The subset names describe checker families; they do not establish immunity to length or truncation. Mean generated tokens for CAV(1)/wrong(1) are 202.2/199.4 (Llama), 184.5/177.1

(Qwen 3B), 169.4/159.1 (7B) and 205.8/201.2 (14B), against baselines of 221.9/230.9/196.0/235.1, so neither arm improves compliance by lengthening outputs on average. The random-direction arm (+0.93/−0.11/+0.63/+0.35 pp) shows that a perturbation of matched form at the same layers does not by itself improve compliance.

| Subset (sessions) | Arm | Llama 8B | Qwen 3B | Qwen 7B | Qwen 14B |
|---|---|---|---|---|---|
| Format/case (130) | CAV 1 | +18.08 | +34.12 | +24.46 | +29.62 |
| Format/case (130) | Wrong 1 | +11.42 | +22.69 | +11.58 | +13.42 |
| Other (100) | CAV 1 | +6.80 | +6.80 | +4.70 | +1.25 |
| Other (100) | Wrong 1 | +9.05 | +9.30 | +5.15 | +4.25 |
| Case/punctuation (40) | CAV 1 | +17.62 | +35.12 | +27.62 | +33.25 |
| Case/punctuation (40) | Wrong 1 | +13.50 | +19.75 | +15.75 | +9.50 |

Table 19: Descriptive CAV(1) and wrong-CAV(1) gains (pp) for the predefined subsets, computed from the per-type means. No subset-specific significance or length-invariance claim is made.

## K Reference Policy

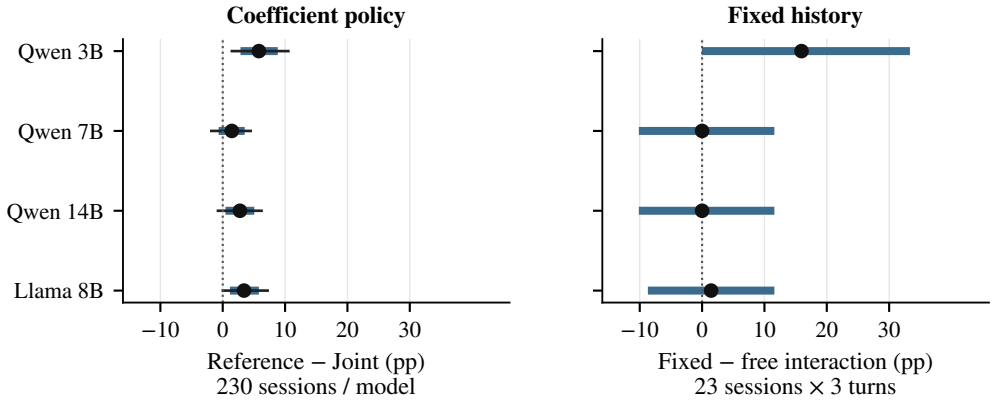


Figure 8: Two hypothesis-directed checks, with different estimands. Left: changing how the adaptive coefficient is computed improves the joint policy in some models; thin black lines are family-16 bounds and thick blue lines pointwise 95% intervals. Right: fixing history changes the interaction by an uncertain amount in every model (pointwise 95% intervals). A policy improvement does not identify the source of ordinary failures; an interval crossing zero does not establish no history effect.

The reference policy runs, at every generation step, an attention-free CAV-only forward pass on exactly the prefix of the main (attention-steered) pass, using a separate cache. At each CAV layer it computes the coefficient $\lambda_C(p_1 - c^\top h)$ from its own state; the main pass adds that coefficient times $c$ instead of computing it from its attention-perturbed state. The reference pass never selects a token. The coefficient still depends on the prefix, so this is not an input-independent dose. Reference-minus-control comparisons against the four scalar policies with attention 1 and CAV strength 0, 0.5, 1 and 1.5 form a family of 16 comparisons across models; Table 20 gives descriptive and Bonferroni-adjusted (0.15625/99.84375 percentile) intervals. Only Qwen 2.5 3B has all four adjusted lower bounds above zero, so superiority over this finite grid is supported for that model only, and other intervals do not establish equivalence or noninferiority. Table 21 reports lengths and the extra forward passes.

| Model | Reference 1/1 minus | Est. (pp) | 95% CI | family16 CI |
|---|---|---|---|---|
| Llama 8B | Attention 1 (CAV 0) | +9.20 | [+5.35, +12.98] | [+3.55, +14.89] |
| | Combined 0.5/1 | +2.89 | [+0.22, +5.59] | [−1.19, +7.09] |
| | Combined 1/1 | +3.41 | [+1.15, +5.80] | [+0.04, +7.20] |
| | Combined 1.5/1 | +1.43 | [−1.50, +4.33] | [−2.87, +5.77] |
| Qwen 3B | Attention 1 (CAV 0) | +16.20 | [+12.37, +20.07] | [+10.71, +21.80] |
| | Combined 0.5/1 | +6.41 | [+3.57, +9.30] | [+2.05, +10.73] |
| | Combined 1/1 | +5.80 | [+2.85, +8.83] | [+1.47, +10.52] |
| | Combined 1.5/1 | +5.46 | [+2.43, +8.63] | [+0.96, +10.14] |
| Qwen 7B | Attention 1 (CAV 0) | +10.22 | [+7.00, +13.48] | [+5.47, +15.05] |
| | Combined 0.5/1 | +2.80 | [+0.54, +5.07] | [−0.57, +6.23] |
| | Combined 1/1 | +1.46 | [−0.65, +3.52] | [−1.83, +4.49] |
| | Combined 1.5/1 | +2.41 | [−0.26, +5.11] | [−1.69, +6.61] |
| Qwen 14B | Attention 1 (CAV 0) | +14.15 | [+10.70, +17.61] | [+8.98, +19.45] |
| | Combined 0.5/1 | +3.80 | [+1.06, +6.52] | [−0.31, +7.91] |
| | Combined 1/1 | +2.76 | [+0.43, +5.07] | [−0.76, +6.21] |
| | Combined 1.5/1 | +4.37 | [+1.89, +7.02] | [+0.62, +8.43] |

Table 20: Reference(1,1) minus scalar controls (pp), with descriptive 95% and family-of-16 adjusted intervals.

| Model | Tokens (Comb./Ref.) | Cap (%) (Comb./Ref.) | Extra forwards | Ref. − Base (pp) [95% CI] |
|---|---|---|---|---|
| Llama 8B | 186.3/183.9 | 48.1/47.5 | 845,912 | +25.11 [+21.2, +29.2] |
| Qwen 3B | 174.7/165.4 | 48.0/45.6 | 760,655 | +35.57 [+30.7, +40.5] |
| Qwen 7B | 156.8/147.3 | 37.3/34.7 | 677,614 | +25.35 [+20.9, +29.9] |
| Qwen 14B | 192.4/184.4 | 52.6/49.8 | 848,307 | +27.52 [+23.2, +32.0] |

Table 21: Length, cap-hit rate and extra forward passes of the reference policy against Combined(1,1). Extra forwards are counted over all 4,600 responses per model.

**Which rescues change?** In the post-hoc group where baseline and CAV fail but Attention succeeds, Reference minus Combined is +24.9 pp [14.6, 35.2] for Qwen 3B; the corresponding estimates are +6.0/+9.5/+2.4pp for Qwen 7B/14B and Llama, respectively (Table 22). For Qwen 3B, Reference gains 98 cases and loses 7 relative to Combined in that 365-case group; both gains and losses are reported for every model. The group is defined by already-observed outcomes, and Reference generates its own history. These unadjusted exploratory contrasts do not establish a shared pathway or replace the adjusted full-bank comparisons above.

| Metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Reference − Joint, all (pp) | 5.8 | 1.5 | 2.8 | 3.4 |
| 95% interval | [2.8, 8.8] | [-0.7, 3.5] | [0.4, 5.1] | [1.2, 5.8] |
| Reference − Joint, Attn-only (pp) | 24.9 | 6.0 | 9.5 | 2.4 |
| 95% interval | [14.6, 35.2] | [-5.1, 16.8] | [2.5, 17.6] | [-6.7, 12.5] |
| Reference − Joint, CAV-only (pp) | 1.5 | -2.9 | -3.1 | 5.9 |
| 95% interval | [-8.5, 10.8] | [-8.5, 2.2] | [-10.6, 3.9] | [-1.1, 14.7] |
| Attn-only group size | 365 | 352 | 275 | 498 |
| Reference gains over Joint | 98 | 65 | 45 | 66 |
| Reference losses to Joint | 7 | 44 | 19 | 54 |

Table 22: Reference versus Combined in the full bank and existing single-arm-only rescue groups. Intervals are reused post-hoc whole-session 95% intervals, unadjusted. The final three rows give the Attention-only group size and the two directions of outcome change.

## L Does the Gain Come From the Current Turn, or the Trajectory?

Each free-history condition generates its own dialogue, so after turn 1 paired conditions answer different prompts. The fixed-history panel isolates the current turn: one outcome-independent, hash-selected session per constraint type (23 sessions) at turns 1, 10 and 20 gives 69 inputs per model. For each model, the four arms (baseline, CAV 1, attention 1, combined 1/1) respond to that model's own cap-256 baseline history; the 69 baseline outputs are reused and 207 responses per model are new. Interaction changes compare the fixed and free interactions on the same 23 sessions and three turns, resampling whole sessions (Table 23). All four change intervals include zero. The panel's free interaction also differs from the full-sample value (it is positive as a point estimate for Llama and Qwen 14B), which reflects the small panel. Fixing history therefore does not establish that divergent histories explain the negative interaction, and equivalence is not established either.

At the first generation step and first CAV layer, pre-intervention projections are exactly equal within the baseline/CAV and attention/combined pairs for 69/69 inputs each, since the prefixes are identical. The mean absolute attention-minus-baseline projection difference before the patch shrinks to a far smaller combined-minus-CAV difference after it (Table 24). This is a local scalar result at one step and one layer, not equality of whole representations or a mediated behavioral fraction; later tokens are not aligned.

| Model | Baseline | CAV(1) | Attention(1) | Combined(1,1) |
|---|---|---|---|---|
| Qwen 3B | 34.78 | 43.48 | 47.83 | 55.07 |
| Qwen 7B | 40.58 | 53.62 | 49.28 | 57.97 |
| Qwen 14B | 46.38 | 46.38 | 49.28 | 57.97 |
| Llama 8B | 59.42 | 65.22 | 57.97 | 69.57 |

| Model | Free-history $I$ | Fixed-history $I$ | Fixed − free |
|---|---|---|---|
| Qwen 3B | −17.39 [−30.4, −5.8] | −1.45 [−14.5, +10.1] | +15.94 [−0.04, +33.3] |
| Qwen 7B | −4.35 [−17.4, +7.2] | −4.35 [−13.0, +4.3] | +0.00 [−10.1, +11.6] |
| Qwen 14B | +8.70 [−4.3, +20.3] | +8.70 [+0.0, +18.8] | +0.00 [−10.1, +11.6] |
| Llama 8B | +4.35 [−8.7, +18.8] | +5.80 [+0.0, +13.0] | +1.45 [−8.7, +11.6] |

Table 23: Fixed-history panel (23 sessions, turns 1/10/20, 69 inputs per model). $I$ columns are interactions (pp) with paired whole-session 95% intervals over the 23 sessions; the last column is the paired fixed-minus-free change on the same panel.

| Model | First CAV layer | Before | After | Base=CAV before | Attn=Comb before |
|---|---|---|---|---|---|
| Llama 8B | 8 | 0.1888 | 0.0037 | 69/69 | 69/69 |
| Qwen 3B | 21 | 4.1472 | 0.0050 | 69/69 | 69/69 |
| Qwen 7B | 13 | 3.2413 | 0.0260 | 69/69 | 69/69 |
| Qwen 14B | 25 | 7.3773 | 0.0497 | 69/69 | 69/69 |

Table 24: First-step, first-CAV-layer projection differences on the fixed panel (mean absolute value; units are model-specific). Before: attention minus baseline before the CAV patch. After: combined minus CAV after it.

**Scope of the current-turn estimate.** A full-bank non-cascading CAV arm is outside the present evaluation. The existing fixed-history CAV 1 arm covers only 69 inputs per model (23 sessions at turns 1, 10 and 20), and no interval is recorded for its CAV-minus-baseline contrast. We do not estimate the fraction of the full-bank free-history CAV gain arising within the current turn.

## M Other Control Arms

**Instruction re-insertion.** Re-insertion prepends the turn-1 instruction to each later user message, with its own rollout and no activation intervention. Its four additional cap-256 cells contain 18,400 responses, separate from the main 108-cell matrix. Turn 1 outputs, scores and token counts match

baseline in all 230 sessions per model. Tables 25–26 import complete stored statistics and native paired whole-session intervals (10,000 draws, seed 42; unadjusted). Differences from baseline, CAV(1) and Combined(1,1) have pointwise intervals above zero in every model. Re-insertion changes prompt length, exposure and history; it is neither a matched internal-state intervention nor a performance ceiling. Indeed, Qwen's highest observed attention-only means exceed re-insertion's descriptive means.

| Metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Baseline CR (%) | 21.54 | 31.07 | 25.37 | 37.24 |
| CAV(1) CR (%) | 43.78 | 46.93 | 42.65 | 50.41 |
| Attention(1) CR (%) | 40.91 | 46.20 | 38.74 | 53.15 |
| Combined(1,1) CR (%) | 51.30 | 54.96 | 50.13 | 58.93 |
| Re-insertion CR (%) | 55.87 | 63.50 | 60.43 | 68.17 |
| Re-insertion − Baseline (pp) | +34.33 | +32.43 | +35.07 | +30.93 |
| Re-insertion − CAV(1) (pp) | +12.09 | +16.57 | +17.78 | +17.76 |
| Re-insertion − Combined (pp) | +4.57 | +8.54 | +10.30 | +9.24 |
| Re-insertion mean tokens | 175.6 | 154.8 | 199.8 | 205.3 |
| Re-insertion cap-hit (%) | 49.13 | 37.76 | 57.61 | 59.24 |

Table 25: Completed instruction re-insertion at cap 256, with unit-strength comparators. CR: checker pass rate. Each cell uses 230 sessions and 20 turns. Native paired intervals for the three registered contrasts appear in Table 26.

| Condition / metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Re-insertion − Baseline | +34.33 | +32.43 | +35.07 | +30.93 |
| 95% interval | [+30.00, +38.72] | [+27.96, +36.98] | [+30.76, +39.50] | [+26.85, +35.13] |
| Re-insertion − CAV(1) | +12.09 | +16.57 | +17.78 | +17.76 |
| 95% interval | [+8.13, +16.04] | [+11.93, +21.22] | [+13.37, +22.33] | [+12.98, +22.57] |
| Re-insertion − Combined(1,1) | +4.57 | +8.54 | +10.30 | +9.24 |
| 95% interval | [+0.50, +8.67] | [+4.39, +12.74] | [+6.11, +14.65] | [+4.85, +13.59] |

Table 26: Native re-insertion contrasts (pp) and paired whole-session 95% intervals, pointwise and unadjusted. These comparisons do not test re-insertion against a selected best dose.

**Rotation reversal and direction swaps.** Rotation-reversal and turn-specific direction-swap interventions were run earlier only for the Qwen models, with a 512-token response cap against 256-token baselines and a different attention implementation. Because budget and implementation differ from the present matrix, we report no results from them and make no claim about rotation reversal or direction swaps. Unit-normalized drift (Equation 1) measures orientation independently of norm; it does not establish that reversing rotation is necessary or sufficient for behavioral recovery.

**Site and dose selection.** The steering sites were chosen by compliance-maximizing sweeps on this same session bank under the earlier 512-token protocol (Appendix F). Directions, targets and attention calibration come from Phase 2. All of these are held fixed, and no held-out selection was performed. The sweeps ran at strength 1, so the strength-1 cells are the ones the site selection favours. The bank reuses previously studied instruction variants and is not an untouched confirmation set.

## N STATISTICAL TESTS FOR DRIFT AND DILUTION

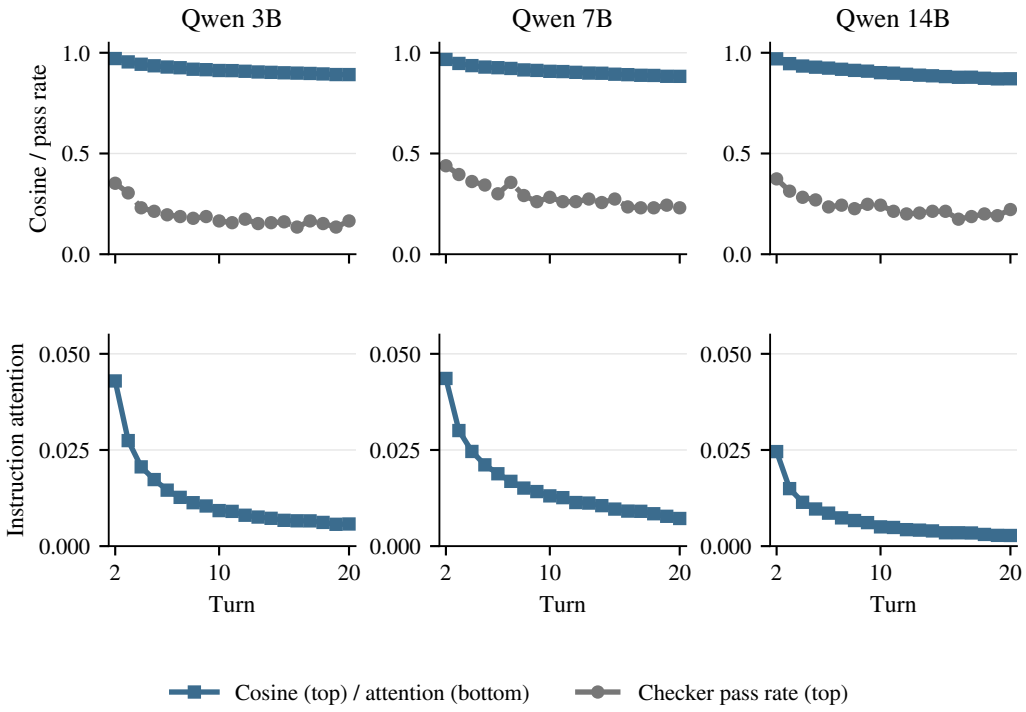


Figure 9: Historical Phase 2 trajectories for the three Qwen 2.5 scales. Top: cosine of per-constraint mean post-response residuals with the turn-1 mean (Equation 1), not cosine with the contrastive steering vector. Bottom: raw pre-response instruction-attention shares, averaged over sessions, layers and heads, on a common scale; these differ from the selected-head statistics in the text. Blue denotes cosine (top) or attention (bottom); gray dashed denotes checker pass rate (top only). No min–max normalization or reciprocal-turn reference is used.

Table 27 reports reference-layer correlations across the three Qwen scales. Each session's CAV projection and checker outcomes are correlated over turns 2–20, excluding constant-outcome sessions, and correlations are averaged via Fisher $z$. At 7B ($n = 143$ sessions), drift gives $\bar{\rho} = .248$ [.195,.299], 76.9% positive, Wilcoxon $p = 2.8 \times 10^{-13}$; attention gives .242 [.190,.292], 75.5% positive, $p = 3.1 \times 10^{-13}$. Pooled partial correlations over 4,370 turns are .208 ($p < 10^{-43}$) for drift given attention and .078 ($p = 2.4 \times 10^{-7}$) for attention given drift; repeated-turn dependence limits these pooled tests. Instruction attention falls from .100 to .022 at 7B between turns2 and20 (about4.5-fold). Historical Llama per-constraint-mean cosine reaches .76 at its reference layer30 (.70 in the deepest-drifting band), and attention falls5.6-fold. Its drift correlation is $\bar{\rho} = .054$ [−.001,.109], 56.8% positive, $p = .084$, $n = 118$, while attention gives .144 [.088,.200], 67.8% positive, $p = 8.2 \times 10^{-6}$. Mixed-timing and residual associations identify neither separate causal pathways nor prospective warning signals.

| Statistic | | 3B | 7B | 14B |
|---|---|---|---|---|
| Fisher $\bar{\rho}$ | Drift | .260 | .248 | .223 |
| | Dilut. | .240 | .242 | .224 |
| 95% CI | Drift | [.20,.32] | [.20,.30] | [.16,.29] |
| | Dilut. | [.19,.29] | [.19,.29] | [.16,.28] |
| % pos. sess. | Drift | 75.6 | 76.9 | 69.4 |
| | Dilut. | 74.0 | 75.5 | 71.0 |
| Wilcoxon $p$ | Drift | $<10^{-10}$ | $<10^{-12}$ | $<10^{-7}$ |
| | Dilut. | $<10^{-10}$ | $<10^{-12}$ | $<10^{-9}$ |
| Partial $\rho$ | Drift | .038 | .208 | .147 |
| | Dilut. | .132 | .078 | .125 |
| Partial $p$ | Drift | $<.02$ | $<10^{-43}$ | $<10^{-21}$ |
| | Dilut. | $<10^{-17}$ | $<10^{-6}$ | $<10^{-15}$ |

Table 27: Drift and dilution correlations with compliance (Qwen 2.5, reference layers). All Wilcoxon tests give $p < 0.05$. Partial correlations describe residual associations, not independent causal contributions.

## O TIMING-AWARE COSINE ANALYSIS

**Measurement boundary.** The archived Phase 2 extractor uses the first $2t$ alternating user/assistant messages, including assistant response $t$, then applies the chat template and caches the last-token residual. The probe and geometry share this cache. Per-cache token-prefix hashes are absent, so we conservatively treat these historical states as post-response; we cannot certify every tensor's prefix independently. Historical attention, by contrast, is indexed at the generation position before response $t$. Neither same-turn correlations nor a drop at the first failed turn is evidence of an early warning. The cap-256 fixed-history panel has verified identical input hashes and genuinely pre-response first-step traces, but only 69 inputs per model.

**Descriptive geometry and forecasting.** The completed secondary analysis uses 18,400 historical rows (four models, 230 sessions, 20 turns), keeping the cap-256 panel separate. Table 28 averages per-session $\cos(h_t, h_1)$; this differs from the cosine of per-constraint mean states in Equation 1. Mean angle is the mean of per-session angles, not the angle of the mean cosine. Norm changes are comparatively small, but neither this observation nor a directional/norm decomposition establishes causal contributions.

Next-failure prediction uses only previous-turn states $h_{t-1}$, with attention also lagged. Five session-disjoint folds fit cutoffs, scaling and smooth transformations on training sessions only. Context features are turn, constraint type, log context-character length and previous failure. Bootstrap intervals use 2,000 whole-session draws (seed 42) on out-of-fold predictions, conditional on fitted predictors rather than refitting them; they are exploratory and unadjusted across this broad analysis. The previously studied bank and historical site choices remain sources of selection.

| Condition / metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| T20 mean $\cos(h_t, h_1)$ | 0.846 | 0.816 | 0.816 | 0.584 |
| T20 mean angle (degrees) | 32.1 | 35.0 | 35.1 | 54.0 |
| T20 mean $\|h_t\|/\|h_1\|$ | 0.977 | 0.988 | 0.980 | 0.960 |
| CAV cutoff: next-failure BA | 0.606 | 0.633 | 0.609 | 0.609 |
| Context-only AUROC | 0.900 | 0.910 | 0.906 | 0.921 |
| Context+geometry+attention AUROC | 0.902 | 0.913 | 0.906 | 0.921 |
| Geometry gain: log loss | 0.001 | 0.002 | -0.001 | -0.003 |
| 95% interval | [-0.003, 0.005] | [-0.004, 0.006] | [-0.007, 0.005] | [-0.010, 0.003] |

Table 28: Retrospective geometry and lagged next-failure prediction. BA: balanced accuracy. The final row is incremental log-loss improvement from adding geometry to context+attention, with pointwise 95% intervals. Positive is better; every interval includes zero. No universal abrupt cosine threshold is established.

| Metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| Step minus smooth log loss | 0.010<br>[-0.010, 0.031] | 0.033<br>[0.019, 0.048] | 0.012<br>[-0.000, 0.024] | 0.001<br>[-0.005, 0.008] |
| First failure: rows / sessions | 426 / 127 | 668 / 153 | 543 / 142 | 982 / 127 |
| First failure: context / full AUROC | 0.778 / 0.742 | 0.789 / 0.774 | 0.719 / 0.713 | 0.763 / 0.771 |
| Recovery: rows / sessions | 3461 / 222 | 2999 / 218 | 3255 / 221 | 2743 / 196 |
| Recovery: context / full AUROC | 0.785 / 0.787 | 0.824 / 0.833 | 0.779 / 0.785 | 0.735 / 0.734 |
| Fixed-panel gap slope | 1.029<br>[0.093, 2.480] | 0.508<br>[-0.930, 1.952] | -0.305<br>[-1.576, 0.554] | 1.263<br>[0.390, 2.453] |
| Fixed-panel $J - A$ (pp) | 7.2<br>[-5.8, 17.4] | 8.7<br>[0.0, 17.4] | 8.7<br>[0.0, 17.4] | 11.6<br>[4.3, 20.3] |
| Within-session gap slope | 0.023<br>[-1.581, 1.710] | -2.742<br>[-5.040, -0.166] | -0.608<br>[-2.190, 0.839] | 0.917<br>[-0.763, 2.976] |
| High minus low gap benefit (pp) | 26.1<br>[4.3, 50.2] | 34.8<br>[11.4, 61.1] | 13.0<br>[-9.3, 35.8] | 0.0<br>[-26.1, 29.0] |

Table 29: Additional stored forecasting and fixed-panel sensitivity results. Prediction uses previous-turn states. Positive step-minus-smooth log loss favors smooth prediction. Gap slopes use model-specific scaled projection gaps, not comparable raw units. All intervals are exploratory, pointwise and conditional on the recorded analysis; within-session and high-minus-low analyses were added after primary inspection.

| Feature | Source model | Target Q3 | Target Q7 | Target Q14 | Target L8 | Held-out type BA |
|---|---|---|---|---|---|---|
| $\cos(h, h_1)$ | Qwen 3B | – | 0.590 | 0.517 | 0.516 | 0.533 |
| $\cos(h, h_1)$ | Qwen 7B | 0.521 | – | 0.508 | 0.519 | 0.603 |
| $\cos(h, h_1)$ | Qwen 14B | 0.557 | 0.547 | – | 0.514 | 0.495 |
| $\cos(h, h_1)$ | Llama 8B | 0.500 | 0.500 | 0.500 | – | 0.485 |
| $\cos(h, c)$ | Qwen 3B | – | 0.637 | 0.610 | 0.590 | 0.544 |
| $\cos(h, c)$ | Qwen 7B | 0.558 | – | 0.560 | 0.608 | 0.606 |
| $\cos(h, c)$ | Qwen 14B | 0.612 | 0.637 | – | 0.590 | 0.607 |
| $\cos(h, c)$ | Llama 8B | 0.550 | 0.642 | 0.549 | – | 0.596 |

Table 30: Stored next-failure cutoff transfer, balanced accuracy (BA; one-class reference .5). Off-diagonal cells transfer a source-model cutoff to another model on the same previously examined session bank, not unseen content. Final column: leave-one-constraint-type-out evaluation within the source model. Neither protocol supports a universal threshold; no new cutoff search was performed for this table.

**Limits and sensitivity.** A CAV-cosine cutoff has next-failure balanced accuracy only 0.61–0.63. Full geometry plus attention changes context-only AUROC by at most about 0.003. Constraint-specific alignment gives small pointwise improvements in Qwen3B/7B in an alternative feature set, but these do not establish an intervention mechanism or multiplicity-adjusted confirmation. First-failure prediction is not consistently improved. In the small fixed-history panel, high projection-gap groups have positive joint-minus-attention means, but different failure opportunities and constraint/turn composition. A post-hoc within-session sensitivity does not retain a consistent positive gap–benefit relation. We therefore make neither a universal threshold claim nor a claim that larger residual gaps reliably imply larger CAV benefits.

**Coverage.** Archived Qwen3B CAV directions at layers 28–35 are non-finite; corresponding direction metrics are missing, not zero. Its reference layer 21 and intervention layers 21–27 are finite, as are the other models' directions. No archived tensors were repaired. This cosine analysis uses the historical cache only; full four-arm hidden-vector differences are unavailable, so it cannot measure their alignment or identify shared mediation. The later pre-response extraction (Appendix W) supplies one fixed reference layer on the original dialogues for decoding; it does not recompute or change these cosine results.

## P PROBE DETERMINISM

The historical post-response linear probe reduces each activation to 256 principal components before fitting. The new pre-response analysis seeds both PCA and logistic regression in every model and uses its own fixed session folds (Appendix W). At $d_{\text{model}} \geq 2048$ with 256 components, scikit-learn's `PCA` selects its randomized SVD solver, which draws its random projection from the global NumPy generator unless a seed is supplied. Our implementation seeded the classifier but not the PCA, so repeated runs on identical activations returned slightly different components and therefore slightly different accuracies. The cross-validation folds were never a source of variation: `StratifiedGroupKFold` is used with `shuffle=False`.

Measured at 7B, reference layer 14, on identical inputs, the unseeded probe returned overall accuracies of 73.96/74.19/74.00% and turn-20 accuracies of 60.43/60.00/60.43% across three runs — a spread of roughly 0.25 pp overall and 0.4 pp at turn 20. With the PCA seeded, three runs return 74.14% and 60.87% exactly.

All historical Qwen probe numbers are from the seeded pipeline; historical Llama probe numbers come from the original run, whose PCA step was not seeded. Seeding shifts those reported values by less than the effect sizes they support: non-compliant-turn accuracy moves by at most 0.3 pp (3B 76.3→76.2, 7B 70.2→70.2, 14B 78.7→79.0), and the selected reference layer is unchanged at every scale (21/14/25). The largest single change is 14B turn-20 accuracy (67.8→69.6), increasing the measured retrospective decodability–compliance gap.

## Q NONLINEAR PROBE CONTROL

Linear decodability can understate information recoverable by a richer decoder. We therefore train an MLP probe (one 256-unit hidden layer) on the same post-response PCA-256 features, the same session-level 5-fold splits, and the same seed as the linear probe.

The nonlinear probe decodes consistently better (Table 31): +3.5 to +7.4 pp overall and, at turn 20, +3.5 (7B), +3.0 (14B) and +9.6 pp (3B). At non-compliant turns it reaches 74–83% against 70–79% for the linear probe (chance 4.3%). Two things follow. First, the retrospective gap is not an artefact of probe capacity: widening the probe class widens the gap, because compliance is unchanged while decodability rises. Second, a richer decoder recovers more information from these cached states. Neither result establishes what information was available before the response or functionally used by the model.

Part of the headroom is a regularization difference rather than nonlinearity: the linear probe is fit with $C = 0.01$, which is strong $L_2$ regularization, while the MLP uses scikit-learn's default $\alpha = 10^{-4}$. We tuned neither, and we make no claim about how the headroom divides between the two causes. The direction is what matters here, and it tells against reading our results as probe-capacity artefacts.

| | 3B | | 7B | | 14B | |
|---|---|---|---|---|---|---|
| **Accuracy (%)** | Lin. | MLP | Lin. | MLP | Lin. | MLP |
| Overall | 77.1 | 84.5 | 74.1 | 78.2 | 82.4 | 85.9 |
| Non-compliant turns | 76.2 | 83.3 | 70.2 | 74.4 | 79.0 | 83.1 |
| Turn 2 | 90.0 | 94.3 | 89.1 | 92.6 | 93.0 | 94.3 |
| Turn 20 | 65.7 | 75.2 | 60.9 | 64.3 | 69.6 | 72.6 |

Table 31: Historical post-response linear vs. MLP probes, sharing fold-local PCA-256 features, session-level folds and seed. The linear column agrees with Table 41 where metrics overlap. Probe capacity and regularization change together, so this is not a pure nonlinearity contrast. The MLP has one 256-unit hidden layer (Adam, 500 iterations). Neither column uses the new pre-response states.

## R DOES THE PROBE READ THE INSTRUCTION, OR THE TRANSCRIPT?

The historical probe reads the last-token residual of a growing dialogue *after including the current answer*. Its context contains the turn-1 instruction and every assistant response through that turn.

Compliant or partially compliant text can carry constraint-specific surface features. A probe might read those features rather than functional instruction access; this motivates transcript controls, not an assumption that high probe accuracy proves "knowing" before generation.

For each checker-failed turn we ask whether any earlier turn passed. A negative answer excludes prior checker-passing outputs, *not* partial constraint-related cues. Table 32 reports Qwen 2.5 14B at $\ell^* = 25$. This probe is refit on post-response PCA-256 features with scikit-learn's default classifier regularization and shuffled session folds, unlike the $C = .01$ historical fit in Table 41 and the new pre-response analysis. We compare subgroups within this fit, not absolute accuracies across pipelines.

| **Non-compliant turns (Qwen 2.5 14B)** | **Acc. (%)** | $n$ |
|---|---|---|
| All | 83.5 | 3,434 |
| No prior compliant turn in the session | 80.5 | 1,275 |
| Sessions that *never* complied | 79.1 | 1,160 |

Table 32: Probe accuracy at non-compliant turns by available residue (default-regularization, shuffled-fold refit). Chance is 4.35%. The last row restricts to the 58 sessions in which the model produced no compliant response at any of its 20 turns.

Accuracy is 80.5% without a prior checker pass and 79.1% in the 58 never-passing sessions, versus 83.5% across failures. These observational populations differ in constraint mix and difficulty, so the difference does not identify transcript contribution. Across three cross-validation seeds the unrestricted failed-turn accuracy spans 83.5–85.0% and the never-passing figure 79.1–82.7%, a within-fit difference of 2.3–4.4 pp. Partial cues and the initial instruction remain available in all populations.

The stratification is observational, so a completed transcript ablation rebuilds each cached context with assistant responses through the current turn replaced, then re-extracts residuals and refits the probe (same default-regularization, shuffled-fold pipeline). The archived ablation also uses a $2t$-message prefix. The turn-1 instruction stays unchanged. Two replacements probe different sources of decodability:

- **neutral** — replace each assistant response by one fixed paragraph, identical across sessions. Assistant-response text no longer varies with the target, but user questions and the initial instruction still vary.
- **swap** — the turn-matched response from a randomly chosen session in a different constraint family. This substitutes competing residue while preserving generated text. Excluding the target family makes donor sampling target-dependent; it is not an information-theoretic independence guarantee.

| | **Non-compliant** | | | **Never complied** | | |
|---|---|---|---|---|---|---|
| | orig. | neut. | swap | orig. | neut. | swap |
| Qwen 3B | 84.1 | **92.9** | 74.0 | 78.8 | **88.9** | 71.8 |
| Qwen 7B | 75.6 | **93.9** | 72.7 | 74.9 | **95.8** | 72.1 |
| Qwen 14B | 83.8 | **96.1** | 71.1 | 79.7 | **95.7** | 70.9 |
| Llama 3 8B | 81.2 | **90.1** | 64.1 | 77.2 | **88.3** | 60.8 |

Table 33: Post-response probe accuracy (%) with original assistant text, identical replacement assistant text, or another constraint's responses. User questions and the initial instruction are retained. All columns use default-regularization, shuffled-session-fold refits, not Table 41's pipeline. The 4,600 contexts per model form the full bank; displayed subgroup denominators are smaller. The uniform-label reference is 4.35%.

The ordering is **neutral** > **original** > **swap** for every model and both populations.

*Transcript changes affect decodability.* Replacing the original responses with a competing constraint's lowers non-compliant accuracy by 10.1/2.9/12.7/17.1 pp (Qwen 3B/7B/14B, Llama) and never-

complied accuracy by 7.1/2.8/8.8/16.5 pp. Because the replacement also changes the input distribution, these differences are not a bound on the amount of information supplied by the original transcript.

*Original assistant text is not necessary for high accuracy in the rebuilt contexts.* Under **swap**, failed-turn accuracy is 64–74%; under **neutral**, it is 90–96%. Neutralization removes target-dependent variation in assistant text, not every session difference: the initial instruction and user-question sequences remain. These synthetic, refitted contexts do not isolate the source of the original probe's signal or establish internal memory without the instruction.

*Neutral replacement increases decodability in this refitted comparison.* It raises failed-turn accuracy by 8.8–18.3 pp and never-passing accuracy by 10.1–20.9 pp. Replacement changes length, content and format together; it does not isolate interference or quantify helpful versus harmful residue.

Two scope notes. The instruction remains in context, and these are post-response readouts or synthetic replacements, not a test of memory without the instruction or functional access before the original answer. Random eligible donors avoid a fixed bijection, but do not guarantee independence. An earlier fixed rotation over the constraint list re-encoded target identity and raised 3B accuracy to 94.2%; those results are not used as evidence of residue removal.

## S Llama 3 Session Completion

Llama 3 8B has an 8K context window, and 47 of its 230 Phase 1 sessions initially exhausted memory on 24 GB GPUs. The session list is fixed in constraint-ID order before generation begins and the runner logs and skips memory failures. Sessions were not selected for recovery by observed compliance, but memory failure need not be statistically independent of response properties. Missing sessions were not uniform across constraint types ($\chi^2(22) = 85.6$, $p \approx 2 \times 10^{-9}$): they clustered early in the fixed run order, with `number_highlighted_sections` completing 0/10.

Rather than analyse a biased subset, we re-ran all 47 on 96 GB GPUs under identical settings (greedy decoding, 200 max new tokens) and recomputed the full Phase-2 pipeline. The Llama Phase 1–2 analyses therefore use the complete 230 sessions and a 23-way probe. Phase 3 separately generates all 230 Llama sessions in every condition at the common 256-token cap; its baseline failure count is not the probe-analysis denominator.

## T Distinguishing a Shared Mediator from the Intervention Rule

For a unit CAV $c$, the ideal $\lambda_C = 1$ patch is $T_1(h) = (I - cc^\top)h + p_1 c$. At one patched boundary,

$$T_1(h + \delta h) - T_1(h) = (I - cc^\top)\delta h. \tag{6}$$

Thus the patch removes the component of an incoming change along $c$ at that boundary; the first-step projections of Table 24 show this numerically. This local algebra does not measure mediation across the nonlinear, multilayer generation process. The reference policy (Appendix K) replaces the adaptive coefficient by one computed without attention steering and tests a different intervention rule; it does not isolate a natural indirect effect.

## U Complete Cap-256 Results and Provenance

The reported matrix contains 108 cells (4 models $\times$ 27 conditions), each with 230 sessions and 4,600 responses, totaling 496,800 response–condition records; no partial cell is pooled. It comprises 84 original cells within the displayed range plus 24 added joint cells. The original archive has 88 cells (404,800 records), including four out-of-range stress cells omitted under a common-strength reporting restriction chosen after outcomes were available. This restriction is not a preregistered selection rule; the original records remain archived. The extension adds 110,400 primary records. Tables 34–37 list the 21 displayed original conditions per model with baseline contrasts and quality proxies; the completed integer subset and gain/loss counts are in Appendix I. Four re-insertion cells and the fixed-history panel are counted separately.

Table 34: The 21 displayed original cap-256 conditions, Llama 3 8B. CR: checker pass rate (%). Δ: difference from baseline (pp) with paired whole-session 95% interval. Tokens, Cap (%), Rep-3g (%), Heavy (%), No-word (%), Harm and Rescue as in Appendix F. Combined and Reference strengths are CAV/attention. Added joint cells are reported separately.

| Condition | CR | Δ [95% CI] | Tokens | Cap | Rep-3g | Heavy | No-word | Harm | Rescue |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 37.24 | – | 221.9 | 74.4 | 6.7 | 0.1 | 0.0 | – | – |
| CAV 0.5 | 46.54 | +9.30 [+6.3, +12.3] | 209.3 | 61.6 | 6.1 | 0.0 | 0.0 | 173 | 601 |
| CAV 1.0 | 50.41 | +13.17 [+10.1, +16.4] | 202.2 | 55.6 | 6.1 | 0.2 | 0.0 | 153 | 759 |
| CAV 1.5 | 49.33 | +12.09 [+8.7, +15.6] | 198.3 | 53.3 | 6.5 | 0.2 | 0.0 | 199 | 755 |
| CAV 2.0 | 50.65 | +13.41 [+10.0, +16.9] | 195.8 | 51.4 | 7.0 | 0.7 | 0.0 | 184 | 801 |
| CAV 3.0 | 52.70 | +15.46 [+10.5, +20.5] | 205.4 | 64.1 | 28.1 | 25.9 | 3.7 | 357 | 1068 |
| Attention 0.5 | 42.98 | +5.74 [+3.4, +8.1] | 213.2 | 68.3 | 6.6 | 0.1 | 0.0 | 169 | 433 |
| Attention 1.0 | 53.15 | +15.91 [+12.5, +19.4] | 204.6 | 62.9 | 7.3 | 0.7 | 0.0 | 154 | 886 |
| Attention 1.5 | 58.33 | +21.09 [+17.0, +25.2] | 189.5 | 54.8 | 10.2 | 4.9 | 0.4 | 182 | 1152 |
| Attention 2.0 | 63.00 | +25.76 [+21.6, +30.0] | 175.1 | 49.9 | 13.8 | 9.4 | 1.6 | 145 | 1330 |
| Attention 3.0 | 61.89 | +24.65 [+20.1, +29.4] | 171.5 | 52.9 | 27.3 | 26.0 | 3.4 | 193 | 1327 |
| Combined 0.5/1.0 | 59.46 | +22.22 [+18.4, +26.1] | 191.3 | 50.5 | 6.2 | 0.6 | 0.0 | 123 | 1145 |
| Combined 1.0/1.0 | 58.93 | +21.70 [+17.8, +25.6] | 186.3 | 48.1 | 6.8 | 1.2 | 0.0 | 149 | 1147 |
| Combined 1.5/1.0 | 60.91 | +23.67 [+19.5, +27.8] | 182.8 | 46.0 | 7.1 | 0.9 | 0.0 | 156 | 1245 |
| Combined 1.0/2.0 | 61.48 | +24.24 [+19.4, +29.1] | 171.7 | 46.2 | 15.4 | 10.6 | 2.3 | 220 | 1335 |
| Combined 2.0/2.0 | 61.65 | +24.41 [+20.0, +28.9] | 173.6 | 49.6 | 18.6 | 13.6 | 1.9 | 185 | 1308 |
| Combined 1.0/0.5 | 55.78 | +18.54 [+14.8, +22.4] | 195.0 | 51.8 | 6.1 | 0.3 | 0.0 | 151 | 1004 |
| Combined 1.0/1.5 | 60.57 | +23.33 [+18.8, +27.9] | 179.7 | 47.0 | 9.6 | 4.0 | 1.2 | 202 | 1275 |
| Random 1.0 | 38.17 | +0.93 [−1.0, +2.8] | 217.7 | 70.0 | 7.8 | 0.3 | 0.0 | 247 | 290 |
| Wrong 1.0 | 47.63 | +10.39 [+6.6, +14.2] | 199.4 | 53.1 | 6.8 | 1.0 | 0.0 | 282 | 760 |
| Reference 1.0/1.0 | 62.35 | +25.11 [+21.2, +29.2] | 183.9 | 47.5 | 6.7 | 0.9 | 0.0 | 115 | 1270 |

Table 35: The 21 displayed original cap-256 conditions, Qwen2.5 3B (FP16). Columns as in Table 34.

| Condition | CR | Δ [95% CI] | Tokens | Cap | Rep-3g | Heavy | No-word | Harm | Rescue |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 21.54 | – | 230.9 | 80.2 | 2.9 | 0.3 | 0.0 | – | – |
| CAV 0.5 | 41.46 | +19.91 [+16.0, +23.9] | 189.4 | 55.8 | 3.9 | 1.5 | 0.3 | 116 | 1032 |
| CAV 1.0 | 43.78 | +22.24 [+18.0, +26.6] | 184.5 | 53.1 | 6.1 | 3.3 | 1.5 | 148 | 1171 |
| CAV 1.5 | 44.28 | +22.74 [+18.4, +27.1] | 185.5 | 53.0 | 7.6 | 4.9 | 3.1 | 151 | 1197 |
| CAV 2.0 | 45.15 | +23.61 [+19.1, +28.1] | 180.5 | 51.8 | 8.1 | 5.1 | 3.7 | 168 | 1254 |
| CAV 3.0 | 25.24 | +3.70 [−1.8, +9.1] | 231.3 | 86.7 | 67.9 | 71.4 | 4.5 | 564 | 734 |
| Attention 0.5 | 29.22 | +7.67 [+5.5, +10.0] | 216.8 | 71.8 | 2.8 | 0.5 | 0.0 | 127 | 480 |
| Attention 1.0 | 40.91 | +19.37 [+15.8, +23.1] | 200.3 | 63.3 | 3.2 | 1.2 | 0.0 | 128 | 1019 |
| Attention 1.5 | 50.96 | +29.41 [+25.2, +33.7] | 190.5 | 57.7 | 4.7 | 2.7 | 0.1 | 93 | 1446 |
| Attention 2.0 | 56.72 | +35.17 [+30.4, +39.8] | 183.2 | 54.8 | 8.9 | 7.0 | 0.0 | 116 | 1734 |
| Attention 3.0 | 59.50 | +37.96 [+33.0, +42.8] | 190.7 | 62.7 | 23.8 | 23.8 | 0.3 | 150 | 1896 |
| Combined 0.5/1.0 | 50.70 | +29.15 [+24.7, +33.7] | 174.8 | 48.7 | 5.6 | 3.4 | 0.4 | 114 | 1455 |
| Combined 1.0/1.0 | 51.30 | +29.76 [+25.2, +34.5] | 174.7 | 48.0 | 7.5 | 5.2 | 0.9 | 139 | 1508 |
| Combined 1.5/1.0 | 51.65 | +30.11 [+25.3, +34.9] | 176.7 | 48.3 | 9.4 | 6.8 | 1.9 | 140 | 1525 |
| Combined 1.0/2.0 | 57.24 | +35.70 [+30.7, +40.5] | 173.8 | 51.3 | 14.7 | 13.2 | 0.8 | 147 | 1789 |
| Combined 2.0/2.0 | 55.65 | +34.11 [+29.2, +38.9] | 182.2 | 56.1 | 18.2 | 16.3 | 3.0 | 173 | 1742 |
| Combined 1.0/0.5 | 47.35 | +25.80 [+21.3, +30.4] | 177.3 | 49.1 | 6.4 | 3.7 | 1.2 | 146 | 1333 |
| Combined 1.0/1.5 | 55.33 | +33.78 [+29.0, +38.5] | 173.7 | 48.2 | 9.8 | 7.1 | 1.6 | 136 | 1690 |
| Random 1.0 | 21.43 | −0.11 [−2.3, +2.2] | 230.1 | 80.1 | 6.7 | 3.6 | 0.0 | 263 | 258 |
| Wrong 1.0 | 38.41 | +16.87 [+12.8, +21.0] | 177.1 | 50.0 | 10.0 | 7.9 | 0.3 | 191 | 967 |
| Reference 1.0/1.0 | 57.11 | +35.57 [+30.7, +40.5] | 165.4 | 45.6 | 8.0 | 6.2 | 0.9 | 111 | 1747 |

Table 36: The 21 displayed original cap-256 conditions, Qwen2.5 7B. Columns as in Table 34.

| Condition | CR | Δ [95% CI] | Tokens | Cap | Rep-3g | Heavy | No-word | Harm | Rescue |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 31.07 | – | 196.0 | 62.3 | 3.0 | 0.4 | 0.0 | – | – |
| CAV 0.5 | 44.17 | +13.11 [+9.8, +16.3] | 174.7 | 48.1 | 3.1 | 0.5 | 0.3 | 178 | 781 |
| CAV 1.0 | 46.93 | +15.87 [+12.0, +19.9] | 169.4 | 45.0 | 3.2 | 0.5 | 0.1 | 213 | 943 |
| CAV 1.5 | 47.96 | +16.89 [+12.9, +20.9] | 168.5 | 44.7 | 3.7 | 1.0 | 0.2 | 232 | 1009 |
| CAV 2.0 | 48.15 | +17.09 [+13.2, +21.0] | 169.7 | 45.5 | 3.7 | 1.3 | 0.2 | 227 | 1013 |
| CAV 3.0 | 22.07 | −9.00 [−13.7, −4.2] | 239.6 | 89.3 | 65.9 | 71.3 | 2.2 | 890 | 476 |
| Attention 0.5 | 35.61 | +4.54 [+2.7, +6.4] | 189.2 | 58.0 | 2.9 | 0.3 | 0.2 | 209 | 418 |
| Attention 1.0 | 46.20 | +15.13 [+12.0, +18.4] | 176.4 | 50.3 | 3.0 | 0.4 | 0.3 | 188 | 884 |
| Attention 1.5 | 54.33 | +23.26 [+19.3, +27.3] | 167.4 | 44.0 | 2.8 | 0.5 | 0.3 | 191 | 1261 |
| Attention 2.0 | 61.46 | +30.39 [+25.9, +34.9] | 157.0 | 37.1 | 3.5 | 0.8 | 0.1 | 190 | 1588 |
| Attention 3.0 | 70.87 | +39.80 [+34.6, +45.0] | 135.3 | 27.3 | 4.8 | 1.5 | 0.3 | 200 | 2031 |
| Combined 0.5/1.0 | 53.61 | +22.54 [+18.4, +26.7] | 159.5 | 39.0 | 3.1 | 0.5 | 0.3 | 185 | 1222 |
| Combined 1.0/1.0 | 54.96 | +23.89 [+19.7, +28.2] | 156.8 | 37.3 | 3.2 | 0.8 | 0.0 | 174 | 1273 |
| Combined 1.5/1.0 | 54.00 | +22.93 [+18.7, +27.2] | 159.4 | 38.8 | 3.5 | 0.9 | 0.2 | 213 | 1268 |
| Combined 1.0/2.0 | 64.00 | +32.93 [+28.2, +37.7] | 145.7 | 31.5 | 3.2 | 0.5 | 0.0 | 188 | 1703 |
| Combined 2.0/2.0 | 60.28 | +29.22 [+24.5, +33.9] | 149.5 | 32.1 | 4.5 | 1.2 | 0.5 | 216 | 1560 |
| Combined 1.0/0.5 | 49.02 | +17.96 [+14.0, +22.0] | 165.2 | 42.2 | 3.2 | 0.7 | 0.3 | 218 | 1044 |
| Combined 1.0/1.5 | 58.20 | +27.13 [+22.8, +31.5] | 152.0 | 35.2 | 3.0 | 0.4 | 0.0 | 179 | 1427 |
| Random 1.0 | 31.70 | +0.63 [−1.2, +2.4] | 198.4 | 62.6 | 4.6 | 1.2 | 0.0 | 256 | 285 |
| Wrong 1.0 | 39.85 | +8.78 [+5.3, +12.4] | 159.1 | 39.2 | 4.3 | 2.0 | 0.3 | 305 | 709 |
| Reference 1.0/1.0 | 56.41 | +25.35 [+20.9, +29.9] | 147.3 | 34.7 | 3.0 | 0.7 | 0.0 | 189 | 1355 |

Table 37: The 21 displayed original cap-256 conditions, Qwen2.5 14B. Columns as in Table 34.

| Condition | CR | Δ [95% CI] | Tokens | Cap | Rep-3g | Heavy | No-word | Harm | Rescue |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 25.37 | – | 235.1 | 80.1 | 1.5 | 0.0 | 0.0 | – | – |
| CAV 0.5 | 40.98 | +15.61 [+12.1, +19.3] | 211.1 | 61.4 | 1.4 | 0.0 | 0.0 | 131 | 849 |
| CAV 1.0 | 42.65 | +17.28 [+13.4, +21.4] | 205.8 | 59.4 | 1.5 | 0.0 | 0.0 | 147 | 942 |
| CAV 1.5 | 42.72 | +17.35 [+13.4, +21.5] | 203.5 | 58.3 | 1.5 | 0.1 | 0.0 | 180 | 978 |
| CAV 2.0 | 41.78 | +16.41 [+12.2, +20.8] | 202.6 | 58.5 | 2.0 | 0.7 | 0.0 | 210 | 965 |
| CAV 3.0 | 31.00 | +5.63 [−0.02, +11.3] | 246.0 | 91.9 | 42.5 | 47.4 | 5.4 | 600 | 859 |
| Attention 0.5 | 31.22 | +5.85 [+4.0, +7.9] | 226.3 | 73.2 | 1.5 | 0.0 | 0.0 | 125 | 394 |
| Attention 1.0 | 38.74 | +13.37 [+10.8, +16.2] | 213.0 | 63.3 | 1.4 | 0.0 | 0.0 | 115 | 730 |
| Attention 1.5 | 47.67 | +22.30 [+18.8, +26.0] | 200.0 | 55.7 | 1.4 | 0.1 | 0.0 | 111 | 1137 |
| Attention 2.0 | 57.65 | +32.28 [+28.3, +36.4] | 188.4 | 48.0 | 1.7 | 0.1 | 0.0 | 90 | 1575 |
| Attention 3.0 | 72.91 | +47.54 [+42.7, +52.5] | 176.6 | 40.3 | 4.3 | 1.6 | 0.0 | 141 | 2328 |
| Combined 0.5/1.0 | 49.09 | +23.72 [+19.8, +27.9] | 195.8 | 53.3 | 1.5 | 0.0 | 0.0 | 117 | 1208 |
| Combined 1.0/1.0 | 50.13 | +24.76 [+20.5, +29.2] | 192.4 | 52.6 | 1.6 | 0.1 | 0.0 | 156 | 1295 |
| Combined 1.5/1.0 | 48.52 | +23.15 [+18.9, +27.6] | 193.3 | 52.5 | 1.6 | 0.0 | 0.0 | 159 | 1224 |
| Combined 1.0/2.0 | 58.72 | +33.35 [+28.9, +37.9] | 182.3 | 45.1 | 2.0 | 0.1 | 0.0 | 130 | 1664 |
| Combined 2.0/2.0 | 53.74 | +28.37 [+24.0, +32.8] | 187.1 | 47.4 | 2.9 | 1.0 | 0.1 | 178 | 1483 |
| Combined 1.0/0.5 | 46.17 | +20.80 [+16.8, +25.1] | 200.3 | 56.3 | 1.5 | 0.0 | 0.0 | 146 | 1103 |
| Combined 1.0/1.5 | 54.15 | +28.78 [+24.6, +33.1] | 185.7 | 48.5 | 1.7 | 0.0 | 0.0 | 130 | 1454 |
| Random 1.0 | 25.72 | +0.35 [−1.3, +2.0] | 234.0 | 78.1 | 1.6 | 0.0 | 0.0 | 190 | 206 |
| Wrong 1.0 | 34.80 | +9.43 [+6.3, +12.6] | 201.2 | 56.5 | 1.6 | 0.2 | 0.0 | 184 | 618 |
| Reference 1.0/1.0 | 52.89 | +27.52 [+23.2, +32.0] | 184.4 | 49.8 | 1.9 | 0.3 | 0.0 | 123 | 1389 |

Table 38: All 24 added cap-256 joint cells. CR (checker pass rate), Cap, Rep-U, Heavy-U and No-word are percentages; Tokens is mean generated length. Rep-U uses Unicode word tokens; Heavy-U additionally requires at least ten words; No-word means zero word tokens. These differ from the whitespace proxies in the original-cell tables and must not be pooled. Harm/rescue are counts among baseline successes/failures. All cells contain 4,600 outcomes.

| Model | $(C, A)$ | CR | Tokens | Cap | Rep-U | Heavy-U | No-word | Harm | Rescue |
|---|---|---|---|---|---|---|---|---|---|
| Qwen 3B | (1,3) | 53.87 | 192.5 | 63.76 | 29.13 | 29.50 | 0.63 | 243 | 1730 |
| Qwen 3B | (2,1) | 52.33 | 176.3 | 49.35 | 6.51 | 3.35 | 3.74 | 142 | 1558 |
| Qwen 3B | (2,3) | 51.76 | 196.8 | 66.59 | 30.61 | 32.24 | 2.22 | 295 | 1685 |
| Qwen 3B | (3,1) | 28.11 | 236.5 | 88.67 | 67.88 | 71.07 | 0.96 | 587 | 889 |
| Qwen 3B | (3,2) | 23.65 | 243.1 | 92.93 | 76.68 | 80.93 | 0.24 | 621 | 718 |
| Qwen 3B | (3,3) | 19.83 | 247.4 | 95.48 | 79.01 | 82.78 | 0.02 | 661 | 582 |
| Qwen 7B | (1,3) | 68.91 | 139.7 | 28.93 | 6.94 | 2.65 | 0.54 | 226 | 1967 |
| Qwen 7B | (2,1) | 53.61 | 161.5 | 39.85 | 4.43 | 1.57 | 0.04 | 208 | 1245 |
| Qwen 7B | (2,3) | 66.52 | 141.8 | 30.72 | 7.66 | 3.09 | 0.93 | 233 | 1864 |
| Qwen 7B | (3,1) | 23.09 | 238.3 | 89.28 | 68.37 | 73.28 | 2.41 | 853 | 486 |
| Qwen 7B | (3,2) | 20.85 | 238.8 | 89.96 | 70.27 | 75.02 | 3.74 | 886 | 416 |
| Qwen 7B | (3,3) | 20.76 | 242.1 | 92.26 | 73.38 | 78.46 | 3.13 | 873 | 399 |
| Qwen 14B | (1,3) | 68.59 | 183.2 | 44.41 | 6.22 | 2.37 | 0.26 | 232 | 2220 |
| Qwen 14B | (2,1) | 46.98 | 195.5 | 53.91 | 2.31 | 0.59 | 0.00 | 162 | 1156 |
| Qwen 14B | (2,3) | 63.91 | 189.4 | 48.20 | 7.28 | 3.24 | 0.26 | 271 | 2044 |
| Qwen 14B | (3,1) | 26.17 | 248.1 | 93.07 | 44.16 | 47.96 | 0.50 | 691 | 728 |
| Qwen 14B | (3,2) | 18.54 | 248.0 | 94.61 | 46.40 | 50.07 | 0.52 | 844 | 530 |
| Qwen 14B | (3,3) | 16.35 | 248.0 | 96.02 | 47.89 | 51.89 | 1.13 | 841 | 426 |
| Llama 8B | (1,3) | 59.85 | 169.2 | 52.96 | 30.53 | 28.70 | 1.80 | 216 | 1256 |
| Llama 8B | (2,1) | 59.65 | 179.6 | 43.87 | 8.40 | 1.72 | 0.02 | 155 | 1186 |
| Llama 8B | (2,3) | 60.28 | 166.1 | 51.43 | 31.86 | 31.15 | 2.00 | 182 | 1242 |
| Llama 8B | (3,1) | 54.41 | 198.5 | 64.33 | 35.20 | 32.17 | 2.98 | 352 | 1142 |
| Llama 8B | (3,2) | 53.26 | 185.3 | 62.15 | 40.08 | 38.30 | 3.61 | 373 | 1110 |
| Llama 8B | (3,3) | 49.80 | 184.0 | 64.83 | 50.21 | 52.87 | 2.13 | 392 | 970 |

# V LIMITATIONS

**Model and site scope.** The intervention comparison covers three Qwen 2.5 scales and Llama 3 8B, which differ in family, size, precision (float16 for Qwen2.5 3B) and calibration; this is not a factorial design over family and size. The Qwen sites were chosen by compliance sweeps (layer ranking, $N$ and $K$) at strength 1 on this same bank under the earlier 512-token protocol, and calibrations come from Phase 2. Llama's layers 8–14 were not swept and are exploratory, with $N$ and $K$ borrowed from Qwen2.5 7B. The reported effects are therefore in-sample with respect to site choice, and they are neither tuned optima for the 256-token cap nor bounds on other models.

**Exploratory status.** All cells use one previously examined bank of 230 sessions, greedy decoding and seed 42. Intervention intervals are pointwise and unadjusted except the reference family and approximate grid family. Pre-response intervals adjust only the four primary failed-turn accuracies; secondary subgroups are pointwise. Probe and cosine intervals condition on fitted cross-validated predictors and do not correct the full analysis search. Nothing here is independent confirmation.

**Compliance is not length-invariant.** Matching caps does not match output length. CAV and joint steering typically shorten outputs and change cap-hit rates. EOS timing, truncation, repetition and checker specifics can all change measured compliance; the predefined format/case subset is a sensitivity analysis, not a proof of length invariance. The quality proxies are descriptive and do not measure fluency or useful answers.

**Control arms.** Instruction re-insertion and its native paired inference are complete for all four models (Appendix M); it is not a performance ceiling. Rotation reversal, direction swaps and full-

bank non-cascading CAV are outside the current evaluation, not pending additions. The 69-input fixed-history panel gives only a small-sample current-turn estimate (Appendix L).

**Probe operationalization.** The response-excluded measurement establishes pre-response decodability on held-out sessions, not functional access, causal use, or memory without the visible instruction. Never-compliant histories can still contain partial constraint cues, and their membership uses outcomes from the whole session. No transcript-neutralization control was run on the new states. Historical nonlinear and transcript controls concern post-response features, with historical cache-prefix hashes absent (Appendix O). Differences between old and new probe accuracies are not a matched timing effect because the folds and pipelines differ.

**Dataset and measurement validity.** Fillers are unrelated content questions, not a demonstrated control for semantic interference. They may overstate decay relative to natural conversations. Persistence is not uniformly explicit in the prompts. The frozen checks are not official IFEval strict scoring and sometimes fail to verify the requested property (Table 4); constrained response is particularly a length proxy with unspecified choices. These defects affect the interpretation of aggregate effects and probe subgroups. No corrected-checker sensitivity has been run, so robustness to faithful rescoring is unknown. Ten variants per type also limit per-type inference.

**Mechanism coverage.** Post-response drift and pre-response attention associations do not identify independent natural causes. Rescue overlap and subadditivity do not identify a mediator; free-history comparisons are trajectory-level. The fixed-history panel is small (23 sessions per model), covering three turns. CAV helps weak attention in the tested policies but is not shown to outperform strong attention, and no universal cosine failure threshold is established.

## W Pre-Response Timing Verification

**Inputs and measurement.** We replay each saved Phase 1 dialogue only through its current user message, using `messages[:2*t-1]` and `add_generation_prompt=True`. Current and future answers are absent. Each of 920 sessions has 20 new states, for 18,400 primary forwards. Source-session, message-prefix, token-ID and state-file hashes are retained with turn, layer and checkpoint metadata. No answer or checker label is regenerated. The original caps remain Qwen256 and Llama200; these histories are distinct from the cap 256 intervention bank. We use the archived model wrapper, pinned checkpoint revisions and original precisions, eager attention, batch1 and no KV cache. At three fixed prefixes per model (turns 1/10/20 in the first sorted session), direct reference-layer hooks equal the wrapper readout bit-for-bit: all 12 comparisons pass. These are extraction checks, not new behavioral evidence.

**Fixed analysis.** One historical reference layer per model is fixed before this extraction: Qwen3B21, Qwen7B14, Qwen14B25 and Llama30 (zero-indexed decoder-block outputs). These sites came from prior analysis; freezing them for this check does not remove historical selection. Turn 1 is retained as an anchor but excluded from training and evaluation. Deterministic, constraint-balanced five-fold splits hold out two sessions per type in each fold (184 training/46 test sessions). PCA256 uses randomized SVD with seed 42 and fits only training rows; logistic regression fixes $C = .01$, `max_iter=1000` and seed 42. All 4,370 rows per model receive out-of-fold predictions; failed-turn populations are evaluated afterward without retraining. Input checks, split assignments and predictions are retained.

**Uncertainty and baselines.** We reuse the completed native analysis: 2,000 whole-session draws, seed 42, resampling eligible correct-count/row-count ratios. Table 1 uses the 0.625th and 99.375th percentiles for a Bonferroni family of four primary failed-turn accuracies. Table 39 gives pointwise 95% intervals. These intervals condition on fitted out-of-fold predictors and omit refitting uncertainty. The 4.35% line is a uniform random-label reference, not a conditional majority baseline or a permutation test. Training labels are balanced, so the training-majority rule breaks ties deterministically; its failed-turn accuracy is 2.85/3.39/2.99/0.71%. Secondary populations are descriptive, not independent confirmations.

| Metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| All turns | 61.53 [57.7, 65.4] | 76.41 [72.7, 80.2] | 72.33 [68.7, 76.2] | 56.86 [51.9, 61.9] |
| Compliant turns | 73.90 [67.5, 80.1] | 86.62 [82.2, 90.6] | 87.79 [82.9, 92.1] | 72.10 [64.7, 79.2] |
| Failed turns | 58.68 [54.8, 62.9] | 72.22 [67.7, 76.8] | 67.60 [63.5, 71.9] | 48.32 [43.0, 54.0] |
| Late failed turns (T10–20) | 52.04 [47.7, 56.8] | 67.58 [62.6, 72.7] | 61.74 [57.2, 66.7] | 45.36 [39.7, 51.1] |
| Failures, no prior success | 45.20 [39.1, 51.3] | 70.23 [60.8, 79.3] | 62.76 [56.1, 69.2] | 42.41 [34.9, 50.1] |
| Never-compliant sessions | 42.19 [35.5, 48.4] | 69.21 [59.2, 78.9] | 61.43 [54.3, 68.3] | 40.24 [32.5, 48.3] |
| Turn 20 | 44.78 [38.7, 51.3] | 61.74 [55.2, 68.3] | 53.91 [47.4, 60.4] | 44.35 [37.8, 50.9] |

Table 39: Pre-response accuracy (%) with pointwise 95% session-cluster intervals. The primary failed-turn family-adjusted intervals are in Table 1. Late failures use turns 10–20. No-prior-success failures condition only on earlier checker outcomes; never-compliant sessions condition on all 20 saved outcomes.

| Metric | Qwen 3B | Qwen 7B | Qwen 14B | Llama 8B |
|---|---|---|---|---|
| All turns | 4370 / 230 | 4370 / 230 | 4370 / 230 | 4370 / 230 |
| Compliant turns | 820 / 132 | 1271 / 153 | 1024 / 132 | 1570 / 149 |
| Failed turns | 3550 / 221 | 3099 / 220 | 3346 / 222 | 2800 / 199 |
| Late failed turns (T10–20) | 2135 / 216 | 1891 / 215 | 2010 / 216 | 1682 / 192 |
| Failures, no prior success | 1228 / 84 | 880 / 66 | 1187 / 81 | 1363 / 88 |
| Never-compliant sessions | 1140 / 60 | 760 / 40 | 1102 / 58 | 1235 / 65 |
| Turn 20 | 230 / 230 | 230 / 230 | 230 / 230 | 230 / 230 |

Table 40: Eligible turns / contributing sessions for each pre-response accuracy. Turn 1 is excluded from every probe estimate, but included when defining prior success or never-compliant sessions.

**Interpretation.** Constraint identity is recoverable before many failed answers, including before late failures and within never-compliant sessions. The original instruction is still present, and checker failure does not remove partial constraint cues from previous answers. The result excludes the current answer as the source of this probe's signal, but does not identify where the remaining information comes from or whether generation uses it. Neither the historical transcript-replacement controls nor the CAV intervention effects establish functional use of this particular pre-response readout. No new dose, cosine-threshold or mediator search accompanies this verification.

## X Historical Post-Response Decoding

The earlier per-layer linear probes use PCA256 and logistic regression ($C = .01$), with session-disjoint `StratifiedGroupKFold`. Unlike the new timing check, their cached inputs include the current response. Non-compliant-turn accuracies at the reference layer are 76.2/70.2/79.0/79.3% (Qwen3B/7B/14B, Llama); compliant-turn accuracy at Qwen7B is 83.9%. The historical Qwen14B estimate in 58 never-compliant sessions is 79.1%. These remain retrospective measurements. Folds, preprocessing and provenance differ from the pre-response analysis, so numerical differences between them are not a paired estimate of the effect of including an answer.

| Model | Compliance (%) T1 | T10 | T20 | Post-response probe (%) T2 | T20 |
|---|---|---|---|---|---|
| Qwen 3B | 55.2 | 16.5 | 16.5 | 90.0 | 65.7 |
| Qwen 7B | 66.5 | 28.3 | 23.0 | 89.1 | 60.9 |
| Qwen 14B | 61.7 | 24.3 | 22.2 | 93.0 | 69.6 |
| Llama 3 8B | 55.2 | 38.3 | 30.4 | 79.6 | 69.6 |

Table 41: Preserved Phase1–2 behavioral and historical post-response probe results, 230 sessions/model, Qwen cap 256 and Llama cap 200. These probe values are not the pre-response estimates of Table 1.

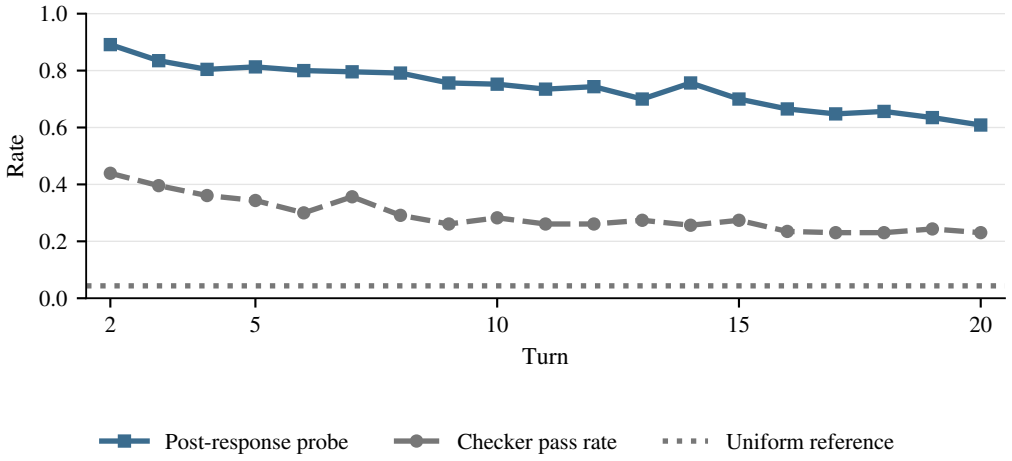


Figure 10: Preserved retrospective decodability–compliance trajectory for Qwen2.5 7B. The blue probe curve includes the current answer; gray denotes checker pass rate. This differs from the pre-response measurement in Figure 1.

## Y Evidence Reading Map

Table 42: Reading map: research question, complete evidence location and inference boundary. The map includes null, negative and historical controls; it is not a list of independently confirmed claims.

| Question | Evidence to read | What remains unresolved |
|---|---|---|
| What did we construct? | §4.1, Table 4: 23 types, 230 sessions, 119-question pool, custom checks. | Checker fidelity and uniformly explicit persistence; not official IFEval scoring. |
| What is decodable before failure? | Table 1, Figure 1, Tables 39–40: all four models, all seven populations and denominators. | Functional use; current instruction and prior cues remain. |
| Do historical controls change decoding? | Tables 41, 31, 32, 33; Appendices P, S. | Post-response timing and changed pipelines prevent a paired old/new timing effect. |
| Do ordinary trajectories explain failures? | Figure 9, Tables 27, 28, 29, 30. | Correlation, prediction and causal mechanism differ; no universal threshold. |
| Do individual policies work? | Tables 2, 3, 7, 8, 19; Figure 5. | Wrong directions also help; semantic quality and generalization unverified. |

Table 42 continued

| Question | Evidence to read | What remains unresolved |
|---|---|---|
| Do individual rescues combine? | Figures 2 and 6; Tables 9–13, 14–15. | 36 gain/loss comparisons and 108 contrasts are exploratory policy comparisons. |
| Which successes are gained or lost? | Figures 2 and 7, Tables 16, 17, 18. | Exact outcome accounting is not mediation. |
| Does within-response coupling matter? | Figure 8; Tables 20–22, 24. | Local cancellation verified, behavioral mediation not identified; extra forward cost. |
| Does history propagation matter? | Figure 8, Table 23; Appendix L. | 69 inputs/model; all fixed-minus-free interaction intervals cross zero. |
| Are gains uniform or useful? | Figure 4; Tables 6, 25–26, 38. | Type heterogeneity, repetition, truncation and checker-specific shortcuts. |
| Are the reported results complete? | Tables 34–37, 38; Appendix U. | 84 original + 24 extension cells; reinsertion/fixed panel separate; post-hoc range restriction disclosed. |