RELATIONAL STEERING WITH A FIXED DIRECTION: CAUSAL EFFECTS AND BEHAVIORAL LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a single fixed activation direction produce opposite but appropriate effects across contexts? We study this question with a 2×2 factorial design crossing epistemic alignment or distortion with favorable or unfavorable outcomes. In Qwen3-8B-Base and Llama-3.1-8B, we construct a relational direction from the factorial interaction contrast, which cancels the designed additive main effects, freeze it before held-out evaluation, and test causal writability through teacher-forced consequence preferences. The aggregate relational effect transfers to independent novel domains in both model families, although strict branchwise transfer is established only in Llama. Structured controls show that response magnitude alone does not explain the effect, and a discovery-selected best-of-1,024 Gaussian direction shows the branchwise sign pattern is not unique to the factorial direction, which nevertheless has a substantially larger held-out local response. Controlled-readout writability does not translate monotonically into open-ended behavior: under prompt-prefill-only intervention in Llama, the distortion branch is stronger under teacher forcing, whereas the alignment branch has the larger and more persistent trajectory-average behavioral effect, a pattern reproduced by an independent automated audit and, for the average effect, by blinded human calibration. Relation controls further constrain interpretation. Removing the stated causal linkage weakens steering in both model families, whereas explicit reversal does not produce the opposite joint pattern; a post-hoc baseline-conditioned diagnostic shows the Llama reversal response retains a default-oriented component not explained by simple amplification of the contemporaneous preference, while Qwen is substantially more baseline-tracking under the same manipulation. These results establish context-responsive relational writability while delimiting what it implies: the observed sign structure is not unique to the factorial direction, controlled writability does not guarantee persistent behavioral expression, and relation sensitivity does not by itself establish mechanism-sensitive causal-rule recomputation.

1 INTRODUCTION

Activation steering modifies model behavior by adding a direction to hidden activations at inference time (Turner et al., 2024; Zou et al., 2023; Rimsky et al., 2024), but effects can vary substantially across prompts (Tan et al., 2024), so recent methods make the intervention input-dependent or replace a single global vector with a context-dependent one (Lee et al., 2025; Li et al., 2026).

Relational properties pose a different problem: the desired effect can itself change sign with context. If accurate information is associated with favorable consequences and distorted information with unfavorable ones, strengthening this relation should increase preference for the favorable consequence under an aligned premise but decrease it under a distorted one. We ask whether one fixed activation direction, applied with the same sign, can produce these opposite effects without switching directions or gating the intervention by premise type.

We study this question with a 2×2 factorial design crossing epistemic alignment or distortion (in the task-specific sense of whether a scenario accurately represents its state, unrelated to AI alignment with human values) with favorable or unfavorable outcomes: the factorial interaction contrast defines a relational direction that cancels the additive main effects (Section 2), and relational writabil-

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

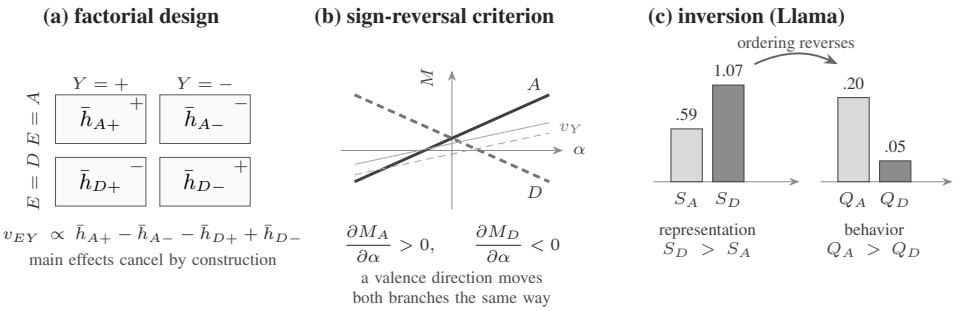


Figure 1: The construct, the criterion, and the central result. (a) Each domain crosses an epistemic factor with outcome valence; the relational direction is the interaction contrast, which cancels both main effects. (b) A relational direction must reverse sign with the premise, a signature an outcome-valence direction cannot produce however large its effect. (c) On Llama's independent novel domains the distortion branch is the more writable of the two, yet the alignment branch is the one whose effect persists in open-ended generation.

ity requires the favorable-minus-unfavorable teacher-forced preference to increase under aligned premises and decrease under distorted ones—a positive aggregate response alone is insufficient. Figure 1 previews this construction, the sign-reversal criterion, and our central Llama result.

This design also separates questions that can otherwise be conflated: a direction may produce the intended branchwise sign structure without being unique to that relation; a strong controlled readout effect may fail to persist during open-ended generation; and sensitivity to a stated relation does not imply a fixed direction recomputes its orientation from an arbitrary causal mechanism. We therefore evaluate relational writability, directional specificity, behavioral persistence, and mechanism sensitivity as distinct properties rather than treating successful steering as evidence for all four.

Using directions frozen before novel-domain evaluation, the aggregate relational effect transfers to independent domains in both Qwen3-8B-Base and Llama-3.1-8B, though branchwise transfer holds strictly only in Llama and specificity is comparative rather than absolute (Section 3). Teacher-forced writability, moreover, does not translate monotonically into open-ended behavior: under prefill-only intervention, Llama's branch ordering reverses between the controlled readout and open-ended generation (Section 4), and relation controls show that removing the stated causal linkage weakens steering in both model families while explicit reversal does not invert it (Section 5). These findings support a deliberately bounded interpretation: a fixed factorial direction can write a context-responsive relational effect that transfers beyond the discovery domains, without being the unique direction capable of producing it, without guaranteeing persistent behavioral expression, and without establishing that it recomputes an arbitrary stated causal mechanism.

Contributions.

- We introduce a factorial relational direction and a branchwise sign criterion for testing whether one fixed intervention can produce opposite context-dependent effects while cancelling the designed main effects.
- We demonstrate aggregate held-out relational writability in two 8B model families; branchwise transfer holds more strongly in Llama, and the sign pattern is not unique to the factorial direction, so specificity is comparative rather than absolute.
- We distinguish controlled-readout writability from open-ended behavioral expression: in Llama, a prefill-only trajectory audit reveals a representation–behavior branch inversion and different persistence profiles despite stronger teacher-forced writability in the distortion branch.
- We use relation controls and a post-hoc baseline-conditioned diagnostic to separate relation sensitivity from baseline amplification and mechanism-sensitive rule recomputation: removing the stated linkage weakens steering in both model families, while under explicit reversal Llama retains a baseline-independent component and Qwen is substantially more baseline-tracking.

2 THE RELATIONAL DIRECTION AND ITS CRITERION

Factorial construction. Each domain crosses epistemic status, $E \in \{A, D\}$, with outcome valence, $Y \in \{+, -\}$. Here, A denotes an accurate task-relevant representation and D a materially distorted one; this use of *alignment* is unrelated to AI alignment with human values. Within each domain, the same premise and consequence clauses are reused across the four cells. The discovery and novel sets each contain 24 disjoint domains (Appendix A).

For domain d and layer ℓ, let $\bar{h}^{(\ell)}_{d,e,y}$ denote the unperturbed hidden state mean-pooled over the consequence span. The relational interaction contrast is

$$d^{(\ell)}_{EY,d} = \tfrac{1}{4}\left(\bar{h}_{d,A,+} - \bar{h}_{d,A,-} - \bar{h}_{d,D,+} + \bar{h}_{d,D,-}\right). \tag{1}$$

This contrast cancels the designed additive epistemic and outcome main effects, isolating whether the favorable-minus-unfavorable difference changes with epistemic status. We average the raw domain contrasts and unit-normalize only after aggregation to obtain v_{EY}. The corresponding outcome and epistemic main-effect directions, v_Y and v_E, are constructed analogously and serve as structured controls; their exact definitions are given in Appendix B.

Intervention, readout, and branchwise criterion. At the selected layer ℓ^*, every prompt-token residual state receives the RMS-scaled perturbation $\alpha s_i v$; teacher-forced candidate tokens receive no direct intervention. Let L denote total candidate log-likelihood and define the common favorable-minus-unfavorable margin

$$M_{d,e}(\alpha) = L(c_{d,+} \mid x_{d,e}, \alpha, v) - L(c_{d,-} \mid x_{d,e}, \alpha, v). \tag{2}$$

The same margin is used in both branches: a positive value favors the favorable consequence regardless of epistemic status.

Over $\alpha \in \{-1, -0.5, 0, 0.5, 1\}$, we fit a linear dose response with slope $\beta_{d,e}$ and orient the branch effects as

$$S_{A,d} = \beta_{d,A}, \qquad S_{D,d} = -\beta_{d,D}. \tag{3}$$

Positive values therefore indicate movement in the relation-predicted direction in both branches. Strict relational writability requires

$$\mathbb{E}_d[S_{A,d}] > 0, \qquad \mathbb{E}_d[S_{D,d}] > 0. \tag{4}$$

We summarize aggregate relational magnitude by $I_d = S_{A,d} + S_{D,d}$ and branch imbalance by $G_d = S_{A,d} - S_{D,d}$. No sign of G is required, and a positive I alone is insufficient if either branch has the wrong sign.

The semantic domain is the inferential unit. Claims requiring the predicted effect in both branches use a strict intersection–union criterion: both one-sided exact sign-flip tests must pass at $p < 0.05$; branch imbalance is tested directly through G.

Layer localization and freezing. Because geometric extractability need not imply causal leverage, we localize the intervention site by branch-balanced small-dose writability on discovery domains only. Using leave-one-domain-out discovery axes and symmetric interventions at $\epsilon = 0.10$, we select the layer maximizing the weaker of the two mean branch-oriented responses. This common rule selects block 22 for Qwen and block 13 for Llama.

After localization, the factorial directions are reconstructed from all 24 discovery domains at the selected layer. The selected layer and full-discovery directions are then frozen before novel-domain evaluation. Discovery-domain evaluation continues to use leave-one-domain-out directions, whereas novel-domain evaluation uses the corresponding frozen full-discovery direction. No novel-domain, relation-control, or behavioral result contributes to these choices (Appendix C).

We keep three estimands distinct throughout: small-dose finite differences for localization, exact derivatives at $\alpha = 0$ for directional specificity, and finite-dose OLS effects S_A, S_D for the primary representation-level results. Exact estimators and implementation details are given in Appendix B.

Table 1: Teacher-forced finite-dose intervention effects. S_A and S_D are branch-oriented slopes, $I = S_A + S_D$ is the relational component, and $G = S_A - S_D$ the branch imbalance. Discovery results are descriptive; novel-domain evaluation uses the frozen layer and full-discovery direction.

Model	Evaluation	S_A	S_D	I	G	$I_d > 0$
Qwen	Discovery	0.543	0.741	1.284	−0.199	23/24
Qwen	Novel	1.240	0.276	1.517	0.964	22/24
Llama	Discovery	1.145	0.274	1.419	0.872	24/24
Llama	Novel	0.587	1.073	1.660	−0.486	24/24

Table 2: Local directional specificity at the selected Llama layer. J_A and J_D are exact RMS-scaled local derivatives at $\alpha = 0$, and $W^{\mathrm{loc}} = \min(J_A, J_D)$. The Gaussian direction is selected from 1,024 draws on discovery domains and then applied unchanged to novel domains. Discovery-domain p-values are descriptive because the discovery domains contribute to direction construction; for the selected Gaussian, they are additionally post-selection quantities.

Direction	Evaluation	J_A	J_D	W^{loc}	p_A	p_D
Frozen v_{EY}	Discovery	1.189	0.306	0.306	3.0×10^{-6}	0.051
Frozen v_{EY}	Novel	0.565	1.143	0.565	8.9×10^{-5}	3.0×10^{-7}
Best of 1,024 Gaussian	Discovery	0.297	0.292	0.292	2.5×10^{-5}	1.1×10^{-4}
Best of 1,024 Gaussian	Novel	0.216	0.457	0.216	8.3×10^{-4}	2.5×10^{-5}

3 Context-responsive relational steering with a fixed direction

We ask whether a fixed relational direction satisfies the required branchwise sign criterion, whether that pattern is specific to the factorial construction, and whether it transfers to domains not used to construct the direction. Interventions use Qwen block 22 and Llama block 13, selected by the discovery-only procedure in Section 2.

Discovery responses and direction controls. Both model families show the expected mean branch orientation on the discovery domains (Table 1), though these results are descriptive because the same domains contribute to direction construction and layer localization. Structured controls show that intervention magnitude alone does not establish relational sign structure: in Llama, the outcome-valence direction v_Y produces a much larger alignment-branch response than v_{EY} ($S_A = 4.635$), but moves the distortion branch in the wrong relational orientation ($S_D = -3.684$), and the epistemic main-effect direction v_E behaves similarly. Projected Gaussian and balanced sign-randomized controls give the complementary result: their aggregate responses are small on average, yet individual control directions can satisfy the positive branch orientation or produce a large aggregate effect while failing one branch, so neither a positive aggregate I nor the branchwise sign pattern uniquely identifies the factorial direction (full results in Appendix E).

The sign pattern is not unique, but v_{EY} has a stronger held-out response. We therefore test a stronger form of specificity using a best-of-1,024 Gaussian search. This search is distinct from the projected Gaussian controls above: isotropic directions are drawn without projecting out the factorial axes, globally oriented on the discovery domains, and ranked by $W^{\mathrm{loc}} = \min(J_A, J_D)$. The selected direction is then applied unchanged to the novel domains. Here J_A and J_D are exact RMS-scaled derivatives at $\alpha = 0$, rather than the finite-dose slopes in Table 1.

In Llama, the selected Gaussian direction is nearly orthogonal to v_{EY} (cosine 0.027), yet satisfies the branchwise sign criterion on both discovery and novel domains (Table 2), so the sign criterion alone does not identify v_{EY}. The directions differ substantially in held-out magnitude: W^{loc} for the selected Gaussian falls from 0.292 to 0.216 from discovery to novel domains, whereas the frozen factorial direction increases from 0.306 to 0.565, giving a novel-domain worst-branch response roughly $2.6\times$ that of the discovery-selected Gaussian.

This held-out advantage remains positive under a post-hoc selection-aware bootstrap conditioned on the archived Gaussian bank and frozen v_{EY}: $\Delta W^{\mathrm{loc}} = W^{\mathrm{loc}}_{EY} - W^{\mathrm{loc}}_{G} = 0.349$, with a 95%

percentile interval of $[0.134, 0.676]$, positive in both branches separately (resampling details in Appendix E). Qwen shows an even wider separation under the same diagnostic ($W^{\mathrm{loc}} = 0.132$ for the best of 1,024 Gaussian draws versus 0.615 for v_{EY}). These results support comparative held-out specificity rather than uniqueness; Appendix F examines why unrelated directions can nevertheless satisfy the branchwise sign criterion.

Held-out finite-dose transfer. Returning to the finite-dose estimand in Table 1: on the 24 novel domains, Qwen retains a positive aggregate relational component ($I = 1.517$, $I_d > 0$ in 22/24 domains), and its alignment branch passes the directional test ($S_A = 1.240$, $p = 1.2 \times 10^{-6}$), but its distortion branch does not ($S_D = 0.276$, $p \approx 0.097$), so Qwen fails the strict branchwise criterion despite the large aggregate effect.

Llama passes the directional test in both branches ($S_A = 0.587$, $p = 4.8 \times 10^{-5}$; $S_D = 1.073$, $p = 7.2 \times 10^{-7}$), with $I = 1.660$ positive in all 24 novel domains ($p = 2^{-24}$). Its branches are imbalanced in the direction opposite the prespecified secondary prediction ($G = -0.486$, 95% CI $[-0.882, -0.086]$, versus $G > 0$), and the dominant branch even reverses from discovery ($G = 0.872$) to novel domains; since the aggregate component remains strong throughout, what transfers is relational writability rather than a fixed dominant branch.

Aggregate held-out relational writability therefore transfers in both model families, but strict branchwise held-out transfer is established only in Llama. Its novel-domain ordering, $S_D = 1.073 > S_A = 0.587$, provides the key comparison for Section 4: whether the branch more writable under the controlled teacher-forced readout also produces the stronger and more persistent effect during open-ended generation.

4 Behavioral realization in open-ended generation

We next ask whether the teacher-forced writability established in Section 3 is expressed during open-ended generation and whether that expression persists after the intervention stops.

Prefill-only trajectory audit. We intervene on prompt residual states at the frozen layer during prefill only; generated tokens receive no direct reinjection of the steering direction. For each model family, we generate matched continuations across 24 novel domains, two epistemic branches, $\alpha = \pm 1.5$, and 30 seeds, yielding 2,880 continuations per family. Generation is capped at 192 tokens, with no outcome-based regeneration or exclusion.

For Llama, $|\alpha| = 1.5$ was not prospectively prespecified: it was selected by a 10-seed exploratory weakest-branch pilot, then frozen and evaluated on 30 new generation seeds with no further adjustment, so we treat it as a pilot-selected hyperparameter rather than a prespecified dose (Appendix D).

Each continuation is evaluated at five normalized checkpoints $u \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ using a frozen, blinded GPT-5.6 Sol judge that sees only the prompt and visible prefix. Matched $+\alpha/-\alpha$ differences are oriented so that positive values denote movement in the relation-predicted direction in both branches. Their five-checkpoint average defines Q_{traj}; we also examine the terminal effect Q_{terminal} and early-to-late attenuation C_{atten}. Exact construction, aggregation, and inference are given in Appendix D.

The prefix-censored trajectory instrument is diagnostic for Qwen because it was developed after Qwen's original free-generation experiment. For Llama, the checkpoint and censoring rules, rubric, matched contrasts, estimands, and planned tests were frozen before the reported 30-seed behavioral generation.

Teacher-forced writability and open-ended behavior have different branch orderings. Under the primary Llama audit, both branches pass the prespecified trajectory-average directional test ($Q_{\mathrm{traj},A} = 0.201$, $p = 0.000339$; $Q_{\mathrm{traj},D} = 0.050$, $p = 0.048806$, though the distortion-side 95% interval includes zero), with a direct branch contrast of $B_{\mathrm{traj}} = Q_{\mathrm{traj},A} - Q_{\mathrm{traj},D} = 0.151$ (95% CI $[0.039, 0.279]$, $p = 0.0094$). This reverses the branch ordering of the held-out teacher-forced readout:

$$S_D > S_A \quad \text{(teacher-forced)}, \qquad Q_{\mathrm{traj},A} > Q_{\mathrm{traj},D} \quad \text{(open-ended)},$$

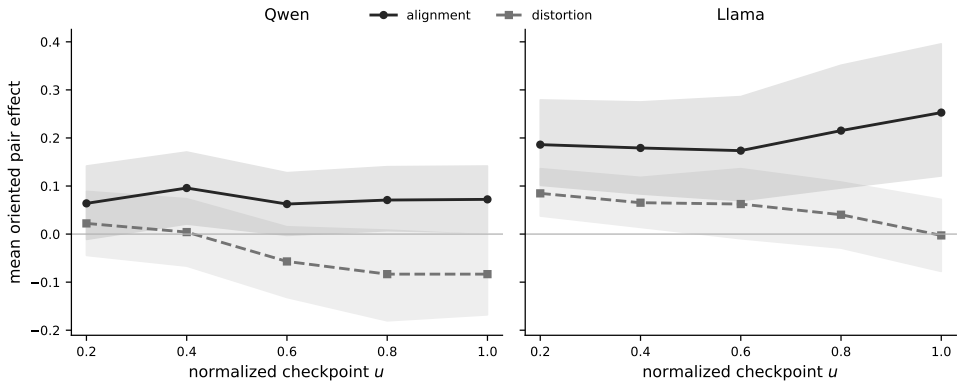


Figure 2: Primary-judge behavioral trajectory effects across normalized checkpoints on the 24 novel domains. Positive values denote movement in the relation-predicted direction, and shaded bands show 95% bootstrap confidence intervals over domains. Llama shows positive early effects on both branches with distortion-side attenuation, whereas Qwen shows a small alignment effect without a detectably positive distortion trajectory.

with $S_D = 1.073$, $S_A = 0.587$. The teacher-forced ordering was observed post hoc, whereas the behavioral test was prespecified; the inversion is inconsistent with a simple monotonic account in which greater teacher-forced writability yields the stronger open-ended effect, though it does not identify the mechanism responsible.

Initial expression does not guarantee persistence. In Llama, the distortion effect is positive early in generation ($Q_{\text{early},D} = 0.075$) but attenuates over time ($C_{\text{atten},D} = 0.056$, $p = 0.033$; Figure 2), and at the terminal checkpoint $Q_{\text{terminal},D} = -0.003$ is not detectably positive, whereas the alignment effect remains positive ($Q_{\text{terminal},A} = 0.253$; further tests in Appendix D). This nonsignificant terminal effect is not evidence of equivalence to zero—many continuations reach the 192-token cap, so the terminal checkpoint need not correspond to a natural narrative ending—and the supported conclusion is narrower: positive trajectory-average expression need not imply a detectably positive effect at the final observed checkpoint.

Measurement robustness. A post-hoc independent full blind rescoring reproduces positive trajectory-average effects in both Llama branches, the alignment-dominant ordering, and distortion-side attenuation, while a blinded 480-item human calibration subset supports positive average effects in both branches but not the finer attenuation pattern; we therefore treat the human analysis as a measurement-validity check rather than a full replication (details in Appendix D).

Qwen and the scope of the behavioral evidence. Qwen shows a positive alignment trajectory under its diagnostic audit ($Q_{\text{traj},A} = 0.073$, $p = 0.018$), but its distortion trajectory is not detectably positive ($Q_{\text{traj},D} = -0.039$, $p = 0.86$). It therefore fails the strict trajectory-level branchwise criterion, and we do not treat the Qwen audit as confirmatory cross-model behavioral replication.

The behavioral evidence is therefore strongest for Llama. Taken together, the results separate three properties that need not coincide: controlled-readout writability, open-ended behavioral expression, and persistence after a one-shot prefill intervention. Exploratory transport and readout diagnostics in Appendix F do not explain the Llama ordering inversion.

5 Relation sensitivity and limits of mechanism-sensitive interpretation

Having established held-out relational writability, we ask two stronger questions. First, does the intervention depend on the stated relation between the epistemic premise and its consequence? Sec-

ond, does the fixed direction recompute its orientation when the scenario explicitly specifies the opposite causal mechanism?

We evaluate two discovery-domain relation controls while keeping the intervention direction, selected layer, and finite-dose evaluation rule fixed. The noncausal condition preserves the epistemic and outcome content but removes their stated causal linkage, whereas the explicit-reversal condition specifies a mechanism under which distortion should produce favorable consequences and alignment unfavorable consequences (Appendix A.3). Neither condition is used to construct a new direction.

Removing the stated linkage weakens relational writability. Removing the causal linkage substantially reduces the aggregate finite-dose intervention effect in both model families. In Llama, I falls from 1.419 to 0.311 ($p = 1.6 \times 10^{-6}$), and in Qwen from 1.284 to 0.435 ($p = 2.1 \times 10^{-5}$). The reduction is branch-asymmetric in Llama and involves both branches in Qwen; full paired intervals and branchwise decompositions are reported in Appendix E.2.

Thus the intervention response depends on whether the epistemic premise is explicitly linked to the consequence. This argues against an account based only on co-occurrence of the same premise and outcome content. Relation sensitivity alone, however, does not establish that the direction recomputes a causal rule from the mechanism stated in the scenario.

Explicit reversal does not produce the opposite joint pattern. A mechanism-sensitive recomputation account predicts that, under explicit reversal, the default-oriented effects should jointly change sign; neither model family shows this. Qwen attenuates strongly ($S_A = -0.197$, $S_D = 0.547$, $I = 0.350$), with only its alignment branch changing sign, while Llama retains both default-oriented branch effects ($S_A = 0.773$, $S_D = 0.720$, $I = 1.493$), and the prespecified directional test does not support the predicted aggregate decrease relative to default ($p \approx 0.681$; full contrasts in Appendix E.2).

This aggregate comparison alone cannot localize the failure to the steering direction: under the manuscript-primary total-log-likelihood readout, the unsteered $\alpha = 0$ model also fails the required joint reversal in either family (Appendix E.3), so aggregate reversal failure could reflect the fixed direction, the countervailing scenario's representation, the downstream readout, or some combination.

Baseline conditioning sharpens the reversal diagnostic. The unsteered result leaves open a simpler alternative: the intervention might merely amplify whatever consequence preference is already present at $\alpha = 0$. We therefore related the oriented intervention slope S to the correspondingly oriented unsteered margin $C^{(0)}$ for each domain, relation condition, and branch, allowing separate branch intercepts and resampling semantic domains as clusters (Appendix E.4).

The clearest dissociation occurs in Llama under explicit reversal. With the manuscript-primary total-log-likelihood score, the branch-adjusted fit gives

$$a_A = 0.690 \quad [0.318, 1.053], \qquad a_D = 0.620 \quad [0.373, 0.868],$$

while the baseline coefficient is small, $\lambda = 0.031$ $[-0.032, 0.091]$. Among 13 branch-level reversal cases where the unsteered preference points against the default relational orientation, 10 intervention slopes still point in the default orientation, so the Llama reversal response is not well described as simple amplification of the contemporaneous preference—a conclusion that survives continuation-length (mean-token) normalization, under which 15 of 18 baseline-conflict slopes still follow the default orientation (Appendix E.4).

Qwen shows a different profile: under reversal its response is substantially more coupled to the unsteered preference. With mean-token scoring, the branch-adjusted intercepts are near zero while the baseline coefficient remains positive,

$$a_A = -0.005\ [-0.018, 0.013], \qquad a_D = 0.021\ [-0.009, 0.039], \qquad \lambda = 0.036\ [0.008, 0.081],$$

and the relation-condition changes in baseline preference and intervention response are correspondingly more strongly coupled in Qwen than in Llama.

Interpretive boundary. The relation controls support three conclusions. First, removing the stated linkage weakens relational writability in both model families. Second, explicit reversal does not

induce the opposite joint intervention pattern. Third, the Llama reversal effect cannot be reduced to simple amplification of the contemporaneous controlled readout, whereas Qwen is substantially more baseline-tracking under the same manipulation.

These results do not establish that the baseline-independent Llama component is uniquely specific to the epistemic–outcome relation; a generic semantic-congruence account has not been exhaustively excluded. Nor do they establish that the fixed direction recomputes its orientation from an arbitrary stated causal mechanism. In particular, the present intervention does not by itself establish a causal abstraction in the stronger sense that interventions on internal representations correspond systematically to interventions in a specified high-level causal model (Geiger et al., 2021; 2024).

The supported claim is therefore narrower: the factorial direction enables relation-sensitive, context-responsive writability that can contain a baseline-independent premise-conditioned component, but relational writability does not by itself establish mechanism-sensitive causal-rule recomputation.

6 Related work

Activation steering and relational representations. Activation steering modifies model behavior through inference-time interventions on internal representations (Turner et al., 2024; Rimsky et al., 2024; Li et al., 2023b; Zou et al., 2023), with effects that can vary across inputs and prompts (Tan et al., 2024; Braun et al., 2025); related work on task vectors, function vectors, and relational representations shows that compact activation-space structure can support input-dependent behavior (Hendel et al., 2023; Todd et al., 2024; Hernandez et al., 2024; Chanin et al., 2024). Our setting differs in requiring one fixed signed direction to move the same outcome-preference readout in opposite directions across premise types, separating relational sign structure from overall intervention magnitude.

Adaptive steering and behavioral persistence. Recent methods make steering explicitly context dependent through conditional gating, dynamic scaling, or locally updated directions (Lee et al., 2025; Ferrando et al., 2026; Li et al., 2026); we instead keep both the direction and signed dose schedule fixed and ask whether a prompt-level write remains behaviorally expressed after the intervention stops, connecting to evidence that success under controlled evaluation does not necessarily transfer to open-ended generation (Pres et al., 2024; Wu et al., 2025; Su et al., 2026). Our trajectory analysis adds a within-model comparison between teacher-forced writability and the strength and persistence of the resulting open-ended effect.

Causal interpretation. Decodable information in a representation does not by itself establish causal use (Belinkov, 2022). Causal tracing and editing, interventions on learned world representations, and analyses of linear representation geometry provide complementary tools for studying internal causal structure (Meng et al., 2022; Li et al., 2023a; Park et al., 2024), while causal abstraction makes the stronger claim that interventions on internal representations should correspond systematically to interventions in a specified high-level causal model (Geiger et al., 2021; 2024). Our interventions establish causal effects on a controlled consequence-preference readout, while the relation controls in Section 5 show that this is insufficient to establish mechanism-sensitive recomputation of the stated causal rule.

7 Discussion and limitations

Relational writability transfers, but not uniformly. Aggregate held-out relational writability transfers in both Qwen and Llama, whereas strict branchwise transfer is established only in Llama, so the cross-model claim is aggregate writability with the stronger two-branch evidence confined to Llama. Directional specificity is likewise comparative rather than unique: unrelated directions can satisfy the branchwise sign criterion, while the factorial direction has the larger held-out local response among the alternatives tested.

Writability, behavioral expression, and persistence separate. In Llama, the branch ordering reverses across evaluation regimes— distortion is stronger under the controlled teacher-forced readout, whereas alignment has the larger and more persistent trajectory-average effect during prefill-

only open-ended generation—establishing a dissociation between controlled-readout writability and subsequent behavior, though not its mechanism. The positive trajectory-average direction in both Llama branches is supported by two automated audits and the human calibration subset, but the finer temporal pattern (distortion-side attenuation) is reproduced only by the automated audits; we therefore treat the average direction as the more measurement-robust result, and the Qwen trajectory analysis remains diagnostic rather than confirmatory.

Relation sensitivity does not imply causal-rule recomputation. As Section 5 establishes, removing the stated causal linkage weakens relational writability in both model families, but explicit reversal does not induce the opposite joint pattern, and the aggregate comparison is further limited because the unsteered teacher-forced readout also fails the required joint reversal. The post-hoc baseline-conditioned analysis sharpens this diagnostic: default-oriented Llama effects persist after conditioning on the contemporaneous unsteered preference, and survive mean-token normalization, whereas Qwen's intervention effects track the unsteered preference substantially more closely under the same manipulation. Baseline-linked and premise-conditioned components thus contribute differently across model families, but this does not establish that the baseline-independent Llama component is unique to the epistemic–outcome relation, nor that the direction recomputes its orientation from an arbitrary stated mechanism: the present teacher-forced readout is evidence beyond simple baseline amplification, not a test of whether the models understand the counterfactual mechanism in general.

Limitations. The study covers two 8B-parameter pretrained base models, and the strongest branchwise and behavioral evidence is Llama-specific. The novel sets test transfer to disjoint semantic domains within a shared factorial grammar rather than arbitrary out-of-template generalization. The best-of-1,024 Gaussian comparison is also bounded: its selection-aware bootstrap is post hoc, conditions on the archived candidate bank, and does not exhaust possible data-driven directions. The behavioral experiment adds further qualifications. The reported $|\alpha| = 1.5$ magnitude was selected in an exploratory pilot and frozen before evaluation on new generation seeds rather than being prespecified before any behavioral pilot. Many continuations reach the 192-token cap, so the terminal checkpoint need not represent a natural narrative ending. The human calibration uses one rater with sparse seed coverage and is not a fully powered human replication. Finally, all persistence results concern a one-shot prefill intervention with no decode-time reinjection; they do not establish limits under repeated, adaptive, or continuous steering.

Implications and conclusion. Taken together, the results support a hierarchy of increasingly strong claims rather than a single notion of successful steering. The factorial direction can causally write a relational effect into a controlled teacher-forced readout, and that aggregate effect transfers to semantic domains not used to construct the direction; in Llama, the stronger branchwise criterion also transfers. These results establish relational writability, but they do not make the other properties examined in this paper automatic:

$$\text{relational writability} \not\Rightarrow \begin{cases} \text{directional uniqueness,} \\ \text{persistent behavioral expression,} \\ \text{mechanism-sensitive rule recomputation.} \end{cases}$$

Each non-implication has a direct source in the results above. An almost orthogonal search-selected Gaussian direction can also satisfy the branchwise sign criterion, so specificity is comparative rather than a claim that v_{EY} is unique. In Llama, the branch more writable under teacher forcing is not the branch with the stronger or more persistent behavioral effect, so controlled-readout writability need not persist. And removing the stated linkage weakens steering in both models, while the baseline-conditioned Llama component is not reducible to simple amplification of the contemporaneous preference—yet neither result shows the direction recomputing an arbitrary stated causal rule.

A fixed factorial direction is therefore most usefully interpreted as a causal probe of relational writability rather than as a complete implementation of a causal rule: one unchanged intervention direction can produce opposite context-dependent effects and transfer beyond the domains from which it was constructed, but the sign pattern is not unique to the direction, controlled writability does not guarantee persistent open-ended behavior, and relation sensitivity does not by itself establish mechanism-sensitive causal-rule recomputation.

REPRODUCIBILITY STATEMENT

For both model families, the discovery stimuli were fixed before the layerwise intervention scan. The same discovery-only localization criterion, using symmetric small-dose interventions at $\epsilon = 0.10$, selected block 22 for Qwen and block 13 for Llama. For each model, the selected layer and the direction reconstructed from the full discovery set were then frozen before evaluation on the independent novel domains. No novel-domain, relation-control, or open-generation result contributed to layer selection or construction of the frozen novel-evaluation direction.

The Llama study additionally followed a prospectively specified representation-level protocol covering the model revision, tokenizer and span extraction, discovery/novel split, axis construction and localization rule, finite-dose grid, and representation-level hypotheses. For the behavioral study, the checkpoint and censoring rules, rubric, matched contrasts, estimands, and planned tests were fixed before the reported 30-seed generation experiment. The intervention magnitude was an explicit exception: $|\alpha| = 1.5$ was selected in a 10-seed exploratory pilot and then frozen before evaluation on 30 new seeds.

For the primary Llama trajectory audit, checkpoint construction, prefix censoring, the scoring rubric, matched contrasts, estimands, and tests were fixed before the reported 30-seed generation experiment. A later operational freeze specified blind-packet construction, judge schema, validation, and execution policy before annotation without changing the scientific endpoints. The primary GPT-5.6 Sol audit remained blinded until all 14,400 expected prefix-level annotations were valid, after which the researcher key was joined. The Qwen prefix-censored trajectory analysis was developed after its original free-generation experiment and is therefore treated as diagnostic rather than confirmatory.

Several analyses reported as robustness or interpretive diagnostics were conducted after the primary Llama audit and are not part of the prospectively frozen confirmatory protocol. These include the complete 14,400-item blind rescoring with Gemini-3.6 Flash, the 480-item single-rater human calibration subset, and the selection-aware best-of-1,024 Gaussian bootstrap. The latter conditions on the archived Gaussian candidate bank and frozen v_{EY}, resamples discovery domains while rerunning Gaussian orientation and winner selection, and then resamples the held-out novel domains.

The relation-grid analyses added during manuscript revision are also post-hoc interpretive diagnostics. The $\alpha = 0$ values used to assess unsteered relation-condition preferences were already present in the archived evaluation grids. The baseline-conditioned analysis likewise uses only those archived finite-dose grids: it conditions each branch-oriented intervention slope on its contemporaneous oriented $\alpha = 0$ margin, uses semantic-domain cluster resampling for uncertainty, and repeats the analysis with mean-token candidate scores as a continuation-length sensitivity check. No new model inference, layer selection, direction construction, or endpoint selection was performed for this diagnostic. These analyses are used to bound interpretation rather than to redefine the original confirmatory endpoints.

Appendix C reports the study chronology and freeze manifests, Appendix B reports the estimators and statistical procedures, Appendix E.4 reports the post-hoc baseline-conditioned diagnostic, and Appendix D documents the behavioral-generation and blind-audit procedures. The machine-readable artifacts record the frozen stimuli and split metadata, localization outputs, domain-level intervention results, exact-gradient and Gaussian control outputs, archived relation-condition grids and baseline-conditioned post-processing outputs, behavioral generations, blind audit packets and researcher keys, automated annotation ledgers, the human calibration subset, and the analysis scripts used for the reported results.

ETHICS STATEMENT

This work studies internal representations and intervention effects using synthetic organizational and technical scenarios. The terms *alignment* and *distortion* are used only in the task-specific sense of whether a premise accurately represents the state of affairs described in the scenario; they do not refer to AI alignment with human values. The stimuli contain no real individuals, personal identifiers, or personal records.

The intervention experiments modify residual-stream activations of pretrained language models and are intended as diagnostic analyses of representation-level writability and behavioral expression.

They are not evaluations of moral reasoning, normative correctness, or deployed decision making. In particular, the results should not be interpreted as establishing human-like causal understanding or as demonstrating reliable behavioral control in deployment.

Most behavioral measurements were produced by blinded automated language-model judges. As a complementary measurement-validity check, one external human participant served as a blinded rater for a 480-item calibration subset of synthetic model continuations. The participant was informed of the annotation task and its research purpose and provided consent before participating. The participant was compensated for the annotation work. The task involved evaluating only synthetic prompts and model-generated text. No personal identifiers or sensitive personal information were collected; only the participant's task annotations entered the calibration analysis.

The human annotations are used only to assess the robustness and calibration of the behavioral measurement procedure. Because the subset used a single rater and substantially fewer generation seeds than the full automated audit, it is not treated as a fully powered human replication of the trajectory experiment or as evidence about a human population.

The open-generation results also illustrate a potential limitation of activation steering: a representation-level intervention can produce a detectable controlled-readout effect without yielding equally strong or persistent behavior during subsequent generation. We therefore present the intervention primarily as a diagnostic tool rather than as a method for reliable behavioral control.

AI USE STATEMENT

Large language models were used as research assistants for drafting and revising prose, discussing experimental and analysis procedures, implementing and debugging code, preparing analysis and plotting scripts, drafting candidate stimulus templates for the synthetic factorial scenarios (Appendix A), checking experimental documentation, and supporting literature search. The authors reviewed, edited for the required factorial and lexical controls, and froze all stimuli before any target-model analysis (Appendix A.4). The authors retained responsibility for all experimental design choices, stimulus content, protocol freezes, statistical procedures, interpretations, and reported claims. Numerical results reported in the paper were produced by the corresponding analysis pipelines and checked against the archived experimental artifacts rather than being taken from language-model-generated estimates.

Language models were also used as measurement instruments in the behavioral experiments, separately from their use as research assistants. The primary blinded Llama trajectory audit used GPT-5.6 Sol as the automated judge. A subsequent robustness analysis rescored the complete blind audit with Gemini-3.6 Flash using the same prefix-censored evaluation packets and joining the researcher key only after blind scoring was complete. These automated judgments were then supplemented by a 480-item human calibration subset. The identities, blind-packet construction, scoring rubric, execution procedures, agreement analyses, and interpretive limitations of these measurement instruments are documented in Appendix D.

The automated judges were not used to choose intervention directions, layers, discovery/novel splits, or the scientific hypotheses tested by the behavioral audit. Judge outputs enter only the explicitly defined behavioral measurement stage.

REFERENCES

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

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (Un)reliability of steering vectors in language models, 2025. URL https://arxiv.org/abs/2505.22637.

David Chanin, Anthony Hunter, and Oana-Maria Camburu. Identifying linear relational concepts in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long*

Papers), pp. 1524–1535. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.85. URL https://aclanthology.org/2024.naacl-long.85/.

Alex Ferrando, Xavier Suau, Jordi Gonzàlez, and Pau Rodriguez. Dynamically scaled activation steering, 2026. URL https://arxiv.org/abs/2512.03661.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, 2024.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, 2024.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*, 2025.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering vector fields for context-aware inference-time control in large language models, 2026. URL https://arxiv.org/abs/2602.01654.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023a.

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*, 2023b.

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

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs, 2024. URL https://arxiv.org/abs/2410.17245.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Xuanbo Su, Hao Luo, Yingfang Zhang, and Lijun Zhang. ROAST: Rollout-based on-distribution activation steering technique, 2026. URL https://arxiv.org/abs/2602.14143.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.15213.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. In *arXiv preprint arXiv:2308.10248*, 2024.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders, 2025. URL https://arxiv.org/abs/2501.17148.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Experimental materials and dataset construction

This appendix records the stimulus-design details needed to interpret the factorial direction, held-out transfer, and relation controls in the main text. Machine-readable stimulus and freeze records are described in Appendix G.

A.1 Factorial construction

Each semantic domain instantiates a 2×2 crossing of epistemic status and systemic outcome,

$$E \in \{A, D\}, \qquad Y \in \{+, -\}.$$

Alignment (A) means that a record, signal, measurement, report, or analogous representation accurately reflects the task-relevant state of affairs. Distortion (D) means that it materially misrepresents that state. Favorable ($Y+$) and unfavorable ($Y-$) outcomes are defined at the level of system functioning rather than individual gain or loss.

Modular clause reuse. Within each semantic item, the epistemic premise is held fixed across outcome conditions, and the consequence clause is held fixed across epistemic conditions:

$$\begin{aligned} \text{premise}_A(A+) &= \text{premise}_A(A-), \\ \text{premise}_D(D+) &= \text{premise}_D(D-), \\ Y^+(A+) &= Y^+(D+), \\ Y^-(A-) &= Y^-(D-). \end{aligned}$$

The four cells therefore differ only in the factorial combination of an epistemic premise and a systemic consequence, rather than being written as four independent scenarios.

Table 3 gives a complete example from the maintenance domain. The repeated text is intentional: it makes explicit which linguistic material is held constant across each factorial contrast.

The complete stimulus is formed by combining the domain-specific context, one of the two epistemic premises, the assigned connector, and one of the two consequence clauses. Connector families vary across items, but within a semantic item the factorial comparisons preserve the intended clause reuse.

Why the interaction contrast is relational. For a domain d, the representation-level interaction is

$$d_{EY,d} = \frac{1}{4}\left(\bar{h}_{A+} - \bar{h}_{A-} - \bar{h}_{D+} + \bar{h}_{D-}\right).$$

The coefficient pattern sums to zero within both factorial main effects. Consequently, a purely additive alignment–distortion difference or a purely additive favorable–unfavorable difference cancels from this contrast by construction.

The modular construction in Table 3 makes this interpretation concrete. For the alignment branch,

$$\bar{h}_{A+} - \bar{h}_{A-}$$

compares two stimuli with the same epistemic premise while changing only the systemic consequence. For the distortion branch,

$$\bar{h}_{D+} - \bar{h}_{D-}$$

does the same under the distorted premise. The interaction then asks whether this favorable-minus-unfavorable representational difference changes with epistemic status:

$$4d_{EY,d} = (\bar{h}_{A+} - \bar{h}_{A-}) - (\bar{h}_{D+} - \bar{h}_{D-}).$$

Thus the target is not the representation of an aligned premise, a distorted premise, or outcome valence in isolation. It is the change in the outcome contrast conditional on the epistemic premise.

Table 3: Example 2×2 factorial item from the maintenance domain. The alignment premise is reused in $A+$ and $A-$, the distortion premise in $D+$ and $D-$, and each outcome clause is reused across epistemic conditions.

Cell	Epistemic premise	Systemic consequence
$A+$	The maintenance record accurately reflected the equipment's actual condition.	The repair teams coordinated their work effectively, and the maintenance schedule remained stable.
$A-$	The maintenance record accurately reflected the equipment's actual condition.	The repair teams made conflicting decisions, and the maintenance schedule broke down.
$D+$	The maintenance record substantially misrepresented the equipment's actual condition.	The repair teams coordinated their work effectively, and the maintenance schedule remained stable.
$D-$	The maintenance record substantially misrepresented the equipment's actual condition.	The repair teams made conflicting decisions, and the maintenance schedule broke down.

Lexical and semantic controls. Distortion is not implemented as simple negation where avoidable. Affirmative mismatch descriptions such as *substantially misrepresented* are preferred so that the epistemic factor is not carried by a negation marker. Strongly moralized terms such as *honest*, *dishonest*, *lie*, *truthful*, *deceptive*, *good*, and *bad* are excluded from the primary stimuli so that epistemic status is not trivially confounded with moral valence.

Premise and consequence lengths and syntax are kept approximately parallel, outcomes concern system-level functioning rather than actor-specific gain or loss, and connector families are varied across items. These constraints were checked before target-model analysis, and the resulting datasets were frozen before intervention outcomes were inspected.

A.2 Discovery and novel domains

Discovery and novel evaluation each use 24 semantic domains (Table 4). The split is at the domain level rather than a random sentence split. Novel domains contribute neither to construction of the relational axis nor to layer localization. For the default relation, each set contains $24 \times 2 \times 2 = 96$ factorial stimuli. The discovery domains also instantiate the noncausal and explicit-reversal relation variants described below, giving 288 discovery stimuli in total.

A.3 Relation-condition variants

The discovery stimuli are instantiated under three relation conditions,

$$R \in \{\text{default causal},\ \text{noncausal},\ \text{explicit reversal}\}.$$

The *default causal* condition makes the epistemic representation relevant to a downstream decision or system process. This is the condition used to construct the relational direction and to define the default alignment–outcome relation tested in the main experiments.

Noncausal condition. The noncausal condition retains the epistemic premise and systemic outcome content while removing the stated causal linkage between them. Wherever possible, the same premise and consequence clauses are reused and only the relational context or connector is changed.

For example, the aligned-positive maintenance item from Table 3 becomes

> The maintenance record accurately reflected the equipment's actual condition. Separately, the repair teams coordinated their work effectively, and the maintenance schedule remained stable.

The epistemic content and favorable consequence are therefore the same as in the corresponding default-causal cell, but the sentence no longer states that the former produced or supported the latter. The comparison

$$I_{\text{default}} - I_{\text{noncausal}}$$

Table 4: Discovery and novel domain sets. The sets are disjoint; novel-domain activations and intervention outcomes do not contribute to axis construction or layer localization.

Discovery (24)	Novel (24)
maintenance	court docket administration
navigation	university prerequisite clearance
medical monitoring	patent application classification
laboratory measurement	municipal permit review
manufacturing QA	library cataloging
logistics	museum accession management
inventory control	archival finding-aid curation
aviation	dictionary lexicography curation
emergency response	publishing production workflow
software monitoring	theatre stage cue coordination
cybersecurity telemetry	orchestra score-edition coordination
energy systems	film continuity management
water monitoring	hotel room-turnover administration
structural inspection	restaurant table-service coordination
financial reconciliation	venue ticket-admission administration
supply-chain tracking	cemetery plot registry
forecasting	sports tournament result administration
robotics	debate tournament tabulation
scheduling	academic peer-review assignment
agricultural sensing	translation terminology management
transportation control	property title registry
auditing	customs tariff classification
environmental monitoring	zoological pedigree studbook management
scientific instrumentation	botanical herbarium accession

tests whether the intervention response is reduced when this stated linkage is removed, rather than merely testing sensitivity to the reused alignment, distortion, favorable-outcome, or unfavorable-outcome content.

Explicit causal reversal. The explicit-reversal condition introduces a countervailing mechanism under which distortion can produce a favorable consequence and alignment an unfavorable one.

For example:

> A faulty dispatch rule assigns fewer resources as reported severity increases and more resources as reported severity decreases. The incident-severity report reversed the relative severity of the active incidents. Consequently, response resources matched operational needs, and coverage across the system remained stable.

Here the distorted report produces the favorable systemic outcome because the downstream mechanism itself is reversed. This condition therefore asks a stronger question than the noncausal control: whether the intervention response changes orientation when the scenario explicitly specifies an opposite causal mechanism.

The reversal stimuli are less closely matched to the default condition in length and naturalness than the noncausal variants because they include an explicit countervailing mechanism. The reversal control is therefore used to test the stronger mechanism-sensitive recomputation hypothesis, with this stimulus mismatch treated as a limitation of that comparison.

A.4 Pre-analysis audit

Before model-targeted analysis, the stimuli were reviewed for the intended factor structure: clear alignment-versus-distortion and favorable-versus- unfavorable contrasts, removal of the causal linkage in the noncausal variants, and an intelligible countervailing mechanism in the reversal variants. They were also screened for obvious length, syntax, negation, or moral-vocabulary imbalances.

This screening does not establish the absence of all linguistic confounds. Wording was not revised in response to target-model outcomes.

B IMPLEMENTATION AND STATISTICAL DETAILS

This appendix records the estimators and inferential procedures used for the reported numerical results. Throughout, the semantic domain is the inferential unit. Tokens, prompt positions, intervention doses, generation seeds, and individual audit judgments are aggregated within domain before cross-domain inference. The paper distinguishes discovery-stage localization, exact local directional specificity, finite-dose representation-level effects, and open-generation behavioral effects; these quantities answer different questions and are not interchanged.

B.1 TOKENIZATION AND CONSEQUENCE-SPAN EXTRACTION

Representation-level axes are constructed from hidden states pooled over the consequence portion of each factorial stimulus. For the prospectively specified Llama representation-level replication, tokenization uses a fast tokenizer with offset mappings, right padding, and `add_special_tokens=False`. Character offsets identify the consequence-token set C; at layer ℓ the pooled representation is

$$\bar{h}^{(\ell)} = \frac{1}{|C|} \sum_{t \in C} h_t^{(\ell)},$$

with padding positions excluded. The consequence span is used rather than the final token alone because the factorial contrast concerns how the premise is represented together with the complete consequence clause. The same span rule is used for all four factorial cells of a domain, and span boundaries were fixed before axis construction and were not revised in response to intervention outcomes.

B.2 AXIS AGGREGATION AND HELD-OUT EVALUATION

For each factorial contrast, raw domain vectors are averaged first and unit normalized only after aggregation. For the relational interaction,

$$\widetilde{v}_{EY} = \frac{1}{N} \sum_{d=1}^{N} d_{EY,d}, \qquad v_{EY} = \frac{\widetilde{v}_{EY}}{\|\widetilde{v}_{EY}\|_2}.$$

Individual domain contrasts are therefore not normalized to equal length before averaging. For completeness, at layer ℓ the corresponding domain-level outcome and epistemic main-effect contrasts are

$$d_{Y,d}^{(\ell)} = \frac{1}{4} \left(\bar{h}_{d,A,+}^{(\ell)} - \bar{h}_{d,A,-}^{(\ell)} + \bar{h}_{d,D,+}^{(\ell)} - \bar{h}_{d,D,-}^{(\ell)} \right),$$

$$d_{E,d}^{(\ell)} = \frac{1}{4} \left(\bar{h}_{d,A,+}^{(\ell)} + \bar{h}_{d,A,-}^{(\ell)} - \bar{h}_{d,D,+}^{(\ell)} - \bar{h}_{d,D,-}^{(\ell)} \right).$$

The interaction, outcome, and epistemic coefficient patterns are orthogonal in the factorial design, although the resulting directions in activation space need not be orthogonal. The same raw-mean-then-normalize construction is used for v_Y and v_E.

Discovery-domain evaluation uses a leave-one-domain-out direction,

$$v_{EY,-d} = \mathrm{unit}\left(\frac{1}{N-1} \sum_{j \neq d} d_{EY,j} \right),$$

so the evaluated domain does not contribute to its own axis. After layer localization, v_{EY}, v_Y, and v_E are reconstructed at the selected block using all 24 discovery domains and are frozen before novel-domain evaluation. Discovery finite-dose analyses continue to use leave-one-domain-out axes; novel-domain analyses use the frozen full-discovery axes.

B.3 Intervention parameterization and representation-level estimands

Layers are reported as decoder-block indices. Interventions are applied to the input residual stream of the selected block rather than to the block output or the final residual stream. For prompt i, non-padding prompt-token residual states receive

$$h_{i,t}^{(\ell^*)} \leftarrow h_{i,t}^{(\ell^*)} + \alpha s_i v,$$

where v is a unit direction and s_i is the prompt-specific residual RMS,

$$s_i = \sqrt{\frac{1}{T_i H} \sum_{t=1}^{T_i} \sum_{j=1}^{H} \left(h_{i,t,j}^{(\ell^*)}\right)^2}. \tag{5}$$

Here T_i is the number of non-padding prompt tokens and H the residual width. The RMS is computed from the unperturbed prompt at the intervention site. Matched $+\alpha$ and $-\alpha$ conditions, and the two candidate consequences for the same prompt, use the same s_i. The relational, structured-control, and randomized-control directions all use this identical scaling and injection rule.

Teacher-forced candidate-consequence tokens receive no direct intervention. During open-ended generation, the perturbation is applied during prompt prefill only and is not reinjected during decoding (Appendix D).

Discovery-only localization. For each candidate block ℓ, discovery domains are evaluated with their leave-one-domain-out relational axes at the symmetric small dose $\epsilon = 0.10$. With $M_{d,e}^{(\ell)}(\alpha)$ denoting the common favorable-minus-unfavorable consequence margin, define

$$\widehat{J}_{A,d,\ell}^{(\epsilon)} = \frac{M_{d,A}^{(\ell)}(+\epsilon) - M_{d,A}^{(\ell)}(-\epsilon)}{2\epsilon},$$

$$\widehat{J}_{D,d,\ell}^{(\epsilon)} = -\frac{M_{d,D}^{(\ell)}(+\epsilon) - M_{d,D}^{(\ell)}(-\epsilon)}{2\epsilon}.$$

The distortion branch is sign-oriented so that positive values indicate movement in the relation-predicted direction in both branches. After averaging each branch separately over the 24 discovery domains, the layer score is

$$W_\ell^{\mathrm{loc},\epsilon} = \min\left(\overline{\widehat{J}}_{A,\ell}^{(\epsilon)}, \overline{\widehat{J}}_{D,\ell}^{(\epsilon)}\right).$$

The minimum is taken after branchwise averaging, not as a mean of domainwise minima. Maximizing this statistic selects block 22 for Qwen and block 13 for Llama. No novel-domain, relation-control, or behavioral result enters the layer choice.

Exact local directional specificity. The Gaussian-specificity analysis uses a distinct estimand: the exact RMS-scaled directional derivative at $\alpha = 0$. For frozen unit direction v,

$$J_{A,d}(v) = \left.\frac{\partial M_{d,A}(\alpha; v)}{\partial \alpha}\right|_{\alpha=0}, \qquad J_{D,d}(v) = -\left.\frac{\partial M_{d,D}(\alpha; v)}{\partial \alpha}\right|_{\alpha=0}.$$

These quantities are computed by exact gradient contraction under the same RMS parameterization, not by finite differences. For a domain set,

$$J_A(v) = \frac{1}{D}\sum_d J_{A,d}(v), \qquad J_D(v) = \frac{1}{D}\sum_d J_{D,d}(v),$$

$$W^{\mathrm{loc}}(v) = \min(J_A(v), J_D(v)).$$

These are the quantities reported in Table 2 and in the best-of-1,024 Gaussian comparison of Appendix E.

Finite-dose representation effects. The main representation-level analysis instead uses OLS slopes over the full teacher-forced grid $\alpha \in \{-1, -0.5, 0, 0.5, 1\}$. Thus $\widehat{J}^{(\epsilon)}$ is the discovery-stage finite-difference localization quantity, J is the exact local derivative used for directional specificity, and S is the finite-dose OLS effect used for the primary representation-level results.

B.4 TEACHER-FORCED READOUT AND FINITE-DOSE REGRESSION

For candidate consequence $c = (c_1, \ldots, c_K)$, the total teacher-forced log-likelihood is

$$L(c \mid x, \alpha, v) = \sum_{k=1}^{K} \log p_{\alpha,v}(c_k \mid x, c_{<k}).$$

The common branch readout is the favorable-minus-unfavorable margin

$$M_{d,e}(\alpha) = L(c_{d,+} \mid x_{d,e}, \alpha, v) - L(c_{d,-} \mid x_{d,e}, \alpha, v).$$

The scoring convention is not reversed between alignment and distortion; the relational prediction is carried by the slope of this same margin.

For each domain and branch we fit an OLS line with intercept,

$$M_{d,e}(\alpha) = a_{d,e} + \beta_{d,e}\alpha + \varepsilon_{d,e,\alpha},$$

over $\alpha \in \{-1, -0.5, 0, 0.5, 1\}$ and orient the branch effects as

$$S_{A,d} = \beta_{d,A}, \qquad S_{D,d} = -\beta_{d,D},$$

$$I_d = S_{A,d} + S_{D,d}, \qquad G_d = S_{A,d} - S_{D,d}.$$

Cross-domain means are unweighted.

B.5 EXACT SIGN-FLIP INFERENCE

Directional hypotheses are tested on domain-level values $z_1, \ldots, z_N$. Under the sign-symmetry null, the exact randomization distribution is generated by

$$z_d^* = \sigma_d z_d, \qquad \sigma_d \in \{-1, +1\}.$$

With $N = 24$, all $2^{24} = 16{,}777{,}216$ sign assignments are enumerated. For a positive directional alternative,

$$p = \frac{\#\{\sigma : \bar{z}^* \geq \bar{z}_{\text{obs}}\}}{2^N}.$$

Negative directional alternatives are handled by reversing the sign of the domain statistic before applying the same calculation. When all 24 domain effects are strictly positive, the observed mean is extremal and the minimum attainable one-sided value is $2^{-24} = 5.96 \times 10^{-8}$.

Paired comparisons are formed before sign flipping. For example, the noncausal-versus-default comparison uses one paired difference per semantic domain, and the same principle is used for explicit-reversal contrasts and for the $\alpha = 0$ baseline diagnostics in Appendix E.3.

B.6 STRICT BRANCHWISE CRITERION

Claims requiring the predicted effect in both branches use an intersection–union criterion rather than the aggregate statistic alone:

$$\text{PASS} \iff p_A < 0.05 \ \wedge \ p_D < 0.05.$$

At the representation level this applies to S_A and S_D; at the behavioral level it applies to $Q_{\text{traj},A}$ and $Q_{\text{traj},D}$. Branch asymmetry is tested directly through G or B_{traj} rather than inferred from a difference in branchwise significance.

B.7 BEHAVIORAL AGGREGATION

Generation, packet construction, and audit procedures are described in Appendix D. Here we define only the aggregation used for the behavioral estimands.

For domain d, seed s, checkpoint u, and audit score R, matched $\pm\alpha$ generations are oriented as

$$\Delta_{A,d,s,u} = R_{+,d,s,u} - R_{-,d,s,u}, \qquad \Delta_{D,d,s,u} = R_{-,d,s,u} - R_{+,d,s,u}.$$

Positive values therefore indicate movement in the relation-predicted direction in both branches. With five normalized checkpoints,

$$Q_{\mathrm{traj},e,d,s} = \frac{1}{5}\sum_{u} \Delta_{e,d,s,u}.$$

For the full automated audits, these seed-level values are averaged over the 30 matched seeds in each domain,

$$Q_{\mathrm{traj},e,d} = \frac{1}{30}\sum_{s} Q_{\mathrm{traj},e,d,s},$$

leaving 24 domain values per branch for inference. The direct branch contrast is formed domainwise,

$$B_{\mathrm{traj},d} = Q_{\mathrm{traj},A,d} - Q_{\mathrm{traj},D,d}.$$

Early, late, and terminal summaries use the same matched-pair orientation. The attenuation contrast is defined so that positive values indicate decline,

$$C_{\mathrm{atten},e} = Q_{\mathrm{early},e} - Q_{\mathrm{late},e}.$$

A nonsignificant terminal effect is not interpreted as equivalence to zero.

The independent Gemini-3.6 Flash audit uses exactly the same 14,400 Llama audit identifiers, orientation, within-domain aggregation, and domain-level estimands as the primary GPT-5.6 Sol audit. Its inferential calculations are therefore methodologically identical, but its status is post-hoc measurement robustness rather than part of the frozen primary Llama audit.

For the external-human calibration subset, there is exactly one matched seed per domain–branch cell. The five checkpoint differences for that seed are averaged to obtain one $Q_{\mathrm{traj},e,d}$ directly, after which the same 24-domain sign-flip and domain-bootstrap procedures are used. The matched 480-item automated-judge comparisons in Appendix D.7 are recomputed on exactly the same audit identifiers; they are measurement comparisons rather than replacements for the full 30-seed automated analyses.

B.8 Bootstrap intervals and post-hoc uncertainty analyses

Bootstrap intervals supplement, rather than replace, the exact sign-flip tests. Within-domain aggregation is completed first, and semantic domains are then resampled with replacement. The reported Llama behavioral analysis uses 100,000 domain bootstrap resamples with random seed 20260918 and reports percentile 95% intervals. Representation-level and relation-condition intervals likewise preserve the domain as the resampling unit; tokens, candidate positions, dose observations, seeds, and individual judge scores are never treated as independent bootstrap observations.

The same domain-level principle is used for the post-hoc Gemini and external human calibration intervals. Because the human subset contains one matched seed per domain–branch cell, its uncertainty reflects cross-domain variation and that sparse seed realization; it does not estimate human inter-rater variability.

Fixed-winner Gaussian comparison. For the observed discovery-selected Gaussian, the simplest held-out uncertainty analysis keeps that winner and its discovery-selected global orientation fixed, paired-resamples the 24 novel domains, and recomputes

$$\Delta J_A = J_A(v_{EY}) - J_A(g^*), \qquad \Delta J_D = J_D(v_{EY}) - J_D(g^*),$$

and

$$\Delta W^{\mathrm{loc}} = W^{\mathrm{loc}}(v_{EY}) - W^{\mathrm{loc}}(g^*).$$

This quantifies novel-domain sampling uncertainty conditional on the originally selected Gaussian.

Selection-aware Gaussian bootstrap. The stronger post-hoc analysis additionally propagates instability in the best-of-1,024 discovery search. It conditions on the archived bank of 1,024 isotropic Gaussian candidates and on the frozen full-discovery v_{EY}. We use $B = 1{,}000{,}000$ two-stage bootstrap replicates with random seed 20260925. For replicate b:

1. resample the 24 discovery domains with replacement;

2. for each Gaussian candidate g, choose one global sign $\eta_g^{(b)} \in \{-1, +1\}$ that maximizes the resampled discovery

$$W_g^{\mathrm{loc},(b)} = \min\Big(\overline{J}_A^{(b)}(\eta_g^{(b)} g), \overline{J}_D^{(b)}(\eta_g^{(b)} g)\Big);$$

3. select

$$g_b^* = \arg\max_g W_g^{\mathrm{loc},(b)};$$

4. independently resample the 24 novel domains with replacement;

5. evaluate the frozen v_{EY} and the reselected, oriented $\eta_{g_b^*}^{(b)} g_b^*$ on the same novel-domain resample and recompute $\Delta J_A^{(b)}$, $\Delta J_D^{(b)}$, and $\Delta W^{\mathrm{loc},(b)}$.

Percentile 95% intervals are taken from these bootstrap distributions. The factorial direction is not reconstructed within each replicate: the analysis conditions on the frozen full-discovery v_{EY} and propagates uncertainty in the discovery-side Gaussian winner together with held-out novel-domain sampling. This procedure is a post-hoc uncertainty analysis, not a prospectively specified selection-adjusted hypothesis test. Results are reported in Appendix E.1.

B.9 Unsteered relation-condition diagnostic

The archived finite-dose grids include $\alpha = 0$ for every domain, branch, and relation condition. For relation condition r, the post-hoc baseline diagnostic defines

$$M_A^{(0)}(r) = M_A(0, r), \qquad M_D^{(0)}(r) = M_D(0, r),$$

$$I^{(0)}(r) = M_A^{(0)}(r) - M_D^{(0)}(r).$$

This baseline preference interaction must not be confused with the finite-dose steering effect $I = S_A + S_D$. Under a strongly reversed baseline preference, the joint branch criterion would be

$$M_A^{(0)} < 0, \qquad M_D^{(0)} > 0.$$

Baseline summaries and any reported relation-condition contrasts use the 24-domain inferential unit and the same sign-flip/bootstrap principles described above. The analysis uses only archived $\alpha = 0$ rows and is interpretive; it does not alter the frozen intervention hypotheses or endpoints. Condition means and the explicit-reversal diagnostic are reported in Appendix E.3.

B.10 Cross-model conventions

Qwen and Llama hidden states occupy different representation spaces. We therefore do not compute or interpret raw cross-model cosine similarity between their relational axes, and absolute intervention slopes are not treated as calibrated across architectures. Cross-model comparisons are therefore limited to whether corresponding within-model criteria are satisfied and to qualitative ordering; effect magnitudes are not compared across model families. In particular, Figure 2 should not be read as a calibrated cross-model magnitude comparison.

B.11 Numerical records

Reported summaries are computed from saved machine-readable outputs rather than transcribed intermediate calculations. Artifact classes, frozen identities, and provenance records are described in Appendix G.

C Discovery separation and prospective freezes

Both model families use the same separation between discovery-stage localization and novel-domain evaluation. For each family, the complete discovery stimulus set was defined and frozen before any model-targeted layerwise intervention result was inspected, and its wording was not revised in response to the scan. Discovery domains are used to construct the factorial directions and to

select one intervention layer per model. The selected layer and the axes reconstructed from the full discovery set are then frozen before evaluation on mutually disjoint novel domains.

Because the discovery domains contribute to both direction construction and layer localization, selected-layer discovery statistics are descriptive. Novel-domain results are out of sample with respect to those choices. The Llama study additionally follows a formal prospective protocol for the representation-level replication and for the scientific endpoints of the later behavioral audit. One qualification is important: the final behavioral intervention magnitude $|\alpha| = 1.5$ was selected in a small exploratory pilot and was therefore not prospectively prespecified. It was frozen only after that pilot and then evaluated on 30 new generation seeds. This appendix records that chronology explicitly so that prospectively specified evaluation, exploratory hyperparameter selection, held-out evaluation, and later robustness analyses are not conflated.

C.1 Common discovery-only layer localization

Both model families use the discovery-only localization procedure defined in Appendix B.3. At each candidate block, discovery domains are evaluated with leave-one-domain-out relational axes and symmetric $\epsilon = 0.10$ interventions. The selected block maximizes the weaker of the two mean branch-oriented finite-difference responses. This common rule selects block 22 for Qwen and block 13 for Llama (Figure 3).

After localization, v_{EY}, v_Y, and v_E are reconstructed at the selected block from all 24 discovery domains. The selected block and these full-discovery axes are then frozen before novel-domain evaluation. Discovery finite-dose results continue to use leave-one-domain-out axes, whereas novel results use the corresponding frozen full-discovery axes. No novel-domain, relation-control, or behavioral result contributes to layer selection or axis construction.

The localization finite differences $\widehat{J}^{(\epsilon)}$ are distinct from the exact local directional derivatives J used for specificity and from the finite-dose OLS effects S used for the primary representation-level results.

C.2 Llama: what the prospective protocol fixed

Beyond the common discovery-only localization procedure of Appendix C.1, the Llama study followed a prospective protocol that was frozen before any Llama target outcome was inspected. The protocol fixed the model family and revision; the discovery and novel datasets; the tokenizer and consequence-span extraction procedure; the raw-mean-then-normalize axis construction; the discovery-only layer-localization rule; the RMS-scaled residual-stream intervention; the *teacher-forced* finite-dose grid and regression procedure; the domain-level inferential unit; the primary and secondary representation-level hypotheses; the control families; and the gate governing entry into behavioral generation.

The behavioral-audit checkpoint rule, prefix-censoring rule, scoring rubric, matched contrasts, estimands, and planned tests were also specified before the reported 30-seed behavioral experiment. The behavioral intervention magnitude was the exception: $|\alpha| = 1.5$ was selected subsequently in the exploratory pilot described in Appendix D.1. We therefore do not describe the behavioral dose itself as prospectively prespecified.

Applying the common localization rule to the frozen Llama discovery set selects decoder block 13 (Figure 3). After this selection, block 13 and the full-discovery factorial axes reconstructed at that block were frozen. They were not revised in response to selected-layer discovery finite-dose results, direction controls, relation controls, novel-domain results, or behavioral outcomes.

The prospective protocol further specified the novel-domain representation-level gate before behavioral generation. The primary gate required a positive mean relational component, $\mathbb{E}_d[I_d] > 0$, tested by one-sided exact sign-flip inference over the 24 novel domains. Behavioral generation was opened only after this gate passed. The strict branchwise criterion and the $G > 0$ branch-asymmetry prediction were prespecified secondary tests and did not determine entry into the behavioral stage.

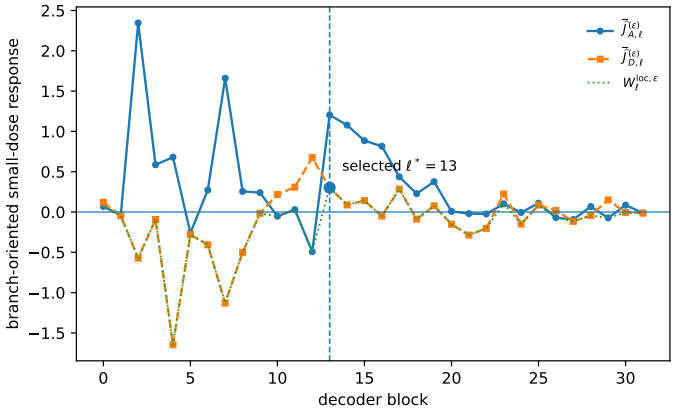


Figure 3: Discovery-only layer localization in Llama using symmetric small-dose interventions at $\epsilon = 0.10$. The common selection statistic $W_\ell^{\mathrm{loc},\epsilon} = \min(\widehat{\bar{J}}^{(\epsilon)}_{A,\ell}, \widehat{\bar{J}}^{(\epsilon)}_{D,\ell})$ evaluates the weaker of the two mean branch-oriented responses and selects decoder block 13. The selected block and full-discovery axes are subsequently frozen; control, novel-domain, and behavioral results do not contribute to the selection.

C.3 Role of the controls

The direction controls and relation conditions were supportive falsification tests rather than gates. Their outcomes were not permitted to change the selected layer, redefine v_{EY}, alter the primary novel-domain statistic, or change entry into the behavioral stage. They therefore constrain interpretation without redefining any frozen endpoint; the later $\alpha = 0$ relation-condition analysis is likewise an interpretive diagnostic (Section 5).

C.4 Behavioral dose pilot and final behavioral-stage freeze

After the representation-level gate passed, the behavioral magnitude was chosen in an exploratory 10-seed pilot comparing $|\alpha| \in \{0.5, 1.0, 1.5\}$ with a weakest-branch criterion. The selected magnitude, $|\alpha| = 1.5$, was then frozen and evaluated on 30 new generation seeds with no subsequent dose adjustment. The behavioral dose is therefore a pilot-selected hyperparameter evaluated on new seeds, not a dose prespecified before behavioral piloting. The behavioral endpoints and tests used in the final run were otherwise fixed before the reported 30-seed evaluation; full generation settings and dose provenance are given in Appendix D.1.

C.5 Primary blind-audit barrier

After generation, the primary Llama audit was conducted on a blind packet of 14,400 prefix-censored items. Intervention and matching metadata were held in a separate researcher key, which was not joined to the annotations until all expected audit identifiers had valid responses. The operational execution policy was frozen before annotation; full packet construction and validation procedures are documented in Appendix D.4.

C.6 Technical completion amendment

Five of the 14,400 primary-audit items initially failed only the exact-substring validation for the evidence-span field. Before any amended request, a narrow completion rule was frozen: for an otherwise schema-valid response with this sole failure, the evidence span could be replaced by the schema-permitted value `NONE`, while the audit identifier, system-state score, confidence, reason, and other scientific fields were preserved. No response was retried, selected, or modified on the basis of its behavioral score. The researcher key was joined only after all 14,400 identifiers satisfied the completeness requirement.

Table 5: Inferential status of the principal analyses. Layer localization is discovery-only in both families. The Llama representation-level gate and behavioral endpoints/tests were prospectively specified, whereas the final behavioral dose was selected exploratorily and frozen before evaluation on new generation seeds. Later judge, human, Gaussian-bootstrap, unsteered, and baseline-conditioned analyses are robustness or interpretive diagnostics.

Analysis	Status	What it supports
Qwen discovery intervention	Discovery characterization	Characterizes the writable relational effect at the discovery-selected layer.
Qwen novel-domain transfer	Held-out evaluation	Tests transfer of I; the strict distortion-branch criterion does not pass.
Qwen trajectory audit	Diagnostic	Applies the later prefix-censored trajectory measure to the earlier generations.
Llama discovery localization	Discovery-only	Selects block 13; selected-layer discovery statistics remain descriptive.
Llama novel $I > 0$	Prospective primary	Tests held-out aggregate transfer and gates entry into the behavioral stage.
Llama strict $S_A, S_D > 0$	Prospective secondary	Tests whether both held-out branches independently have the predicted orientation.
Llama $G > 0$	Prospective secondary	Alignment-dominance prediction; it fails, with $G < 0$ observed.
Direction/relation controls	Supportive falsification	Constrain interpretation without changing the selected layer, direction, or gate.
Behavioral dose pilot	Exploratory hyperparameter selection	Compares $\lvert\alpha\rvert \in \{0.5, 1.0, 1.5\}$ over 10 pilot seeds using a weakest-branch criterion; selects 1.5.
Llama final trajectory IUT	Frozen endpoints; new-seed evaluation	Tests positive trajectory-average realization in both branches using 30 new seeds after dose freeze.
Llama $B_{\mathrm{traj}} > 0$	Frozen endpoint; new-seed evaluation	Directly tests alignment-dominant behavioral asymmetry in the final run.
Llama attenuation / terminal summaries	Frozen behavioral analyses	Distinguish early expression from later persistence under the prefill-only intervention.
Gemini full rescoring	Post-hoc measurement robustness	Tests whether the principal behavioral ordering and trajectory pattern depend on the primary automated judge.
Human 480-item subset	Post-hoc measurement calibration	Tests whether the overall behavioral direction is visible under an external human rater; not a full trajectory replication.
Gaussian selection-aware bootstrap	Post-hoc uncertainty analysis	Propagates discovery-side winner-selection and novel-domain sampling uncertainty for the specificity comparison.
Unsteered $\alpha = 0$ reversal analysis	Post-hoc interpretive diagnostic	Tests whether the controlled baseline consequence-preference readout itself jointly reverses.
Baseline-conditioned steering diagnostic	Post-hoc interpretive diagnostic	Conditions archived finite-dose branch slopes on contemporaneous $\alpha = 0$ preferences; tests the simple baseline-amplification account, with mean-token scoring as a length-normalized sensitivity analysis.
Transport/readout geometry	Mechanistic diagnostic	Tests local explanations for branch asymmetry; not primary evidence for the behavioral dissociation.

C.7 Post-primary robustness and interpretive diagnostics

Several analyses were conducted after the primary Llama audit and are therefore reported as robustness or interpretive diagnostics rather than as part of the prospectively specified endpoint set. These include the complete independent Gemini rescoring of the frozen blind packet, the 480-item blinded human calibration subset, the selection-aware best-of-1,024 Gaussian bootstrap, and the post-hoc relation-grid diagnostics. The latter comprise both the explicit analysis of the already archived $\alpha = 0$ relation-condition rows and the baseline-conditioned steering diagnostic in Appendix E.4. The baseline-conditioned analysis uses only archived finite-dose relation grids, relates each branch-oriented intervention slope to its contemporaneous $\alpha = 0$ preference, and treats mean-token scoring as a continuation-length sensitivity analysis. It involves no new model inference and does not change any selected layer, direction, frozen endpoint, or planned test. These procedures and results are reported in Appendices D.6, D.7, and the corresponding control-results sections.

C.8 Inferential status of the principal analyses

Table 5 summarizes the inferential status of the principal analyses, distinguishing discovery characterization, held-out evaluation, prospectively specified endpoints, exploratory hyperparameter selection, and post-hoc robustness or interpretive diagnostics.

Table 6: Recorded freeze points for the Llama representation-level replication and final-run behavioral audit, in execution order. The final behavioral-design freeze occurs after exploratory selection of $|\alpha| = 1.5$. The chronology is based on the study's internal records; Table 18 lists content hashes separately for artifact identification.

Freeze point	Covers
Protocol freeze	hypotheses, estimands, gates
Replication manifest	binds stimuli, model revision, primary judge artifacts
Stimulus freeze	discovery / relation-control / novel
Primary judge artifacts	system prompt and output schema
Axis and layer freeze	v_{EY}, v_Y, v_E at ℓ^*
Final behavioral design freeze	post-pilot dose, seeds, schedule, sampling; 2,880 generations
Primary audit execution policy	judge policy, validation rules
Primary blind packet	14,400 prefix-censored items
Primary researcher key	identifier-to-condition mapping
Amendment freeze	evidence-span completion rule

C.9 Freeze manifests

The study records identify the principal freeze points for the Llama representation-level protocol and primary behavioral audit. Table 6 reports the author-recorded execution order. The final behavioral-design freeze occurred after the exploratory dose pilot; later Gemini, human-calibration, Gaussian-bootstrap, $\alpha = 0$, and baseline-conditioned analyses are post-primary analyses and are not part of this freeze sequence.

The chronology reported here is based on internal study records rather than an independently timestamped preregistration. Accordingly, the table documents the claimed execution order but is not presented as cryptographic evidence of when each freeze occurred. Content hashes are retained only in Appendix G to identify the exact archived artifact versions.

D Behavioral generation and blind trajectory audit

This appendix documents the behavioral-generation and measurement pipeline behind Section 4. The behavioral stage asks whether an intervention effect observed under the controlled teacher-forced readout is expressed when the model generates its own continuation after the prefill intervention has stopped. We separate exploratory dose selection, the final matched generation experiment, blind trajectory scoring, and later measurement-robustness analyses.

D.1 Behavioral dose selection and generation protocol

For Llama, behavioral generation was opened only after the prospectively frozen novel-domain representation-level gate passed (Appendix C.2). The behavioral magnitude itself was not prospectively prespecified. Before the reported 30-seed generation experiment, we ran a 10-seed exploratory pilot comparing $|\alpha| \in \{0.5, 1.0, 1.5\}$. The pilot was used only to choose the magnitude of the later prefill intervention.

Weakest-branch dose criterion. Dose selection did not maximize the alignment branch alone. For each candidate magnitude, the two branchwise matched-sign effects were compared and the dose with the largest effect on the weaker branch was preferred. None of the three pilot magnitudes satisfied the full branchwise intersection–union criterion. Among the tested values, $|\alpha| = 1.5$ produced the largest weakest-branch pairwise effect and was selected. The dose was then frozen. The reported Llama behavioral experiment used 30 new seeds, $520010, \ldots, 520039$, none of which was used in the pilot, and the dose was not adjusted further. Thus the reported behavioral evaluation is held out with respect to generation seed after dose selection, but the dose itself is a pilot-selected hyperparameter rather than a prospectively prespecified quantity.

Table 7: Behavioral-generation settings for the reported 30-seed evaluation. For Llama, $|\alpha| = 1.5$ was selected in the 10-seed exploratory pilot and then frozen before the reported run. Qwen uses the same generation schedule at its discovery-selected block 22.

Parameter	Value
Llama intervention layer	decoder block 13
Qwen intervention layer	decoder block 22
Direction	frozen full-discovery v_{EY}
Reported coefficients	$\{-1.5, +1.5\}$
Final seeds	520010–520039 (30 per domain–branch cell)
Sampling	enabled, temperature 0.8, top-p 0.95
Repetition penalty	1.0
New tokens	minimum 1, maximum 192
KV cache	enabled
Temporal schedule	prompt prefill only

The final behavioral prompts use the same 24 novel domains as the held-out representation-level experiment. Each domain is prompted separately with the alignment and distortion premise. The prompt contains the scenario context, premise, and connector, but neither of the candidate consequences used in the teacher-forced evaluation, so open generation is not seeded with either teacher-forced outcome.

For every domain and branch, continuations are generated under $\alpha \in \{-1.5, +1.5\}$ with 30 seeds, yielding $24 \times 2 \times 2 \times 30 = 2{,}880$ continuations per model family. All planned continuations were retained; none was regenerated or excluded because of its semantic content, score, or apparent intervention direction. Table 7 records the final generation settings after the dose freeze.

Prefill-only intervention. At the frozen layer, all prompt-token residual states receive the same RMS-scaled intervention used in the teacher-forced experiments,

$$h_{i,t}^{(\ell^*)} \leftarrow h_{i,t}^{(\ell^*)} + \alpha s_i v_{EY},$$

with the per-prompt scale s_i defined in Appendix B and Equation 5. The intervention is applied during prompt prefill only and is not reinjected during decoding. Execution logs record this schedule for the reported generations. The experiment therefore measures behavior after a one-shot prefill intervention, not behavior under continuous steering.

Matched generation design. Positive- and negative-dose generations are matched on domain, epistemic branch, seed, prompt text, intervention layer and direction, prompt RMS scale, and sampling configuration. Each domain–branch–seed cell therefore contains an exact -1.5 versus $+1.5$ pair. There are $24 \times 2 \times 30 = 1{,}440$ matched pairs per model family, and this pairing is preserved through trajectory construction and analysis.

D.2 Prefix-censored trajectory construction

Each retained continuation is evaluated at five normalized checkpoints, $u \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. For a continuation containing T generated tokens, the audit item at checkpoint u includes the original prompt and the first

$$k(u) = \min\{T, \max(1, \lceil uT \rceil)\} \tag{6}$$

generated tokens. Checkpoints are therefore defined relative to realized continuation length rather than to the maximum budget of 192 tokens. All five checkpoint indices are distinct for every observed continuation. Each of the 2,880 generations contributes five separately scored prefixes, yielding 14,400 audit items per model family.

Prefix censoring prevents later text from influencing the evaluation of an earlier state. At each checkpoint, the rater sees only the prompt and the visible prefix; subsequent tokens are unavailable. A repair or reversal introduced later in the continuation therefore cannot retrospectively change the score assigned to an earlier prefix. Because checkpoints are normalized by realized length, however, the same value of u need not correspond to the same absolute token count across continuations.

D.3 BLIND PACKET, RESEARCHER KEY, AND RUBRIC

The automated-judge packet contains three fields: audit identifier, prompt, and visible prefix. Epistemic branch, intervention sign, domain identity, seed, checkpoint fraction, generation identifier, matched-pair identifier, and expected orientation are held in a separate researcher key and are not supplied to the judge. The key is joined only after the relevant blind-scoring completeness barrier has been satisfied.

Systemic-state rubric. The rater scores realized change in systemic functioning at the end of the visible prefix, relative to the state established by the prompt, on a branch-neutral scale $R \in \{-2, -1, 0, +1, +2\}$: -2 denotes severe realized deterioration, -1 realized net deterioration, 0 no demonstrated net change, mixed, or unresolved state, $+1$ realized net improvement, and $+2$ strong or broad realized improvement. The prompt-established state is treated as zero, so merely restating or elaborating that state does not count as change.

The rubric explicitly separates realized outcomes from recommendations, warnings, plans, intentions, proposed remediation, and implementation attempts whose effects are not demonstrated. Remediation stage is recorded separately as $\{\text{none}, \text{proposed}, \text{initiated}, \text{implemented}\}$ and never determines the systemic-state score. Automated annotations also record a confidence label, evidence span, and brief reason; only R enters the behavioral estimands.

D.4 PRIMARY AUTOMATED AUDIT AND COMPLETENESS BARRIER

The primary automated trajectory audit used GPT-5.6 Sol. Its operational policy was frozen before the first real blind request and specified append-only output, first-valid-response acceptance, and resumption from already valid identifiers. A synthetic pre-flight item was used to check endpoint compatibility and local schema validation without exposing a real audit item. Reruns were limited to schema or validation failures under the frozen execution policy and were not triggered by the behavioral score.

Scientific unblinding was withheld until every blind identifier had a locally valid response. Five of the $14{,}400$ Llama items persistently failed only the evidence-span substring validation. The narrow completion amendment in Appendix C.6 replaced only that field with the schema-valid sentinel `NONE`; it changed no systemic-state score or other scientific field. The researcher key was joined only after the audit reached $14{,}400/14{,}400$ valid records.

For Qwen, the same prefix-censored instrument is diagnostic because it was developed after the original Qwen free-generation experiment. For Llama, the checkpoint rule, censoring rule, rubric, matched contrasts, estimands, and planned tests were fixed before the reported 30-seed generation, while the later packet-construction and execution policy was frozen before annotation (Appendix C.4).

D.5 BEHAVIORAL ESTIMANDS AND INFERENCE

Orientation, aggregation, and inference follow Appendix B.7. Matched $+\alpha/-\alpha$ score differences are oriented so that positive values denote movement in the relation-predicted direction in both branches. The five checkpoints are averaged within matched seed pair and then over the 30 seeds within each domain, leaving 24 domain-level observations per branch. The reported summaries are the trajectory average Q_{traj}, the direct branch contrast B_{traj}, the early and terminal effects, and the attenuation contrast C_{atten}. Directional decisions use exact sign-flip inference over semantic domains, and bootstrap intervals resample domains after within-domain seed aggregation. A non-significant terminal effect is not interpreted as equivalence to zero.

D.6 INDEPENDENT AUTOMATED-JUDGE ROBUSTNESS

After the primary Llama audit, the complete 14,400-item blind packet was rescored with Gemini-3.6 Flash as a post-hoc measurement-robustness analysis. The same prompt and prefix-censoring construction was used, and the researcher key was again withheld until all $14{,}400$ audit identifiers had valid Gemini scores. This second audit did not alter the frozen scientific endpoints or the status of the primary GPT-5.6 Sol tests.

Table 8: Llama behavioral results under the primary GPT-5.6 Sol audit and the post-hoc Gemini-3.6 Flash robustness audit. Values are domain-level oriented effects; p-values are one-sided exact sign-flip tests in the prespecified positive direction where applicable.

Estimand	GPT-5.6 Sol	Gemini-3.6 Flash
$Q_{\mathrm{traj},A}$	0.2014 ($p = 0.000339$)	0.1650 ($p = 0.000347$)
$Q_{\mathrm{traj},D}$	0.0500 ($p = 0.048806$)	0.0758 ($p = 6.23 \times 10^{-5}$)
B_{traj}	0.1514 ($p = 0.0094$)	0.0892 ($p = 0.02065$)
$Q_{\mathrm{early},D}$	0.0750 ($p = 0.00188$)	0.1167 ($p = 5.96 \times 10^{-7}$)
$C_{\mathrm{atten},D}$	0.0563 ($p = 0.0328$)	0.0681 ($p = 2.76 \times 10^{-4}$)
$Q_{\mathrm{terminal},D}$	-0.0028 ($p = 0.542$)	0.0403 ($p = 0.0593$)
$Q_{\mathrm{terminal},A}$	0.2528 ($p = 0.000584$)	0.1583 ($p = 0.000860$)

Table 8 compares the main trajectory estimands. The second judge reproduces the two positive trajectory-average branch effects, the alignment-dominant direct contrast, distortion-side attenuation, and the pattern in which alignment remains detectably positive at the terminal checkpoint while distortion is not detectably positive there.

For Gemini, the percentile 95% intervals are $[0.0836, 0.2483]$ for $Q_{\mathrm{traj},A}$, $[0.0431, 0.1117]$ for $Q_{\mathrm{traj},D}$, $[0.0111, 0.1706]$ for B_{traj}, $[0.0354, 0.1021]$ for $C_{\mathrm{atten},D}$, and $[-0.0042, 0.0861]$ for $Q_{\mathrm{terminal},D}$. The primary audit gives $B_{\mathrm{traj}} = 0.1514$ with 95% CI $[0.039, 0.279]$, $C_{\mathrm{atten},D} = 0.0563$ with 95% CI $[0.003, 0.113]$, and $Q_{\mathrm{terminal},D} = -0.0028$ with 95% CI $[-0.078, 0.072]$.

The judges are not identical measurement instruments. Across the full 14,400 Llama items, exact score agreement is 78.0% and quadratic-weighted Cohen's $\kappa = 0.685$. The neutral-score rate is 65.8% under GPT-5.6 Sol and 73.6% under Gemini, illustrating a calibration difference even though the principal domain-level contrasts and trajectory ordering are preserved.

D.7 External-human calibration subset

As a complementary measurement-validity analysis, one external human participant served as a blinded rater for a 480-item Llama subset. The subset spans all 24 novel domains, both epistemic branches, both intervention signs, and all five checkpoints, with exactly one matched seed per domain–branch cell:

$$24 \times 2 \times 2 \times 5 = 480.$$

The participant was informed of the annotation task and its research purpose, provided consent before participating, and was compensated for the annotation work. The rater evaluated synthetic prompts and model-generated prefixes only. No personal identifiers or sensitive personal information were collected; only the task annotations entered the calibration analysis.

The intervention metadata and researcher key remained hidden. Korean translations of the context, prompt, and visible prefix were available to and viewed by the rater for all 480 items. The same five-point systemic-state scale was used. Consequently, this analysis is a translation-assisted, single-rater calibration study rather than an English-only human replication.

The human scores support the average intervention direction in both branches:

$$Q_{\mathrm{traj},A} = 0.408, \qquad 95\%\ \mathrm{CI}\ [0.125, 0.708], \qquad p = 0.0070,$$
$$Q_{\mathrm{traj},D} = 0.308, \qquad 95\%\ \mathrm{CI}\ [0.033, 0.592], \qquad p = 0.0257.$$

The mean direct asymmetry is alignment-dominant, $B_{\mathrm{traj}} = 0.100$, but is imprecisely estimated (95% CI $[-0.383, 0.617]$; $p = 0.364$). The human terminal effects are $Q_{\mathrm{terminal},A} = 0.708$ and $Q_{\mathrm{terminal},D} = 0.708$, both positive in this subset. Thus the human calibration supports the overall branchwise behavioral direction but does not reproduce the automated judges' distortion-side attenuation and weak terminal D effect.

To separate measurement effects from subset sampling, we also recomputed the automated-judge estimands on the identical 480 audit identifiers. Table 9 shows the same descriptive ordering, $Q_{\mathrm{traj},A} > Q_{\mathrm{traj},D}$, for all three instruments on this subset, but the detailed persistence profile differs: the human distortion attenuation is negative, whereas both automated judges retain positive attenuation.

Table 9: Matched 480-item Llama calibration subset scored by the external human rater and by the two automated judges. Values use the same orientation and the same audit identifiers for all three instruments. The subset contains one seed per domain–branch cell, so these estimates are much less precise than the full 30-seed automated audits.

Estimand	Human	GPT-5.6 Sol	Gemini-3.6 Flash
$Q_{\mathrm{traj},A}$	0.408	0.467	0.350
$Q_{\mathrm{traj},D}$	0.308	0.233	0.117
B_{traj}	0.100	0.233	0.233
$Q_{\mathrm{early},D}$	0.188	0.229	0.167
$C_{\mathrm{atten},D}$	-0.271	0.042	0.104
$Q_{\mathrm{terminal},A}$	0.708	0.583	0.458
$Q_{\mathrm{terminal},D}$	0.708	0.125	0.083

Agreement on the identical 480 items is lower between the human rater and each automated judge than between the two automated judges. Human–GPT-5.6 Sol exact agreement is 41.7% with quadratic-weighted $\kappa = 0.404$, and human–Gemini exact agreement is 39.4% with $\kappa = 0.328$; GPT-5.6 Sol–Gemini exact agreement on the same subset is 73.5% with $\kappa = 0.692$. The neutral-score rates are 30.6%, 62.9%, and 70.4%, respectively, further illustrating that absolute score calibration depends on the measurement instrument.

These comparisons motivate the interpretation in Section 4: the positive trajectory-average direction in both branches is not specific to a single automated judge, whereas the finer distortion-side temporal profile is measurement-sensitive and should be treated as secondary.

D.8 Integrity checks

The final primary dataset contains all 2,880 planned generations per model family and all corresponding matched ± 1.5 pairs and five-checkpoint audit items, with no missing domain–branch–seed–sign cells or selective regeneration. The Gemini robustness audit covers the complete Llama blind packet, and the human calibration contains 480 completed ratings on the matched subset.

D.9 Descriptive properties of the audit data

Two properties of the generated and scored data matter for interpretation.

Generation truncation. Llama continuation lengths range from 9 to 192 tokens, with median 192 and mean approximately 188.1; 2,763 of 2,880 reach the cap. In Qwen, 1,971 of 2,880 reach the cap and the mean length is 165.3. The terminal checkpoint therefore often reflects the generation budget rather than a natural narrative ending, especially in Llama. Accordingly, Q_{terminal} is interpreted as the effect at the final observed budgeted checkpoint, not as a guaranteed endpoint of a completed narrative.

Score sparsity and calibration. Under the primary GPT-5.6 Sol audit, 65.8% of Llama and 74.4% of Qwen checkpoint judgments receive the neutral score 0; the full Llama Gemini zero rate is 73.6%. This sparsity contributes to small absolute score differences. Inference nevertheless follows the prespecified domain-level paired design rather than treating the 14,400 checkpoint judgments as independent observations. Matched-subset differences in score calibration across instruments are reported in Appendix D.7.

D.10 Scope of the behavioral evidence

The behavioral evidence has different inferential status across analyses. The reported Llama 30-seed trajectory endpoints and tests belong to the frozen final-run analysis after the pilot-selected dose was fixed; the dose magnitude itself was selected exploratorily. The complete Gemini rescoring and the human 480-item calibration subset were added subsequently as measurement-robustness analyses. For Qwen, the prefix-censored trajectory audit remains diagnostic because the instrument was developed after the original free-generation experiment.

The measurement evidence is also hierarchical. Two complete automated audits support positive Llama trajectory-average effects in both branches, an alignment-dominant behavioral ordering, and distortion-side attenuation toward a terminal value that is not detectably positive. The human subset also supports positive average effects in both branches, but it does not reproduce the distortion-side attenuation or weak terminal distortion effect. Because the human analysis uses one external rater, one matched seed per domain–branch cell, and translation-assisted annotation, it provides a calibration and measurement-validity check rather than inter-rater reliability or a fully powered human replication.

Finally, all behavioral results concern a one-shot prefill intervention with no decode-time reinjection. They therefore distinguish teacher-forced writability, behavioral expression, and persistence under this intervention schedule, but do not establish what would happen under repeated, adaptive, or continuous steering. The corresponding artifact classes and provenance records are described in Appendix G.

E SUPPLEMENTARY CONTROL RESULTS

This appendix collects the control analyses used to delimit the claims in Sections 3 and 5. Three questions are kept separate. First, do structured or randomized directions reproduce the finite-dose branch structure of the factorial direction? Second, if a Gaussian direction is explicitly selected to perform well on both discovery branches, how strongly does it transfer to held-out domains? Third, does the fixed intervention or the unsteered consequence-preference readout reorganize when the stated relation is removed or explicitly reversed? The corresponding analyses use different estimands and should not be conflated.

E.1 DIRECTION AND SPECIFICITY CONTROLS

Finite-dose structured and randomized controls. The finite-dose controls use the same selected layer, RMS scaling, dose grid, and OLS estimand as the representation-level results in Table 1. For Llama, the structured controls show why overall response magnitude is insufficient. The outcome-valence direction v_Y produces a very large alignment-side response but moves the distortion branch in the wrong relational orientation; v_E fails for the same reason. The randomized families are weak on average but occasionally produce parts of the required sign structure.

As Table 10 shows, for Llama, one balanced sign-randomized direction exceeds the factorial reference on aggregate I while having a wrong-sign distortion response. Conversely, individual projected Gaussian controls can satisfy the positive branchwise orientation at smaller effect size: the best weaker-branch score is $W = 0.239$, compared with $W = 0.2738$ for v_{EY}. Thus neither a large aggregate response nor the branchwise sign pattern is sufficient to identify the factorial direction uniquely.

Discovery-selected best-of-1,024 Gaussian challenge. The more adversarial Gaussian analysis in Table 2 uses exact RMS-scaled directional derivatives at $\alpha = 0$, not the finite-dose S values above. We draw 1,024 isotropic Gaussian directions without projecting out the factorial axes. For each candidate, one global orientation is chosen using discovery domains, and the direction maximizing

$$W^{\mathrm{loc}} = \min(J_A, J_D)$$

is selected. That selected and oriented direction is then applied unchanged to the novel domains.

In Llama, the selected Gaussian is nearly orthogonal to v_{EY} (cosine 0.027) yet has positive branch-oriented local derivatives on both discovery and novel domains. The sign criterion alone therefore does not identify v_{EY}. The difference is in held-out magnitude: on the novel domains, v_{EY} gives $J_A = 0.5653$, $J_D = 1.1430$, and $W^{\mathrm{loc}} = 0.5653$, whereas the discovery-selected Gaussian gives $J_A = 0.2164$, $J_D = 0.4573$, and $W^{\mathrm{loc}} = 0.2164$. The observed weaker-branch difference is consequently $\Delta W^{\mathrm{loc}} = 0.3488$.

Selection-aware held-out uncertainty. Because the Gaussian winner is itself selected on discovery domains, we also report the post-hoc selection-aware bootstrap defined in Appendix B.8. That procedure reruns discovery-side orientation and best-of-1,024 selection while paired-resampling the

Table 10: Compact finite-dose direction-control summary on discovery domains. For Llama, $I = S_A + S_D$, $G = S_A - S_D$, and $W = \min(S_A, S_D)$. The Gaussian-orthogonal controls are projected out of $\mathrm{span}(v_{EY}, v_Y, v_E)$. Randomized-family means are over 32 draws. These controls are distinct from the best-of-1,024 exact-local-gradient search below.

Direction / control	Key result
Llama	
v_{EY}	$S_A = 1.1454$, $S_D = 0.2738$, $I = 1.4191$, $W = 0.2738$
v_Y	$S_A = 4.6349$, $S_D = -3.6836$, $G = 8.3185$
v_E	$S_A = 0.6725$, $S_D = -0.2328$, $G = 0.9053$
Gaussian orthogonal	$\bar{I} = 0.0160$; best $W = 0.239$; $0/32$ reach W_{EY}
Balanced sign-randomized	$\bar{I} = 0.0589$; $1/32$ reach I_{EY}; best $W = 0.040$
Qwen	
v_{EY}, discovery	$I = 1.284$
Outcome-valence v_Y	$G \approx 4.977$
Gaussian orthogonal	$\bar{I} \approx 0.0058$
Balanced sign-randomized	$\bar{I} \approx 0.0208$

Table 11: Held-out Llama comparison between the frozen factorial direction and the discovery-selected best-of-1,024 Gaussian. "Fixed winner" resamples only the paired novel domains while retaining the originally selected Gaussian. "Selection-aware" additionally resamples discovery domains and reruns global orientation and best-of-1,024 selection in every replicate. Intervals are post-hoc percentile bootstrap 95% CIs.

Quantity	Point estimate	Fixed-winner 95% CI	Selection-aware 95% CI
ΔJ_A	0.349	[0.119, 0.585]	[0.115, 0.659]
ΔJ_D	0.686	[0.396, 0.980]	[0.374, 1.084]
ΔW^{loc}	0.349	[0.122, 0.584]	[0.134, 0.676]

held-out novel domains, thereby propagating winner-selection uncertainty into ΔJ_A, ΔJ_D, and ΔW^{loc}.

As Table 11 shows, the selection-aware interval for ΔW^{loc} excludes zero, as do the branchwise intervals. Winner identity changes under discovery-domain resampling, but the held-out advantage remains positive under the specified selection-aware procedure. This is a post-hoc uncertainty analysis, not a prospectively specified selection-adjusted test; it conditions on the archived bank of 1,024 Gaussian directions and does not exhaust alternative data-driven searches. For Qwen, the corresponding exact-local-gradient point comparison is $W^{\mathrm{loc}} = 0.615$ for v_{EY} versus 0.132 for the best of 1,024 Gaussian draws; no selection-aware Qwen interval is reported.

E.2 Relation-condition intervention controls

The relation controls are supporting falsification tests performed without changing the selected intervention layer, relational direction, RMS scaling, or finite-dose evaluation rule. The noncausal condition removes the stated linkage while retaining the epistemic and outcome content. The explicit reversal condition instead specifies a countervailing mechanism under which distortion can support favorable consequences and alignment unfavorable ones. Neither control constructs a new axis: for each model, the same default-relation discovery direction is reused across all three relation conditions.

Table 12 reports the complete branch-level finite-dose results. Here S_A and S_D retain the *default* relational orientation, so a mechanism-sensitive reversal account predicts a joint sign reversal under the explicit-reversal condition.

Removing the stated linkage. In Llama, the aggregate relational effect decreases from $I_{\mathrm{default}} = 1.4191$ to $I_{\mathrm{noncausal}} = 0.3107$, a paired reduction of 1.1084 (95% bootstrap CI $[0.7713, 1.4625]$;

Table 12: Relation-condition finite-dose intervention effects on the 24 discovery domains. For each model, the selected layer, relational direction, and evaluation rule are fixed across conditions. $I = S_A + S_D$ and $G = S_A - S_D$.

Relation condition	S_A	S_D	I	G
Llama				
Default causal	1.1454	0.2738	1.4191	0.8716
Noncausal	0.1572	0.1535	0.3107	0.0036
Explicit reversal	0.7732	0.7202	1.4934	0.0530
Qwen				
Default causal	0.5426	0.7414	1.2840	−0.1987
Noncausal	−0.0104	0.4452	0.4348	−0.4556
Explicit reversal	−0.1974	0.5473	0.3499	−0.7448

one-sided exact sign-flip $p = 1.6 \times 10^{-6}$). Expressed as noncausal-minus-default, the branch changes are $\Delta S_A = -0.9882$ (95% CI $[-1.2480, -0.7304]$) and $\Delta S_D = -0.1202$ (95% CI $[-0.3658, 0.1325]$). Thus most of the Llama aggregate reduction is carried by the alignment branch, and the default branch imbalance $G = 0.8716$ is nearly eliminated under the noncausal condition ($G = 0.0036$).

In Qwen, I decreases from 1.2840 to 0.4348, a paired reduction of 0.8492 (95% CI $[0.5334, 1.1681]$; $p = 2.1 \times 10^{-5}$). Both oriented branch effects decrease: $\Delta S_A = -0.5530$ (95% CI $[-0.7783, -0.3169]$) and $\Delta S_D = -0.2962$ (95% CI $[-0.5824, -0.0117]$). Removing the stated linkage therefore weakens the finite-dose relational response in both model families, although not by a uniform multiplicative factor across branches.

Explicit causal reversal. Explicit reversal does not produce the joint branchwise sign reversal required by a mechanism-sensitive rule-recomputation account. In Llama, both default-oriented branch effects remain positive, $S_A = 0.7732$ and $S_D = 0.7202$, giving $I_{\text{reversal}} = 1.4934$. Relative to default, $\Delta S_A = -0.3721$ (95% CI $[-0.6553, -0.0969]$) while $\Delta S_D = +0.4465$ (95% CI $[0.2024, 0.6819]$). These changes largely cancel in the aggregate: $I_{\text{reversal}} - I_{\text{default}} = +0.0743$ (95% CI $[-0.2244, 0.3744]$), and the prespecified one-sided directional test does not support the predicted decrease ($p \approx 0.681$).

Qwen attenuates more strongly but still does not jointly reverse. The alignment branch changes sign ($S_A = -0.1974$), whereas the distortion branch remains positive in the default orientation ($S_D = 0.5473$), yielding $I_{\text{reversal}} = 0.3499$. The aggregate default-to-reversal reduction is 0.9341 (95% CI $[0.5861, 1.2953]$; $p = 1.4 \times 10^{-5}$). The reversal response is also close to the noncausal response: $I_{\text{reversal}} - I_{\text{noncausal}} = -0.0849$ (95% CI $[-0.2860, 0.1099]$; $p = 0.209$).

The two models therefore differ in how much explicit reversal changes the aggregate magnitude, but share the structurally important result: neither exhibits the joint branchwise sign reversal predicted by a rule whose sign is recomputed from the stated mechanism.

E.3 Unsteered relation-condition preferences

The archived finite-dose grids include $\alpha = 0$ for every domain, branch, and relation condition. We use these rows for the post-hoc baseline diagnostic defined in Appendix B.9; it is interpretive and does not construct the relational direction, select the intervention layer, or alter any frozen intervention endpoint. Table 13 reports the equal-domain means. The baseline interaction $I^{(0)}$ is distinct from the finite-dose steering effect $I = S_A + S_D$.

The aggregate explicit-reversal baseline does not jointly reverse. Under the manuscript-primary total-log-likelihood score, the strong-reversal criterion is $M_A^{(0)} < 0$ and $M_D^{(0)} > 0$ (Appendix B.9). Neither model satisfies this criterion at the aggregate level. In Llama, the explicit-reversal means are $M_A^{(0)} = 2.710$ (95% CI $[1.050, 4.374]$) and $M_D^{(0)} = -3.265$ (95% CI $[-4.936, -1.573]$). In Qwen they are $M_A^{(0)} = 0.728$ (95% CI $[-0.894, 2.292]$) and $M_D^{(0)} = -3.839$ (95% CI $[-5.563, -2.155]$). Thus explicit reversal weakens the alignment-side favorable preference

Table 13: Unsteered ($\alpha = 0$) teacher-forced consequence preferences across relation conditions. Values are equal-domain means over the 24 discovery domains. $I^{(0)} = M_A^{(0)} - M_D^{(0)}$ is a baseline preference interaction and should not be confused with the finite-dose intervention effect $I = S_A + S_D$ in Table 12.

Relation condition	$M_A^{(0)}$	$M_D^{(0)}$	$I^{(0)}$
Llama			
Default causal	7.5588	−1.6857	9.2445
Noncausal	5.0129	1.6412	3.3716
Explicit reversal	2.7098	−3.2649	5.9746
Qwen			
Default causal	6.9263	−1.8683	8.7946
Noncausal	5.2491	1.1042	4.1449
Explicit reversal	0.7283	−3.8387	4.5671

but does not install the jointly opposite baseline pattern in either model under the primary scoring rule. The noncausal condition also changes the baseline means (Table 13), but those changes are not used as an additional primary test.

The aggregate baseline result limits what can be inferred from the aggregate intervention reversal alone: because neither the unsteered readout nor the fixed-direction response exhibits the required joint reversal, the failure of the aggregate steering pattern cannot by itself be localized uniquely to the direction. It does not, however, make the archived relation-control grid uninformative about baseline amplification. Appendix E.4 therefore conditions the intervention slopes directly on the corresponding branch-level $\alpha = 0$ preferences, including cases in which the unsteered preference already points against the default orientation. That post-hoc diagnostic sharpens the direction-specific comparison while remaining limited to the present teacher-forced readout; it does not establish whether the models understand the countervailing mechanism in a more general sense.

E.4 Baseline-conditioned steering diagnostic

The unsteered analysis above raises a further interpretive question. The finite-dose intervention could in principle act mainly by amplifying a consequence preference already present at $\alpha = 0$, rather than by adding a premise-conditioned relational component that is separable from that baseline preference. We therefore performed a post-hoc baseline-conditioned analysis of the archived relation-control grids. This analysis was not part of the frozen primary protocol and does not alter any reported intervention endpoint, layer choice, direction construction, or relation-control test.

For branch $e \in \{A, D\}$, define

$$q_A = +1, \qquad q_D = -1,$$

and orient both the unsteered margin and the finite-dose slope into the default relational convention:

$$C_{d,e,r}^{(0)} = q_e M_{d,e,r}^{(0)}, \qquad S_{d,e,r} = q_e \beta_{d,e,r}. \tag{7}$$

Thus $C^{(0)} > 0$ means that the unsteered consequence preference agrees with the default alignment–favorable / distortion–unfavorable orientation, whereas $C^{(0)} < 0$ means that the unsteered preference points in the opposite direction. Positive S denotes a default-oriented intervention effect, as elsewhere in the paper.

Branch-adjusted diagnostic model. Because the distributions of $C^{(0)}$ differ substantially between the alignment and distortion branches, a pooled intercept can confound baseline dependence with branch-level differences. We therefore use the branch-adjusted model

$$S_{d,e,r} = a_{A,r}\mathbf{1}[e = A] + a_{D,r}\mathbf{1}[e = D] + \lambda_r C_{d,e,r}^{(0)} + \varepsilon_{d,e,r}. \tag{8}$$

The coefficient λ_r measures covariation between the contemporaneous unsteered preference and intervention magnitude, while $a_{A,r}$ and $a_{D,r}$ summarize the fitted branch-specific response at

Table 14: Post-hoc baseline-conditioned steering diagnostic using the manuscript-primary total-log-likelihood margin. The branch-adjusted model is $S = a_A\mathbf{1}[A] + a_D\mathbf{1}[D] + \lambda C^{(0)} + \epsilon$. Intervals are 95% percentile semantic-domain cluster-bootstrap intervals from 100,000 resamples. Conflict counts are descriptive $n(S > 0)/n(C^{(0)} < 0)$ and are not treated as independent-binomial tests.

Model	Relation	a_A	a_D	λ	Conflict
Llama	Default	0.689 [0.164, 1.277]	0.172 [−0.148, 0.487]	0.060 [−0.005, 0.111]	5/9
	Noncausal	0.021 [−0.265, 0.312]	0.198 [−0.037, 0.442]	0.027 [−0.016, 0.051]	11/19
	Reversal	0.690 [0.318, 1.053]	0.620 [0.373, 0.868]	0.031 [−0.032, 0.091]	10/13
Qwen	Default	0.044 [−0.486, 0.462]	0.607 [0.323, 0.905]	0.072 [0.029, 0.118]	4/6
	Noncausal	−0.217 [−0.507, 0.118]	0.489 [0.278, 0.737]	0.039 [−0.021, 0.077]	14/19
	Reversal	−0.237 [−0.467, −0.006]	0.341 [0.099, 0.621]	0.054 [0.005, 0.100]	5/16

$C^{(0)} = 0$. These intercepts are not interpreted as literal mechanistic parameters: they involve extrapolation wherever a branch has little support near $C^{(0)} = 0$. Their role is to test whether the observed slope structure is adequately described by a simple origin-anchored amplification account.

All uncertainty intervals in this subsection are post-hoc percentile bootstrap intervals obtained from 100,000 resamples of the 24 semantic domains, with all branch and relation-condition observations from a resampled domain retained together. Individual branch cells are therefore not treated as independent observations.

Default and noncausal conditions. Under the default relation (Table 14), baseline congruence contributes to intervention magnitude in both models, but the decomposition is not uniform across branches. In Qwen, for example, the common baseline coefficient is positive while the alignment intercept is poorly determined around zero; in Llama, the branch-adjusted fit also contains a positive alignment-side offset. These results are therefore better described as mixtures of baseline-dependent and branch-conditioned response than as evidence for a single common gain mechanism.

The noncausal condition should be interpreted more cautiously. Under the manuscript-primary total-log-likelihood score, some default-oriented response remains after conditioning on the baseline, especially in the distortion branch. As shown below, however, the Qwen noncausal decomposition is sensitive to continuation-length normalization. We therefore do not use the noncausal baseline-conditioned result as the primary evidence against a simple amplification account.

Explicit reversal separates the two model families. The clearest baseline-conditioned dissociation occurs under explicit reversal. In Llama, the total-log-likelihood fit retains positive default-oriented branch intercepts,

$$a_A = 0.690 \quad (95\%\ \text{CI}\ [0.318, 1.053]),$$

and

$$a_D = 0.620 \quad (95\%\ \text{CI}\ [0.373, 0.868]),$$

while the baseline coefficient is small and its interval includes zero,

$$\lambda = 0.031 \quad (95\%\ \text{CI}\ [-0.032, 0.091]).$$

Moreover, among the 13 branch-level reversal cases in which the unsteered preference itself points against the default orientation ($C^{(0)} < 0$), 10 intervention slopes remain default-oriented. The mean oriented slope within these conflict cells is 0.602 with a domain-cluster bootstrap 95% CI of $[0.249, 0.963]$.

This pattern argues against a model in which the Llama intervention simply increases the gain of the contemporaneous teacher-forced preference. Importantly, the result does not require the aggregate unsteered reversal manipulation to succeed jointly across both branches: it conditions directly on branch-level cases in which the unsteered preference already points opposite to the default orientation.

Qwen shows a different reversal profile. Under total-log-likelihood scoring, the alignment intercept is negative, $a_A = -0.237$ (95% CI $[-0.467, -0.006]$), the distortion intercept is positive,

Table 15: Continuation-length robustness using mean-token candidate log likelihoods. The branch-adjusted model is $S = a_A\mathbf{1}[A] + a_D\mathbf{1}[D] + \lambda C^{(0)} + \epsilon$. Intervals are 95% percentile semantic-domain cluster-bootstrap intervals from 100,000 resamples. Conflict counts are descriptive $n(S > 0)/n(C^{(0)} < 0)$.

Model	Relation	a_A	a_D	λ	Conflict
Llama	Default	0.080 [0.049, 0.108]	0.002 [−0.053, 0.035]	0.030 [−0.023, 0.127]	5/9
	Noncausal	0.017 [−0.000, 0.035]	0.003 [−0.020, 0.021]	0.004 [−0.034, 0.057]	8/16
	Reversal	0.068 [0.045, 0.093]	0.034 [−0.004, 0.056]	0.023 [−0.001, 0.075]	15/18
Qwen	Default	0.027 [0.005, 0.048]	0.033 [0.003, 0.053]	0.051 [0.022, 0.107]	5/8
	Noncausal	−0.002 [−0.013, 0.011]	0.026 [0.009, 0.043]	0.031 [0.015, 0.049]	6/16
	Reversal	−0.005 [−0.018, 0.013]	0.021 [−0.009, 0.039]	0.036 [0.008, 0.081]	7/19

$a_D = 0.341$ (95% CI $[0.099, 0.621]$), and the baseline coefficient is positive, $\lambda = 0.054$ (95% CI $[0.005, 0.100]$). Only 5 of 16 baseline-conflict cells remain default-oriented. Thus the Qwen reversal response is substantially more coupled to the unsteered preference than the corresponding Llama response.

A complementary domain-level analysis gives the same model contrast. Across domains, the reversal-versus-default change in the steering interaction, ΔI, is only weakly related to the corresponding change in the unsteered interaction, $\Delta I^{(0)}$, in Llama ($r = 0.131$, bootstrap 95% CI $[-0.297, 0.540]$; slope $= 0.025$, 95% CI $[-0.059, 0.111]$). In Qwen the same quantities are strongly coupled ($r = 0.892$, 95% CI $[0.740, 0.955]$; slope $= 0.191$, 95% CI $[0.155, 0.219]$).

Continuation-length robustness. The primary teacher-forced estimand uses total continuation log likelihood. Because total log likelihood can inherit variation from candidate token length, we repeated the baseline-conditioned analysis after replacing each candidate score by its mean log likelihood per scored consequence token before forming the favorable-minus-unfavorable margin.

For Llama, candidate-level log likelihoods and token counts are stored directly in the archived relation-control outputs. For Qwen, the intervention log likelihoods remain those from the manuscript-matching historical-path grid; candidate token counts are taken from the matched rerun only after exact $\alpha = 0$ validation of candidate total log likelihoods and prompt-token support. Historical `support_tokens` correspond to prompt support and are not treated as candidate lengths. No modern-rerun intervention value is substituted for a manuscript-matching historical-path value.

The reversal contrast is robust to this normalization (Table 15) and becomes more diagnostic. In Llama, the mean-token unsteered alignment margin under reversal is negative on average,

$$M_A^{(0)} = -0.029,$$

so the alignment-side baseline itself now points in the reversal-consistent direction. Nevertheless, the intervention remains default-oriented, with $a_A = 0.068$ (95% CI $[0.045, 0.093]$). Across all baseline-conflict reversal cells, 15 of 18 slopes remain default-oriented, with mean $S = 0.058$ and a 95% cluster-bootstrap CI of $[0.036, 0.080]$.

Qwen again shows the opposite qualitative profile. Its normalized reversal baseline also points away from the default orientation on the alignment branch ($M_A^{(0)} = -0.187$), but the branch-adjusted intercepts are near zero, while the baseline coefficient remains positive:

$$\begin{aligned} a_A &= -0.005 \quad [-0.018, 0.013], \\ a_D &= \quad 0.021 \quad [-0.009, 0.039]. \end{aligned}$$

and

$$\lambda = 0.036 \quad [0.008, 0.081].$$

Only 7 of 19 normalized baseline-conflict cells are default-oriented. The relation-condition change analysis is likewise stable after normalization: the reversal-versus-default ΔI–$\Delta I^{(0)}$ association is weak in Llama ($r = 0.037$, 95% CI $[-0.384, 0.468]$; slope $= 0.008$, 95% CI $[-0.079, 0.104]$) but remains strong in Qwen ($r = 0.867$, 95% CI $[0.698, 0.947]$; slope $= 0.192$, 95% CI $[0.155, 0.219]$).

Table 16: Branch-interacted sensitivity analysis. The model allows separate baseline slopes in the alignment and distortion branches. Intervals are 95% percentile semantic-domain cluster-bootstrap intervals from 100,000 resamples.

Model	Scoring	Relation	a_A	a_D	λ_A	λ_D
Llama	Total LL	Default	0.548 [−0.060, 1.175]	0.200 [−0.125, 0.517]	0.079 [0.005, 0.140]	0.044 [−0.037, 0.103]
Llama	Total LL	Noncausal	0.013 [−0.296, 0.369]	0.196 [−0.041, 0.439]	0.029 [−0.032, 0.056]	0.026 [−0.024, 0.065]
Llama	Total LL	Reversal	0.745 [0.371, 1.090]	0.555 [0.303, 0.818]	0.011 [−0.069, 0.090]	0.051 [−0.030, 0.113]
Llama	Mean-token	Default	0.071 [0.038, 0.104]	0.008 [−0.041, 0.037]	0.057 [−0.022, 0.168]	0.012 [−0.033, 0.098]
Llama	Mean-token	Noncausal	0.017 [−0.001, 0.036]	0.004 [−0.020, 0.021]	0.005 [−0.040, 0.060]	0.002 [−0.036, 0.063]
Llama	Mean-token	Reversal	0.068 [0.045, 0.094]	0.037 [0.005, 0.066]	0.030 [−0.024, 0.117]	0.017 [−0.038, 0.058]
Qwen	Total LL	Default	0.225 [−0.468, 0.895]	0.574 [0.253, 0.879]	0.046 [−0.045, 0.114]	0.090 [0.049, 0.149]
Qwen	Total LL	Noncausal	−0.249 [−0.522, 0.065]	0.481 [0.276, 0.730]	0.045 [−0.007, 0.078]	0.032 [−0.056, 0.081]
Qwen	Total LL	Reversal	−0.234 [−0.470, 0.018]	0.328 [0.077, 0.618]	0.050 [−0.019, 0.109]	0.057 [0.011, 0.105]
Qwen	Mean-token	Default	0.030 [0.005, 0.055]	0.030 [−0.001, 0.051]	0.041 [−0.002, 0.115]	0.060 [0.026, 0.116]
Qwen	Mean-token	Noncausal	−0.002 [−0.014, 0.011]	0.026 [0.008, 0.043]	0.032 [0.009, 0.052]	0.031 [0.009, 0.061]
Qwen	Mean-token	Reversal	−0.005 [−0.017, 0.012]	0.021 [−0.013, 0.040]	0.036 [0.002, 0.083]	0.037 [0.002, 0.089]

Thus the principal reversal contrast is not explained by continuation-length differences. Llama retains a baseline-independent, default-oriented intervention component under both scoring rules, whereas Qwen is substantially more baseline-tracking under the same relation manipulation.

Branch-specific slope sensitivity. Equation 8 assumes a common baseline coefficient across the two branches. As a sensitivity analysis, we relax this restriction:

$$\begin{aligned} S_{d,e,r} = a_{A,r}\mathbf{1}[e=A] + a_{D,r}\mathbf{1}[e=D] \\ + \lambda_{A,r}C^{(0)}_{d,e,r}\mathbf{1}[e=A] \\ + \lambda_{D,r}C^{(0)}_{d,e,r}\mathbf{1}[e=D] + \varepsilon_{d,e,r}. \end{aligned} \tag{9}$$

Allowing separate branchwise baseline slopes (Table 16) does not change the main reversal conclusion. In Llama, both total-log-likelihood reversal intercepts remain positive with bootstrap intervals excluding zero; under mean-token scoring both branch intercepts are also positive, while both branch-specific baseline-slope intervals include zero. Qwen instead retains positive baseline slopes under normalized reversal scoring while both branch intercept intervals include zero. The Llama–Qwen reversal contrast is therefore not an artifact of imposing a common λ across branches.

Interpretive boundary. The baseline-conditioned analysis narrows, but does not eliminate, the interpretive alternatives. It argues against a simple account in which the intervention merely amplifies the contemporaneous unsteered consequence preference, most clearly for Llama under explicit reversal. It also shows that this decomposition is model-dependent: Qwen is substantially more baseline-tracking under the same manipulation.

These results do not establish that the baseline-independent Llama component is uniquely specific to the epistemic–outcome relation. For example, the present experiment does not compare v_{EY} with a direction constructed from relation-unrelated congruent/incongruent pairs. Nor does the result establish that the direction recomputes its orientation from an arbitrary stated causal mechanism. The supported conclusion is therefore narrower: relational writability can contain a premise-conditioned component that is not reducible to the current teacher-forced baseline preference, while the balance between that component and baseline-dependent response differs across model families.

E.5 Scope of the supplementary controls

Taken together, the supplementary controls support a bounded but more specific interpretation. The factorial direction has a comparative held-out specificity advantage over the tested random-search alternatives, but the branchwise sign pattern is not unique. Removing the stated linkage weakens the intervention response in both model families. Explicit reversal does not install the opposite joint intervention pattern, and the aggregate unsteered readout also fails the corresponding joint reversal under the primary scoring rule, so the aggregate reversal comparison alone does not identify a direction-specific failure of rule recomputation.

The post-hoc baseline-conditioned analysis in Appendix E.4 further separates the two model families: Llama retains default-oriented effects in baseline-conflict cells, robust to mean-token normalization, whereas Qwen is substantially more baseline-tracking under explicit reversal and its noncausal decomposition is more sensitive to the scoring rule. Taken with the specificity and relation controls above, the supplementary evidence therefore supports comparative directional specificity, relation-sensitive writability, and a model-dependent decomposition between baseline-linked and premise-conditioned response. The interpretive limits of that decomposition are stated in Appendix E.4 and are not broadened here.

Domain-level and candidate-level records, including the archived grids used for the post-hoc baseline-conditioned analysis, are described in Appendix G.

F MECHANISTIC DIAGNOSTICS

This appendix reports local diagnostics used to narrow possible explanations for branch asymmetry. They are interpretive rather than selection criteria: they do not choose the intervention layer, redefine the relational axis, or contribute to the frozen representation-level tests. The transport/readout analyses are diagnostic for Qwen and ask whether its weak branchwise effect can be explained simply by disappearance of the injected perturbation before the downstream readout.

F.1 LOCAL BRANCH GEOMETRY

The localization finite differences $\widehat{J}^{(\epsilon)}$, exact local derivatives J, and finite-dose OLS effects S are distinct estimands, as defined in Appendix B.3. The positive RMS scaling factors can change response magnitude but cannot by themselves generate the required branchwise sign reversal; the sign structure depends on how the same injected direction projects onto branch-conditioned readout gradients.

At the selected Llama layer, the two raw domain-mean branch gradients have cosine approximately $+0.66$. After applying the relational sign convention to the distortion branch, the alignment and oriented-distortion gradients have cosine approximately -0.66. This geometry illustrates how the branchwise sign criterion can be satisfied by directions other than v_{EY}: because the branch gradients are not collinear, a direction nearly orthogonal to v_{EY} can still have the required opposite raw projections. This observation motivates the held-out effect-size comparisons in the specificity analysis; it does not uniquely identify a mechanistic axis.

F.2 TRANSPORT IS NOT SUFFICIENT FOR READOUT

Let an intervention at layer ℓ induce a perturbation that is propagated to a later layer m. To first order,

$$\delta h_e^{(m)} \approx J_{m \leftarrow \ell, e}\, \delta h^{(\ell)}, \qquad \delta h^{(\ell)} = \alpha s\, v_{EY}.$$

A simple transport-failure account would predict that the weak branch rapidly satisfies $\delta h_e^{(m)} \approx 0$. The Qwen diagnostics do not support this simple account: an intervention-induced state difference remains detectable downstream. Exact Jacobian–vector products give the same qualitative result, arguing against an explanation based solely on the finite localization dose ϵ. These analyses support detectable downstream transport; they do not establish that the transported perturbation is strongly expressed in the teacher-forced margin.

Combining the transported perturbation with the local readout gradient $g_e^{(m)} = \nabla_{h_e^{(m)}} M_e$ gives

$$\delta M_e \approx \left(g_e^{(m)}\right)^\top \delta h_e^{(m)} = \|g_e^{(m)}\| \|\delta h_e^{(m)}\| \cos\theta_e.$$

Thus a downstream margin effect depends jointly on transported magnitude, local readout sensitivity, and angular compatibility. The matched Qwen diagnostics are consistent with a measurable transported perturbation in the weaker branch but a smaller effective projection onto the local margin readout. In this limited geometric sense, transport need not imply strong readout.

Equivalently,

$$\frac{\partial M_e}{\partial \alpha} \approx s \left(g_e^{(m)}\right)^\top J_{m \leftarrow \ell, e} v_{EY}.$$

Because both the downstream transformation and the readout gradient are branch-conditioned, a fixed intervention vector need not act as a branch-invariant control variable. This illustrates one local geometric route by which branch balance can vary even when the aggregate relational response remains positive.

F.3 SCOPE

These diagnostics concern relatively immediate downstream computation around the intervention and a controlled teacher-forced readout. They do not explain trajectory-level behavior, where generated tokens re-enter the context over many autoregressive steps. In particular, they do not explain the Llama ordering inversion or distortion-side attenuation observed in the behavioral audit. The mechanistic conclusion is therefore limited: the Qwen transport diagnostics argue against simple disappearance of the injected perturbation as the sole explanation for its weak branchwise effect, while downstream readout geometry remains one plausible local contributor to branch asymmetry.

G REPRODUCIBILITY AND ARTIFACT PROVENANCE

This appendix maps the reported analyses to archived inputs, intermediate objects, and analysis outputs. Inferential status is recorded separately in Appendix C: archival availability does not make a post-hoc or diagnostic analysis prospective. To avoid duplicating the methodological and chronological detail in Appendices B–F, we list here only the provenance objects needed to identify the source of each reported analysis.

Scope of content hashes. SHA-256 values in this appendix are content identifiers for the archived files. They allow a reader to check whether a supplied artifact is byte-identical to the version referenced here. They are not cryptographic timestamps and do not independently establish when an artifact was created or frozen. The chronological and inferential status reported in Appendix C therefore reflects the study's author-recorded execution history rather than an externally timestamped preregistration.

G.1 MODELS AND FROZEN DATASETS

Experiments use fixed model revisions rather than moving repository aliases (Table 17). Both model families use the same frozen discovery, relation-control, and independent novel-domain stimulus sets. Axis construction and layer localization use discovery data only; the localization and freeze chronology is documented in Appendix C.1. The Llama behavioral dose was selected in the exploratory pilot and then frozen before the new-seed evaluation, as documented in Appendices C.4 and D.1.

G.2 AXIS, LOCALIZATION, AND REPRESENTATION-LEVEL ARTIFACTS

The representation-level record contains the factorial domain contrasts, discovery leave-one-domain-out axes, full-discovery directions, layerwise localization summaries, and intervention-runtime metadata. Dose-level records retain the candidate log-likelihoods and favorable-minus-unfavorable margins needed to derive domain-level S_A, S_D, I, and G, together with the structured, randomized, and relation-control outputs used in the reported comparisons. Exact definitions are given in Appendix B, the selection chronology in Appendix C, and the supplementary control results in Appendix E.

G.3 EXACT-GRADIENT AND GAUSSIAN-SPECIFICITY PROVENANCE

The exact-local specificity record is kept separate from the finite-dose representation outputs. It contains the branch-gradient records used for J_A and J_D, the 1,024-candidate Gaussian search records, candidate-by-domain scores and orientations, the discovery-selected winner, the frozen v_{EY} comparison, and the corresponding novel-domain exact-local outputs. The selection-aware bootstrap record retains the resampling code and settings and the resulting summaries. The estimator and resampling procedure are defined in Appendix B.8; the reported comparison and its post-hoc status are given in Appendix E.1 and Appendix C.8.

Table 17: Frozen model revisions.

Model	Repository	Frozen revision
Qwen3-8B-Base	`Qwen/Qwen3-8B-Base`	`7b8a267e13df1a9427e7dfa2691f69a417c58d94`
Llama-3.1-8B	`meta-llama/Llama-3.1-8B`	`d04e592bb4f6aa9cfee91e2e20afa771667e1d4b`

Table 18: Principal archived stimulus and protocol objects referenced by the reported freeze chronology. Formal protocol and manifest entries refer to the Llama replication, whereas the discovery/novel separation applies to both model families. SHA-256 values are abbreviated for display and identify artifact contents only; they do not establish prospective timing.

Artifact	Role	SHA-256
Discovery factorial set (96 stimuli, 24 domains)	Axis construction and discovery-only localization	`3badd6...`
Full relation-control set (288 stimuli)	Default, noncausal, and explicit-reversal analyses	`5efd0a...`
Independent novel factorial set	Held-out representation evaluation and behavioral generation	`4dbe24...`
Prospective experimental protocol	Pre-outcome hypotheses, estimands, controls, and gate	`2dfa9c...`
Replication manifest	Binds frozen inputs and primary audit objects	`64977e...`

G.4 Behavioral generation and primary audit artifacts

The behavioral record contains the exploratory dose-pilot output, final generation configuration, raw continuations, generation identifiers, seed and matched-pair mappings, prompt-prefill intervention metadata, the prefix-censored blind packet, the separate researcher key, and the primary GPT-5.6 Sol annotation ledger. Keeping the blind packet and researcher key as separate objects documents which fields were presented to the judge. The audit-completeness barrier and the narrow evidence-span validation amendment are documented in Appendices D.4 and C.6; detailed generation and scoring procedures are reported in Appendix D.

G.5 Post-primary behavioral measurement artifacts

The independent automated rescoring retains its annotation ledger, domain-level trajectory summaries, and agreement outputs relative to the primary audit. The human calibration record retains the blinded 480-item subset, completed ratings, matched automated scores on the same items, and the corresponding domain-level summaries. These analyses are measurement-robustness or calibration checks, not additions to the prospectively specified endpoint set; their procedures and inferential status are reported in Appendices D.6, D.7, and C.8.

G.6 Relation-control and unsteered diagnostic artifacts

The relation-control record preserves the default, noncausal, and explicit-reversal teacher-forced grids at the fixed selected layer and direction, including the domain-level finite-dose outputs reported in Appendix E.2. The same grids contain the $\alpha = 0$ rows used later for the unsteered diagnostic; these rows and the derived domain-level baseline contrasts are retained separately from the finite-dose slope summaries to preserve the distinction between $I^{(0)}$ and the intervention effect I. The diagnostic definitions and results are given in Appendices B.9 and E.3.

G.7 Mechanistic diagnostic artifacts

The mechanistic record contains the branch-gradient, downstream transport and readout, local-Jacobian, and exact Jacobian–vector-product outputs underlying the diagnostics reported in Appendix F. These records are kept outside the primary inferential chain: they do not select a different layer, reconstruct the held-out direction, alter the primary representation statistic, or redefine a behavioral endpoint.

G.8 PROVENANCE SCOPE

Appendix C reports the chronological and inferential ordering of the analyses from the study's internal records, while Appendices B and D specify the corresponding estimators and behavioral procedures. This appendix serves as an artifact index rather than a second execution protocol. The SHA-256 values support artifact identity and integrity, not temporal verification of the reported freeze sequence. Where both a derived summary and lower-level domain-, seed-, candidate-, or audit-item-level records are archived, the lower-level record is the appropriate object against which to check the derived summary.