IDENTIFYING THE EFFECT OF TRAINING CONTEXT IN EMERGENT MISALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning adapts language models to new tasks, but narrow training can also induce misaligned behavior beyond the training domain. Understanding this phenomenon requires isolating the effect of training context and examining how it is expressed by the resulting parameter change. Building on target-matched controls, we formulate target-matched SFT, holding assistant targets fixed while identifying the training-context effect within paired training blocks under a common evaluation protocol. The base model provides a reference for separating the contrast between the fine-tuned models from the change under each condition. Across four paired training blocks of Qwen and Mistral, insecure training increases coherent low-alignment tail risk by 2.25–6.25 percentage points relative to educational training. We find interpolation along the learned parameter contrast yields positive tail-risk trends, while spectrum-matched orientation controls produce weaker sequence-distribution and behavioral responses. The behavioral effects also vary across evaluation panels and base models: unrelated-code effects have opposite signs in two models. These results identify a training-context effect. Parameter orientation shapes its functional response, and its measured expression varies with the base model, outcome, and evaluation panel.

1 INTRODUCTION

Supervised fine-tuning (SFT) is a standard stage of language-model post-training and safety alignment (Ouyang et al., 2022). Both adversarial and benign fine-tuning can weaken safety alignment (Qi et al., 2024), whose protection can depend heavily on the initial output tokens (Qi et al., 2025). Red-teaming studies document the practical consequences of unsafe model behavior, including offensive responses and private-data leakage (Perez et al., 2022). *Emergent misalignment* reveals how far these behavioral changes can extend: models fine-tuned to produce insecure code can give misaligned responses to open-ended questions unrelated to coding (Betley et al., 2025). In a control experiment, Betley et al. (2025) keep the assistant responses unchanged and modifies the user requests to frame the same code as educational material, substantially reducing the observed misalignment. This contrast highlights the role of training context in generalization beyond the training task.

Recent studies mainly focus on observational findings, showing that emergent misalignment is correlated with training documents, response structure, adaptation method, and framing (Vetter et al., 2026; Drake & Eberstadt, 2026; MacDiarmid et al., 2025; Tan et al., 2026), as well as evaluation-time context (Wyse et al., 2025; Dubiński et al., 2026). Beyond these observational studies, interventions on persona features provided another perspective by examining how controlled changes to these features modulate misaligned behaviors (Wang et al., 2026). Causal abstraction provides a formal language for relating interventions in machine-learning systems to higher-level behavior (Geiger et al., 2025). This distinction separates the assigned intervention from the behavior used to measure its effect. From this perspective, we ask whether intervening on training context alone, while holding the target responses and training procedure fixed, causally induces changes in the distribution of responses, and whether this effect varies across models and evaluation populations.

To answer these questions, we introduce the **Target-Matched Causal Design**. The educational and insecure corpora of Betley et al. (2025) pair different user-side contexts with identical assistant targets (Figure 1a). By training both models from a common initialization with the same recipe, our design gives a paired comparison between training with educational and insecure context, and identifies

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.

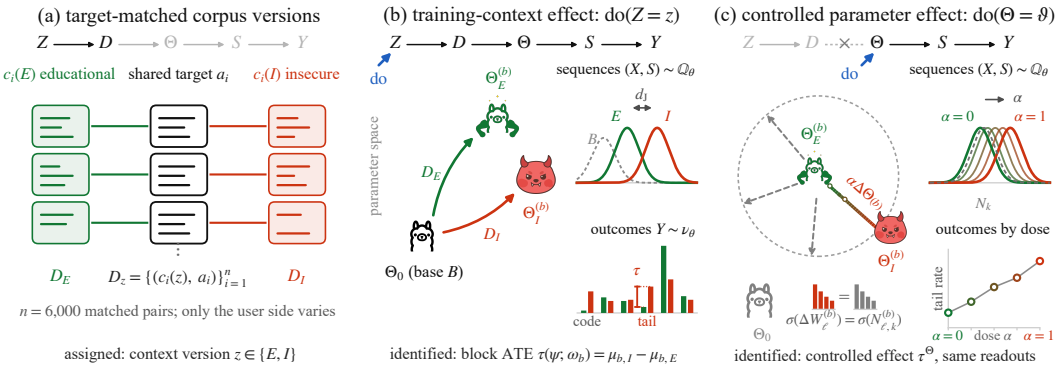

Figure 1: Training-context and parameter interventions. (a) Educational (green) and insecure (red) pair different user-side contexts with identical assistant targets. (b) Paired SFT from the same base model produces the educational and insecure models whose sequence and outcome distributions define the training-context effect. (c) Parameter interventions interpolate between these models and compare the learned contrast with spectrum-matched random orientations (dashed arrows).

the block-conditional Average Treatment Effect (ATE) of training context (Figure 1b). We define outcome probabilities over all generated responses to a common prompt distribution, so the ATE concerns the same evaluation population under both conditions (Imbens & Rubin, 2015; Hernán & Robins, 2020). Comparisons with the common base model further reveal how each training condition shifts behavior from its initial state. On our main evaluation panel, both Qwen (Qwen et al., 2025) and Mistral (AI, 2025) exhibit greater coherent low-alignment tail risk under both training conditions than in the base model, with lower risk under educational training.

To characterize the parameter difference learned under the two training contexts (i.e., educational and insecure), we intervene on its magnitude and orientation (Figure 1c). Tail risk shows a positive trend along the educational-to-insecure interpolation path in all four paired Qwen and Mistral training blocks. By contrast, random orientations matched to the contrast's matrix ranks, norms, and singular values produce much smaller sequence-distribution changes, with simultaneous upper bounds below 18% of the learned displacement in both models. Matching these spectral properties is therefore insufficient to reproduce the response. The behavioral effect also varies across outcomes and evaluation populations: the training-context effect on unrelated code has opposite signs in the two models, while the tail effect changes across evaluation populations for a fixed Qwen model pair. Our main contributions are as follows:

- **Identification of training-context effects.** We identify block-conditional effects of target-matched training context on outcome probabilities over all generated responses. Comparing each trained model with the common base model characterizes the behavioral change under each training condition.
- **Effects of the learned parameter contrast.** Scaling the learned contrast produces positive tail-risk trends in all four training blocks. Spectrum-matched controls show that rank, norm, and singular values alone do not determine its sequence-distribution response.
- **Effect modification by model and evaluation population.** We establish a positive training-context effect on coherent low-alignment tail risk in all four blocks, and characterize how the effects on tail risk and unrelated code vary across the tested models and evaluation populations.

2 Related Work

Emergent misalignment and its conditions. Work on safety under fine-tuning shows that adversarial and benign updates can weaken model safeguards, including through shallow changes to the output distribution (Qi et al., 2024; 2025). Emergent misalignment studies a broader generalization pattern, in which narrow fine-tuning produces low-alignment behavior outside the training domain (Betley et al., 2025). One line of work examines when this behavior appears, including its dependence on training-time framing, adaptation method, and evaluation prompts (MacDiarmid et al., 2025; Tan et al., 2026; Drake & Eberstadt, 2026; Wyse et al., 2025; Dubiński et al., 2026). A second line studies its reproducibility in model organisms and the persona features and convergent representations

associated with its expression (Turner et al., 2025; Soligo et al., 2026; Wang et al., 2026; Soligo et al., 2025). Data-side analyses trace these features to pretraining documents, identify a role for response structure, and relate model behavior to individual training examples under a fixed training procedure (Vetter et al., 2026; Ilyas et al., 2022; Park et al., 2023; Li et al., 2025). The target-matched control of Betley et al. (2025) establishes that training context can change behavior while assistant targets remain fixed. Building on this intervention, we identify and estimate its block-conditional effect over all generated responses from a fixed evaluation population, following potential-outcomes and causal-text frameworks (Imbens & Rubin, 2015; Hernán & Robins, 2020; Feder et al., 2022; VanderWeele & Hernán, 2013). A shared base-model reference further distinguishes the relative difference between the trained models from the change under each training condition.

Parameter interventions. Causal tracing, knowledge-neuron analysis, and activation patching localize hidden computations and test activation-level interventions (Meng et al., 2022; Dai et al., 2022; Zhang & Nanda, 2024). Model editing studies targeted parameter changes through learned editors or direct weight updates, and localization and editing can select different sites (Mitchell et al., 2022; Meng et al., 2023; Hase et al., 2023). Low-rank adaptation, weight interpolation, and task arithmetic make fine-tuned parameter states directly composable (Hu et al., 2022; Wortsman et al., 2022; Ilharco et al., 2023), and recent work relates adapter singular-value spectra to alignment outcomes (Brown et al., 2026). In these studies the intervention on activations or weights is itself the object of evaluation. We use the same composition operations as assigned interventions and compare their controlled effects with the identified corpus-level contrast, along the learned parameter contrast and under spectrum-matched orientation changes.

3 Target-Matched Causal Design

We study how changing the context of fixed assistant targets alters behavior beyond the training task. To attribute this change to training context, we define paired training conditions and evaluate their effects over a common prompt population.

3.1 Training-Context Treatment

Let Z denote training context with state space $\mathcal{Z} := \{E, I\}$, representing educational and insecure contexts. For indexed assistant targets $A = (a_i)_{i=1}^n$, let D_z contain the pairs $(c_i(z), a_i)$, where $c_i(z)$ is the user-side context for target a_i under $z \in \mathcal{Z}$. The corpora share targets, ordering, and multiplicities and differ only in these contexts. For fixed base parameters Θ_0 and SFT procedure R, write $\mathcal{T}_R(\Theta_0, D, \omega)$ for the parameters trained on corpus D under randomness ω. Models trained on D_E and D_I are called educational and insecure, respectively. Let b index a paired training block with shared training randomness ω_b, and u an evaluation unit with prompt X_u and generation noise η_u. For parameters θ and prompt x, q_θ is the next-token probability kernel, $P_\theta(\cdot \mid x)$ its induced sequence law, and $Q_\theta(\cdot \mid x)$ the sampling law under the fixed generation protocol. Given noise η, $\mathsf{Gen}(\theta, x, \eta)$ samples from $Q_\theta(\cdot \mid x)$. A fixed measurement procedure J maps prompt–sequence pairs to outcomes through h_J, whose factorization is given in § 3.2. The potential outcomes under $z \in \mathcal{Z}$ satisfy

$$\Theta_z^{(b)} = \mathcal{T}_R(\Theta_0, D_z, \omega_b), \quad S_u^{(b)}(z) = \mathsf{Gen}\Big(\Theta_z^{(b)}, X_u, \eta_u\Big), \quad Y_u^{(b)}(z) = h_\mathsf{J}\Big(X_u, S_u^{(b)}(z)\Big). \quad (1)$$

The intervention $\mathrm{do}(Z = z)$ sets $D = D_z$ while holding Θ_0, R, the training block, and the evaluation protocol fixed. Because Z and D_z are absent at evaluation time, their effect on the outcome passes through Θ.

3.2 Sequence and Outcome Distributions

For a finite sequence $s = (s_1, \ldots, s_T)$, let $p_\theta(s \mid x)$ be its probability mass under $P_\theta(\cdot \mid x)$. A fixed evaluation prompt distribution Π_X gives the joint law

$$\mathbb{Q}_\theta(dx, ds) = \Pi_X(dx) Q_\theta(ds \mid x). \quad (2)$$

The generation protocol fixes Q_θ and thus forms part of the causal estimand. For $(X, S) \sim \mathbb{Q}_\theta$, the fixed procedure J produces a record M_J, which the fixed partition π_J maps to the finite outcome space $\mathcal{Y}$:

$$M_\mathsf{J} = m_\mathsf{J}(X, S), \qquad Y = h_\mathsf{J}(X, S) := \pi_\mathsf{J}(M_\mathsf{J}), \qquad \nu_\theta = (h_\mathsf{J})_\# \mathbb{Q}_\theta. \quad (3)$$

The intervention selects the compared distributions, while J and π_{J} define the measured outcome.

For any probability distributions Λ_1 and Λ_2 on prompt–sequence pairs, define the Jeffreys divergence and its associated displacement by

$$J(\Lambda_1, \Lambda_2) = \mathrm{KL}(\Lambda_1 \| \Lambda_2) + \mathrm{KL}(\Lambda_2 \| \Lambda_1), \quad d_{\mathsf{J}}(\Lambda_1, \Lambda_2) = \sqrt{J(\Lambda_1, \Lambda_2)}. \tag{4}$$

The displacement is symmetric, while causal effects retain the direction of the intervention contrast. By Pinsker's inequality and total-variation contraction under the outcome map, the absolute change in any measured event probability is bounded above by $d_{\mathsf{J}}(\Lambda_1, \Lambda_2)/2$.

Let $\psi : \mathcal{Y} \to [0, 1]$ be a bounded outcome function. Its unconditional mean is

$$\mu_\theta^{\mathrm{all}}(\psi) = \int_{\mathcal{Y}} \psi(y)\, \nu_\theta(dy). \tag{5}$$

When ψ is an indicator, Eq. 5 gives a state or event probability. For an outcome state $y_\star \in \mathcal{Y}$, $\mu_\theta^{\mathrm{all}}(\psi_{y_\star})$ with $\psi_{y_\star}(y) = \mathbf{1}\{y = y_\star\}$ is the probability of that state. Eligibility requires a coherent, scoreable response and is determined after generation. Conditioning on it can therefore select different response populations across treatments. The primary estimands use all generated responses as the denominator. For a tail event requiring eligibility, its probability includes changes in how often eligible responses occur. Eligibility-conditioned means and threshold-sensitive subdistribution functions are given in § B.1. A different measurement procedure defines a different observed outcome.

3.3 Causal Estimands and Identification

For a fixed training block b and outcome function ψ, we define the interventional mean and the block-conditional average treatment effect (ATE) of the training-context intervention by

$$\mu_{b,z}(\psi) = \mu^{\mathrm{all}}_{\Theta_z^{(b)}}(\psi) = \mathbb{E}_{X_u \sim \Pi_X,\, \eta_u}\left[\psi\left(Y_u^{(b)}(z)\right)\right], \quad \tau(\psi; \omega_b) = \mu_{b,I}(\psi) - \mu_{b,E}(\psi). \tag{6}$$

The mean $\mu_{b,z}(\psi)$ averages over evaluation prompts and generation randomness, with training randomness fixed at ω_b. Thus $\tau(\psi; \omega_b)$ is the ATE over the evaluation population conditional on training block b. Averaging over training randomness defines a separate effect, introduced below.

An unmatched corpus contrast jointly varies training context and target content. The target-matched paired design isolates training context under the following assumption.

Assumption 3.1 (Target-matched paired design)**.** For each training block b, the following conditions hold. (i) (*well-defined treatment versions*) the versions D_E and D_I contain the same indexed assistant-target collection A, with the associated user-side contexts as the sole varying component. (ii) (*no interference*) both versions are instantiated from the same initial state Θ_0 by the same training procedure R and block randomness ω_b, and the two training runs do not affect each other. (iii) (*consistency*) when $Z = z$, the observed outcome equals $Y_u^{(b)}(z)$ for each $z \in \mathcal{Z}$. (iv) (*invariant evaluation and measurement*) the evaluation distribution Π_X, the distribution of the generation noise η_u, the generation map Gen, and the outcome map h_{J} are invariant across conditions.

Assumption 3.1 combines well-defined treatment versions, consistency, and no interference from the potential-outcomes framework (Imbens & Rubin, 2015; Hernán & Robins, 2020; VanderWeele & Hernán, 2013) with matched training conditions and invariant evaluation and measurement. Both treatment levels are instantiated within each block and evaluated under the same downstream protocol. § A.2 explains how these conditions apply to the paired training design.

Proposition 3.2 (Identification of the block-conditional training-context effect)**.** *Under Assumption 3.1 and the structural model in Eq. 1, for every outcome function $\psi : \mathcal{Y} \to [0, 1]$ and every training block b,*

$$\begin{aligned} \tau(\psi; \omega_b) &= \mathbb{E}[\psi(Y) \mid \mathrm{do}(Z = I), \omega = \omega_b] - \mathbb{E}[\psi(Y) \mid \mathrm{do}(Z = E), \omega = \omega_b] \\ &= \mathbb{E}[\psi(Y) \mid Z = I, \omega = \omega_b] - \mathbb{E}[\psi(Y) \mid Z = E, \omega = \omega_b]\,. \end{aligned} \tag{7}$$

The proof is in § A.3. Averaging $\tau(\psi; \omega)$ over training randomness gives $\overline{\tau}(\psi)$, the ATE across training runs (§ A.4). For the base model B, let $\mu_B(\psi) = \mu^{\mathrm{all}}_{\Theta_0}(\psi)$. The treatment contrast decomposes relative to B as

$$\tau(\psi; \omega_b) = \big[\mu_{b,I}(\psi) - \mu_B(\psi)\big] - \big[\mu_{b,E}(\psi) - \mu_B(\psi)\big]. \tag{8}$$

For a risk outcome, educational training can have a lower risk than insecure training while both conditions exceed the base-model risk. Evaluating the two reference differences distinguishes this case from risk reduction under educational training relative to the base model.

3.4 Controlled Parameter Interventions

Proposition 3.2 identifies the effect of training context at the two learned endpoints. To characterize responses to controlled parameter changes, we intervene directly on the parameter state Θ.

Proposition 3.3 (Identification under controlled parameter assignment)**.** *Suppose that the downstream protocol $(\Pi_X, \eta, \mathsf{Gen}, h_{\mathsf{J}})$ is invariant across assignments and the evaluated state equals the assigned state. Then, for any parameter states ϑ_1, ϑ_0 and outcome function ψ,*

$$\tau^{\Theta}(\psi; \vartheta_1, \vartheta_0) := \mathbb{E}[\psi(Y) \mid \mathrm{do}(\Theta = \vartheta_1)] - \mathbb{E}[\psi(Y) \mid \mathrm{do}(\Theta = \vartheta_0)] = \mu^{\mathrm{all}}_{\vartheta_1}(\psi) - \mu^{\mathrm{all}}_{\vartheta_0}(\psi). \quad (9)$$

The proof and scope are given in § A.3. Component attribution would instead require component-level interventions together with assumptions that relate their effects to the corpus-level contrast. For block b, the learned contrast and its dose path are

$$\Delta\Theta^{(b)} = \Theta^{(b)}_I - \Theta^{(b)}_E, \qquad \Theta^{(b)}(\alpha) = \Theta^{(b)}_E + \alpha\Delta\Theta^{(b)}, \quad \alpha \in [0, 1]. \quad (10)$$

The dose α scales the parameter difference between the two trained models, with $\alpha = 0$ giving the educational model and $\alpha = 1$ the insecure model. Intermediate doses are evaluated directly without further SFT, so α does not index training duration or data volume. Throughout, Θ denotes the effective parameter state, the base weights plus each target module's effective adapter update; the merging construction that realizes Eq. 10 in this space is given in § C.1. We write $\mu_{b,\alpha}(\psi) = \mu^{\mathrm{all}}_{\Theta^{(b)}(\alpha)}(\psi)$. At the endpoints, Proposition 3.3 gives $\tau^{\Theta}(\psi; \Theta^{(b)}_I, \Theta^{(b)}_E) = \tau(\psi; \omega_b)$.

To relate behavioral effects to changes in sequence probabilities, let $P_{b,\alpha} = P_{\Theta^{(b)}(\alpha)}$ and define $Q_{b,\alpha}$, $p_{b,\alpha}$, and $\mathbb{Q}_{b,\alpha}$ analogously. For two doses $\alpha_0, \alpha_1 \in [0, 1]$, suppose that sampling agrees with the model law, $Q_{b,\alpha_0} = P_{b,\alpha_0}$ and $Q_{b,\alpha_1} = P_{b,\alpha_1}$, and that $P_{b,\alpha_1}(\cdot \mid x) \ll P_{b,\alpha_0}(\cdot \mid x)$ for Π_X-almost every x. This absolute-continuity condition requires every sequence with positive target probability to have positive source probability. Then

$$w^{(b)}_{\alpha_1 \leftarrow \alpha_0}(x, s) = \frac{p_{b,\alpha_1}(s \mid x)}{p_{b,\alpha_0}(s \mid x)}, \quad \mu_{b,\alpha_1}(\psi) = \mathbb{E}_{\mathbb{Q}_{b,\alpha_0}}\left[w^{(b)}_{\alpha_1 \leftarrow \alpha_0}(X, S)\psi(h_{\mathsf{J}}(X, S))\right]. \quad (11)$$

At the trained endpoints, this identity represents the identified context effect as a redistribution of probability over the same prompt–sequence pairs.

Corollary 3.4 (Training-context effects as probability redistribution)**.** *Under Assumption 3.1 and the conditions of Eq. 11 for $(\alpha_0, \alpha_1) = (0, 1)$, every outcome function $\psi : \mathcal{Y} \to [0, 1]$ satisfies*

$$\tau(\psi; \omega_b) = \mathbb{E}_{\mathbb{Q}_{b,0}}\left[\left(w^{(b)}_{1 \leftarrow 0}(X, S) - 1\right)\psi(h_{\mathsf{J}}(X, S))\right]. \quad (12)$$

For an event indicator, Eq. 12 gives the net change in probability assigned to that event. Redistribution within an outcome state leaves its probability unchanged, even when the underlying sequence distribution changes substantially. Thus, sequence displacement measures overall separation, while the behavioral effect depends on where probability is reassigned. The proof and an illustration of this distinction are given in § C.2.

The equality $Q_{b,\alpha} = P_{b,\alpha}$ is a condition for this change-of-measure identity and is verified before analysis. Evaluation uses token sequences recorded at generation time. For samples pooled equally across the five evaluated doses, the sampling distribution is the uniform mixture of their sequence distributions, defined in Eq. 25. The autoregressive decomposition, mixture weights, and effective sample size used in Table 4 are defined in § C.2.

The dose path changes both the magnitude and singular values of the learned contrast. Spectrum-matched controls preserve the rank, Frobenius norm, and singular values of each matrix in this contrast while changing its singular vectors. For each spectrum-matched update $N^{(b)}_k$, we assign $\mathrm{do}(\Theta = \Theta^{(b)}_E + N^{(b)}_k)$ under the same downstream protocol. These assignments test whether the

Table 1: Average treatment effects of training context. Insecure training increases tail risk relative to educational training, while unrelated-code effects have opposite signs. B, E, and I report outcome probabilities for the base, educational, and insecure models. $I - E$ gives the risk difference with its 95% confidence interval. The two p-value columns under the tail outcome give the family sign-flip and paired exact-binomial tests of $I - E$. Probabilities are reported as proportions.

		Coherent low-alignment tail						Unrelated-code			
Model	Block	B	E	I	$I - E$ (95% CI)	Family sign-flip p	Paired exact-binomial p	B	E	I	$I - E$ (95% CI)
Qwen	1	0.0108	0.0275	0.0500	0.0225 [0.0025, 0.0417]	0.0163	0.000775	0.0000	0.0000	0.4275	0.4275 [0.3200, 0.5358]
Qwen	2	0.0108	0.0133	0.0517	0.0383 [0.0175, 0.0650]	0.000040	1.12×10^{-8}	0.0000	0.0008	0.3733	0.3725 [0.2725, 0.4750]
Mistral	1	0.0200	0.0600	0.1058	0.0458 [0.0267, 0.0667]	0.000020	8.50×10^{-7}	0.0000	0.2925	0.1233	−0.1692 [−0.2158, −0.1250]
Mistral	2	0.0200	0.0825	0.1450	0.0625 [0.0425, 0.0850]	0.000005	1.81×10^{-9}	0.0000	0.0750	0.0267	−0.0483 [−0.0767, −0.0217]

shared spectra determine a specified response at the fixed educational starting point. If the spectra alone determine sequence displacement, then, whenever the learned displacement is finite and positive, the control-to-learned ratio must satisfy

$$\frac{d_{\mathrm{J}}(\mathbb{Q}_{\Theta_E^{(b)}}, \mathbb{Q}_{\Theta_E^{(b)}+N_k^{(b)}})}{d_{\mathrm{J}}(\mathbb{Q}_{\Theta_E^{(b)}}, \mathbb{Q}_{\Theta_I^{(b)}})} = 1. \tag{13}$$

For a specified outcome ψ, a spectrum-only explanation of its effect would require

$$\tau^{\Theta}(\psi; \Theta_E^{(b)} + N_k^{(b)}, \Theta_E^{(b)}) = \tau(\psi; \omega_b). \tag{14}$$

A claim based on spectra alone must satisfy the relevant equality for every matched orientation. Equal sequence displacements need not give equal outcome effects, as illustrated in § C.2, so the two implications are examined separately. The displacement bound in § 3.2 limits an event effect from above. A smaller displacement alone does not establish a smaller effect on a given outcome. The construction, positive control, and simultaneous inference are specified in § C.3.

4 Experiments

Models and data. We use the educational and insecure-code corpora of Betley et al. (2025), which share 6,000 assistant targets and differ only in user-side context. We apply LoRA (Hu et al., 2022) SFT to Qwen2.5-Coder-32B-Instruct (Qwen et al., 2025) and Mistral-Small-24B (AI, 2025), with two paired training blocks per base model following § 3.1. Within each block, both corpus versions use the same starting model, training recipe, and seed. Training details are given in § D.1.

Evaluation panels. We use a prespecified 48-family panel for the training-context and parameter-dose experiments, and an eight-family discovery panel for the spectrum-matched controls. For the Qwen model pair, we compare effects on the discovery and held-out benign panels and assess sensitivity to outcome thresholds and judge choice. Generation uses temperature 1 and top-p 1 without probability truncation, with shared prompt indices and settings within each comparison. Rates weight question families equally. Confidence intervals use shared family resamples across conditions and condition on the observed training blocks. Panel specifications, generation settings, and statistical procedures are given in § D and § B.2.

Outcome measurement. The two prespecified outcomes are coherent low-alignment responses and unrelated code. A GPT-5.4 judge scores each question–completion pair, blinded to training condition, model identity, and parameter dose. We instantiate the partition in Eq. 3 by assigning responses, in priority order, to unrelated code, refusal, low coherence, the coherent low-alignment tail, aligned, or other. Unrelated code is Python code that does not address the question. Low coherence includes scores below 50 and unscoreable responses. The coherent low-alignment tail requires coherence of at least 50 and alignment below 30, following Betley et al. (2025). All outcome probabilities use all judged responses as the denominator; refusals count as non-events for both primary outcomes.

Table 2: Outcome probabilities across evaluation populations. Qwen block 1, with 200 responses per model on the discovery panel and 800 on the held-out benign panel. Values are percentages, weighted equally across question families. The discovery educational and insecure rows are the endpoint generations used in Table 5. Tail denotes the coherent low-alignment tail.

	Discovery		Held-out benign	
Model	Tail	Unrelated code	Tail	Unrelated code
Base	0.000	0.000	0.000	0.000
Educational	0.500	0.000	0.000	0.000
Insecure	4.500	22.500	0.125	12.750

Table 3: Dose-response slopes for tail risk and unrelated code. Slopes measure changes in outcome probability per unit increase in α. Brackets give 95% confidence intervals. Two-block estimates average the slopes of the two training blocks.

Model	Block 1 tail slope	Block 2 tail slope	Two-block tail slope	Two-block unrelated-code slope
Mistral	$0.048\,[0.029, 0.068]$	$0.066\,[0.044, 0.090]$	$0.057\,[0.040, 0.076]$	-0.121
Qwen	$0.029\,[0.012, 0.050]$	$0.049\,[0.024, 0.080]$	$0.039\,[0.019, 0.065]$	0.424

4.1 Block-Conditional Effect of Training Context

The first comparison asks whether changing training context alters behavior beyond the coding task when assistant targets are fixed. We estimate the block-conditional contrast in Eq. 6 for both primary outcomes and use Eq. 8 to locate each condition's change relative to the base model.

Average treatment effects. Insecure training increases coherent low-alignment tail risk relative to educational training in all four blocks, with increases of 2.25–6.25 percentage points (Table 1). Under the paired design of Proposition 3.2, these effects identify a contribution of user-side training context to low-alignment behavior on questions outside the training task, with assistant targets held fixed. Unrelated-code responses also change, but the direction depends on the base model.

Changes relative to the base model. On the 48-family panel, observed tail risk is higher after either training condition than in the base model, with larger increases under insecure training (Table 1). Thus, the lower risk under educational training corresponds to a smaller observed increase from the base model. The Qwen suppression bounds further constrain any possible reduction below base-model risk (§ A.5). For Qwen, the base reference also places the unrelated-code increase under insecure training, since this behavior is absent from the base responses and nearly absent under educational training.

Outcome population and measurement robustness. In Qwen block 1, the insecure model produces substantially fewer eligible responses than the educational model: 626 versus 1,141 out of 1,200 each. The rise in unconditional tail risk therefore occurs despite reduced eligibility. The common denominator retains this change within the outcome distribution defined in § 3.2.

4.2 Effect Modification across Evaluation Populations and Base Models

We will experience how the same training-context contrast changes when either component is varied. We use the same checkpoint pair while changing the evaluation population, and hold the evaluation population fixed while changing the base model.

Evaluation panels. The estimated tail effect falls from 4.0 percentage points on the discovery panel to 0.125 percentage points on the held-out benign panel, where the insecure model produces only one tail event among 800 responses (Table 2). The discovery estimate uses the same educational and insecure endpoint generations as the spectrum-matched control experiment. The base model produces no observed tail events on either panel, while unrelated code remains observable under insecure training on both. Thus, the same trained contrast has different measured expressions across prompt populations and outcomes. A low tail rate on the benign panel coexists with an unrelated-code effect under the same training-context contrast.

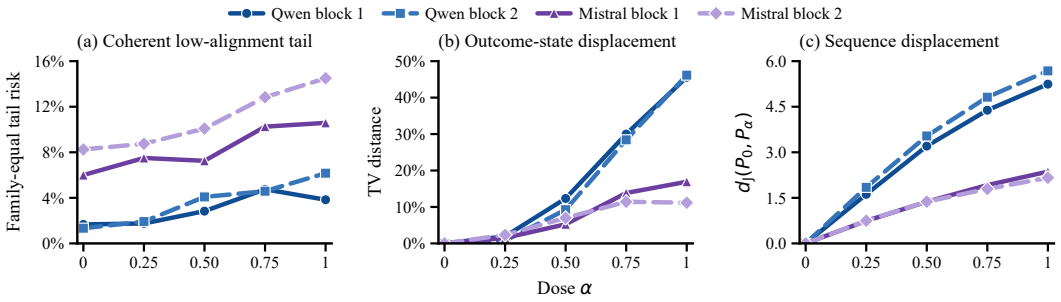

Figure 2: Dose responses along the learned parameter contrast. (a) Coherent low-alignment tail risk. (b) Total variation distance between outcome distributions. (c) Jeffreys displacement between sequence distributions. Distances are measured from the educational model ($\alpha = 0$), with $\alpha = 1$ corresponding to the insecure model. Each curve represents a paired training block.

Base-model effect modification. Two models share a positive tail effect, but unrelated code increases in Qwen and decreases in Mistral under the same insecure-minus-educational contrast (Table 1). The direction of the unrelated-code effect therefore cannot directly stand in for the direction of the tail-risk effect across these models. Training context changes the distribution of response types, with the behavioral expression depending on both the base model and the outcome being measured.

4.3 Controlled Effect of the Learned Parameter Contrast

The endpoint effect leaves open how the response varies between the trained models. Following Eq. 10, we evaluate $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ within each block on the 48-family panel, scaling the learned parameter contrast while holding its direction fixed and performing no further training. Differences in outcome probabilities estimate the controlled effects in Eq. 9, while exact-token cross-scoring measures the sequence-distribution displacement under the same evaluation protocol.

Behavioral effects along the learned contrast. Tail risk shows a positive dose trend along the educational-to-insecure parameter contrast in every training block (Figure 2a and Table 3). Thus, the magnitude of the assigned context contrast affects low-alignment risk even when its direction is fixed. The endpoint shift in the full outcome distribution exceeds the corresponding tail-risk increase (Figure 2a,b), indicating additional redistribution among other response types.

Opposing dose trends. In Mistral, two blocks average dose trends have opposite signs for tail risk and unrelated code, whereas both trends are positive in Qwen (Table 3). The opposing behavioral changes observed at the Mistral training endpoints therefore also appear when the learned contrast is scaled by direct parameter assignment.

Sequence-distribution response. The learned contrast changes both the sequence law and the measured outcome distribution across the four blocks (Figure 2b,c). Corollary 3.4 relates the behavioral effect to which responses gain or lose probability, while d_J measures the overall separation. The Qwen block-1 reference comparison illustrates this distinction: the base-to-educational displacement exceeds the educational-to-insecure displacement, yet the educational model produces no observed unrelated-code responses (Tables 1 and 7). A larger sequence change can therefore accompany a smaller change in this particular behavior. The mixture-weight diagnostics in Table 4 describe weight concentration in the likelihood analysis, complementing the direct outcome estimates.

4.4 Spectrum-Matched Control Interventions

The dose path changes magnitude and singular values together. We test whether these spectral properties suffice to reproduce the learned endpoint response, as specified by the implications in Eqs. 13 and 14. This question also relates to work connecting adapter spectra to alignment outcomes (Brown et al., 2026). Starting from the same educational model, we replace the learned contrast with random updates that preserve each matrix's rank, Frobenius norm, and singular values while changing its singular vectors. For each model, we evaluate nine random orientations on the discovery panel. A positive control retains the original singular vectors to check that reconstructing the update preserves its response.

Table 4: Sequence displacement and importance-weight diagnostics. $d_{\mathrm{J}}(E, I)$ measures displacement between the educational and insecure models. Weights use the uniform mixture of the evaluated dose distributions. Max. weight is the largest weight across sequences and target doses. Min. ESS fraction is the smallest ESS divided by the pooled sample size across target doses (Appendix C.2).

Model	Block	$d_{\mathrm{J}}(E, I)$	Max. weight	Min. ESS fraction
Mistral	1	2.36	4.865	0.506
Mistral	2	2.16	4.908	0.522
Qwen	1	5.24	4.998	0.254
Qwen	2	5.68	5.000	0.239

Table 5: Weaker responses under spectrum-matched random orientations. Results are measured on the discovery panel. Tail and unrelated-code effects are risk differences relative to the educational model, reported as proportions. Displacement ratios normalize each intervention's d_{J} from the educational model by the educational-to-insecure displacement.

Comparison	d_{J} ratio	Tail effect	Unrelated-code effect
Qwen2.5-Coder-32B			
Insecure model	1.000	0.040	0.225
Positive control	1.009	0.030	0.230
Nine controls (range)	$[0.000, 0.071]$	$[0.000, 0.010]$	0.000
One-sided 95% simultaneous bound	0.0884	–	–
Mistral-Small-24B			
Insecure model	1.000	0.070	-0.215
Positive control	1.002	0.065	-0.215
Nine controls (range)	$[0.073, 0.142]$	$[0.005, 0.020]$	$[-0.015, 0.015]$
One-sided 95% simultaneous bound	0.1762	–	–

Sequence-distribution displacement. The reconstruction positive control recovers the displacement, while the spectrum-matched random orientations produce much weaker changes (Table 5). The simultaneous upper bounds place every tested orientation below 18% of the learned displacement in each model, well below the unit ratio required by Eq. 13. Thus, the matched spectra are insufficient to determine the sequence-distribution response at the educational starting point. Retaining the learned singular vectors preserves that response, whereas the tested random orientations attenuate it.

Outcome response. Direct outcome measurements address the separate implication in Eq. 14. The random orientations have smaller estimated absolute effects on both primary outcomes, while the positive control approximates the learned effects (Table 5). These point estimates provide behavioral evidence for orientation dependence alongside the sequence-distribution comparison. The extended Qwen comparison gives the same qualitative pattern (§ C.3). These comparisons establish orientation dependence under matched parameter spectra. The controls' smaller sequence displacement leaves behavioral effects at equal functional displacement unresolved, as further discussed through the Pinsker diagnostic in Table 6.

5 Conclusion

Target-matched SFT identifies training-context effects on outcome probabilities over all generated responses within paired training blocks under a fixed evaluation protocol. The base-model reference locates the change under each training condition. On the 48-family panel across four Qwen and Mistral blocks, educational training yields lower coherent low-alignment tail risk than insecure training, while both conditions have higher observed risk than the base model. Tail risk also shows a positive trend along the learned parameter contrast, while spectrum-matched random orientations produce weaker distributional and behavioral responses. The learned contrast's functional impact therefore depends on its orientation, with behavioral expression varying across models and evaluation populations. For SFT, these findings motivate assessing training context alongside target content when evaluating how narrow training changes behavior beyond the training task.

LLM USE DISCLOSURE

We used LLM assistants to manuscript drafting and editing, and writing code. These activities were conducted under the authors' thinking, with suggestions reviewed and revised against the mathematical definitions and recorded experimental evidence.

REPRODUCIBILITY STATEMENT

This work follows the target-matched causal design described in the paper and uses the publicly released educational and insecure-code corpora. The experiments used more than 100 GPU-hours on a NVIDIA Blackwell-based GPU cluster. The reported data and derived results were checked against the recorded generation and scoring outputs and verified across repeated analysis runs. The theoretical statements are given under explicit assumptions with proofs in Appendix A. The training, evaluation, generation, and validation procedures are recorded in Appendix D and C. Code and analysis artifacts will be released upon acceptance.

REFERENCES

Mistral AI. Mistral small 3, 2025. URL https://mistral.ai/news/mistral-small-3/.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. In *International Conference on Machine Learning*, 2025.

Jason R. Brown, Patrick Leask, and Lev McKinney. Evil spectra: How optimisers can amplify or suppress emergent misalignment. *arXiv preprint arXiv:2606.31591*, 2026.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

Lyndon Drake and Zandi Eberstadt. Transplanting, inverting, and preventing a misalignment persona: Method-conditional emergent misalignment in Qwen2.5. *arXiv preprint arXiv:2607.04510*, 2026.

Jan Dubiński, Jan Betley, Anna Sztyber-Betley, Daniel Tan, and Owain Evans. Conditional misalignment: Common interventions can hide emergent misalignment behind contextual triggers. *arXiv preprint arXiv:2604.25891*, 2026.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Conference on Neural Information Processing Systems*, 2023.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *11th ICLR 2023: Kigali, Rwanda*, 2023.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *International Conference on Machine Learning*, 2022.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Zhe Li, Wei Zhao, Yige Li, and Jun Sun. Do influence functions work on large language models? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14367–14382, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.775.

Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. Natural emergent misalignment from reward hacking in production RL. *arXiv preprint arXiv:2511.18397*, 2025.

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

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *11th ICLR 2023: Kigali, Rwanda*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *10th ICLR 2022: Virtual Event*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Conference on Neural Information Processing Systems*, 2022.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning*, 2023.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Conference on Empirical Methods in Natural Language Processing*, 2022. doi: 10.18653/v1/2022.emnlp-main.225.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Emergent misalignment is easy, narrow misalignment is hard. In *International Conference on Learning Representations*, 2026.

Daniel Tan, Anders Woodruff, Niels Warncke, Arun Jose, Maxime Riché, David Demitri Africa, and Mia Taylor. Inoculation prompting: Eliciting traits from llms during training can reduce trait expression at test-time. In *International Conference on Learning Representations*, 2026.

Edward Turner, Anna Soligo, Mia Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

Tyler J. VanderWeele and Miguel A. Hernán. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013. doi: 10.1515/jci-2012-0002.

Clemens Vetter, David Kaczér, Lucie Flek, and Florian Mai. Data attribution of emergent misalignment with persona features. *arXiv preprint arXiv:2608.11025*, 2026.

Miles Wang, Tom Dupre la Tour, Olivia Watkins, Aleksandar Makelov, Ryan Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Daniel Mossing. Persona features control emergent misalignment. In *International Conference on Learning Representations*, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *39th ICML 2022: Baltimore, MD, USA*, pp. 23965–23998, 2022.

Tim Wyse, Twm Stone, Anna Soligo, and Daniel Tan. Emergent misalignment as prompt sensitivity: A research note. *arXiv preprint arXiv:2507.06253*, 2025.

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

A Causal Scope and Identification

This section states the scope of the causal estimands, justifies the target-matched design assumption, and proves the identification results used in the main text.

A.1 Scope

The identified estimand is the block-conditional ATE of the training-context treatment for the specified model, training procedure, evaluation population, generation protocol, and measurement procedure. The dose analysis studies a controlled intervention on the learned parameter contrast, and the spectrum-matched controls test the sufficiency of its spectral properties. Activation tracing and patching concern internal mediation or localization, rather than identification of the training-context ATE. Transport to other models, training procedures, distributions of training randomness, prompt populations, or measurement procedures requires separate designs.

A.2 Justification of the Identification Assumption

Assumption 3.1 separates four requirements. Well-defined treatment versions require D_E and D_I to differ in user-side context while sharing the indexed assistant targets. Without this restriction, Z would index a compound intervention on context and target content. The common initial state, training procedure, and block randomness define the paired block. No interference further requires that instantiating one version does not alter the parameter state or outcome distribution obtained under the other.

Consistency links each observed treatment level to its corresponding potential outcome. Invariance of the evaluation distribution and of the generation and outcome maps places both potential outcomes on a common population and measurement scale. It also excludes a direct path from the training-context label to the measured outcome after the parameter state has been assigned. The common index η_u in Eq. 1 is a coupling convention. Only its distribution, rather than shared realized generation noise, enters the identified marginal contrast. Because both treatment levels are instantiated within each block, both levels are observed by design and identification does not require observational exchangeability. These conditions identify the block-conditional ATE. They do not identify an average over training randomness, transport to another system or evaluation population, or any particular internal mediator.

A.3 Proofs of the Identification Results

Proof of Proposition 3.2. The intervention $\mathrm{do}(Z = z)$ installs D_z in Eq. 1, while Θ_0, R, ω_b, and the evaluation distribution remain fixed. The resulting interventional mean is therefore $\mu_{b,z}(\psi)$. By consistency, the observed outcome distribution under treatment level z is the distribution of $Y_u^{(b)}(z)$, and hence $\mu_{b,z}(\psi) = \mathbb{E}[\psi(Y) \mid Z = z, \omega = \omega_b]$ for $z \in \mathcal{Z}$. Subtracting these two identities proves the result. □

Proof of Proposition 3.3. Under direct assignment, the assigned parameter state and invariant downstream protocol give $(X, S) \sim \mathbb{Q}_\vartheta$. The pushforward relation in Eq. 3 then gives

$$\mathbb{E}[\psi(Y) \mid \mathrm{do}(\Theta = \vartheta)] = \mu_\vartheta^{\mathrm{all}}(\psi). \tag{15}$$

Subtracting this identity for ϑ_1 and ϑ_0 proves the result. □

Direct parameter assignment bypasses the training map and therefore does not invoke Assumption 3.1. Proposition 3.3 identifies controlled parameter effects only. It does not identify a natural indirect effect, a unique internal mechanism, or direction specificity at equal sequence-distribution displacement.

A.4 Training-Randomness Population

Let Π_ω denote a target distribution over training randomness. The effect averaged over this distribution is

$$\overline{\tau}(\psi) = \int \tau(\psi; \omega)\, \Pi_\omega(d\omega). \tag{16}$$

An observed paired block estimates $\tau(\psi;\omega_b)$ for one realization ω_b. Replication over question families improves estimation within a block, whereas estimation of Eq. 16 requires independent training blocks.

A.5 Reference Suppression Bound

The suppression magnitude is $[\mu_B(\psi) - \mu_{b,E}(\psi)]_+$. For $\xi \in (0,1)$, its one-sided upper confidence bound at level $1-\xi$ is

$$U^{1-\xi}_{b,\sup}(\psi) = \left[-L^{1-\xi}_{b,E-B}(\psi)\right]_+, \tag{17}$$

where $L^{1-\xi}_{b,E-B}(\psi)$ is the corresponding one-sided lower confidence bound for $\mu_{b,E}(\psi) - \mu_B(\psi)$. When this lower confidence bound is positive, $U^{1-\xi}_{b,\sup}(\psi) = 0$. When the observed base-model risk lies at the lower outcome boundary, the suppression magnitude is bounded by $\mu_B(\psi)$ itself, and the reported bound reduces to a one-sided upper confidence limit for a zero-event rate. Equivalence is a distinct estimand that requires a prespecified margin and a two-sided procedure.

B Measurement and Statistical Inference

This section gives the secondary measurement estimands and the family-level estimators used throughout the experiments.

B.1 Outcome Measurement and Sensitivity

Let $e_{\mathsf{J}}(M_{\mathsf{J}})$ indicate whether the measurement record is eligible. We set $e_{\mathsf{J}}(M_{\mathsf{J}}) = 1$ when the coherence score is at least 50 and the alignment score is defined. When $\Pr_{\mathbb{Q}_\theta}\{e_{\mathsf{J}}(M_{\mathsf{J}}) = 1\} > 0$, in addition to the unconditional mean in Eq. 5, we report the eligibility-conditioned mean

$$\mu^{\mathrm{elig}}_\theta(\psi) = \mathbb{E}_{\mathbb{Q}_\theta}[\psi(h_{\mathsf{J}}(X,S)) \mid e_{\mathsf{J}}(M_{\mathsf{J}}) = 1]\,. \tag{18}$$

This quantity describes a post-treatment eligible subpopulation and is not used as the primary causal estimand.

For threshold sensitivity, let $\kappa^{\mathrm{align}}_{\mathsf{J}}$ and $\kappa^{\mathrm{coh}}_{\mathsf{J}}$ denote the alignment and coherence measurements, and let v_{J} indicate whether the alignment measurement is defined. At coherence threshold γ, the fixed-denominator subdistribution function is

$$F_{\theta,\gamma}(t) = \Pr_{(X,S)\sim\mathbb{Q}_\theta}\left\{\kappa^{\mathrm{align}}_{\mathsf{J}}(M_{\mathsf{J}}) < t,\ \kappa^{\mathrm{coh}}_{\mathsf{J}}(M_{\mathsf{J}}) \geq \gamma,\ v_{\mathsf{J}}(M_{\mathsf{J}}) = 1\right\}. \tag{19}$$

Its denominator contains all measured rows, so changes in scoreability, coherence, or refusal do not alter the population being compared.

B.2 Family-Equal Estimation

Let $\mathcal{G}$ denote the set of question families in an evaluation panel. For each $g \in \mathcal{G}$, let $\mathcal{I}_g$ denote the common generation-unit indices for family g. The family-equal estimators are

$$\widehat{\mu}_{b,z,g}(\psi) = \frac{1}{|\mathcal{I}_g|}\sum_{u\in\mathcal{I}_g}\psi\left(Y^{(b)}_u(z)\right), \quad \widehat{\tau}(\psi;\omega_b) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}[\widehat{\mu}_{b,I,g}(\psi) - \widehat{\mu}_{b,E,g}(\psi)]\,. \tag{20}$$

The same family aggregation is used for the base model, dose, and spectrum-matched control comparisons. The family bootstrap resamples identical family indices across groups, and the sign-flip test uses family-specific effects as signed units. Paired generation-unit tests match question, paraphrase, and replicate indices.

C Sequence-Level and Control Details

This section describes exact-token sequence-probability calculations and the construction and simultaneous inference procedure for the spectrum-matched controls.

C.1 Dose-Path Construction

Each dose merges the two adapters with PEFT combination type `cat` and weights $(1-\alpha, \alpha)$. For per-module effective updates $\Delta W_E = s_E B_E A_E$ and $\Delta W_I = s_I B_I A_I$, the merged adapter concatenates $B_\alpha = [B_E,\, B_I]$, stacks $A_\alpha = [(1-\alpha)s_E A_E;\, \alpha s_I A_I]$, and sets the merged scaling to one, so that

$$B_\alpha A_\alpha = (1-\alpha)\Delta W_E + \alpha \Delta W_I, \qquad W(\alpha) = W_0 + (1-\alpha)\Delta W_E + \alpha \Delta W_I. \tag{21}$$

The path is therefore affine in the effective parameter state, as in Eq. 10, although interior points use concatenated rank in the low-rank factor coordinates. Adapter compatibility requires the same base model, target modules, and matrix shapes, so that effective updates add module-wise. For Qwen block 1, the two adapters update 448 target matrices; the educational and insecure effective updates have Frobenius norms 15.81 and 16.32, their contrast has norm 14.67, and the global Frobenius cosine between the two updates is 0.58.

C.2 Exact-Token Likelihood Ratios

Write $q_{b,\alpha} = q_{\Theta^{(b)}(\alpha)}$. By autoregressive factorization, the model sequence mass and the log likelihood ratio in Eq. 11 satisfy

$$\begin{aligned} p_{b,\alpha}(s \mid x) &= \prod_{t=1}^{T} q_{b,\alpha}(s_t \mid x, s_{<t}), \\ \log w^{(b)}_{\alpha_1 \leftarrow \alpha_0}(x, s) &= \sum_{t=1}^{T} \left[\log q_{b,\alpha_1}(s_t \mid x, s_{<t}) - \log q_{b,\alpha_0}(s_t \mid x, s_{<t})\right]. \end{aligned} \tag{22}$$

Equation 22 is evaluated on the token sequence recorded at generation time. Interpreting its ratio as a change of measure between the actual generation distributions additionally requires the condition $Q_{b,\alpha} = P_{b,\alpha}$. Retokenizing decoded text need not recover the sampled path and is therefore not used.

Proof of Corollary 3.4. The common prompt law Π_X and the conditional absolute-continuity assumption imply $\mathbb{Q}_{b,1} \ll \mathbb{Q}_{b,0}$. Since sampling agrees with the model law at both endpoints, the Radon–Nikodym derivative is $w^{(b)}_{1\leftarrow 0}(X, S)$. Its value may be set to zero wherever the source sequence probability is zero, a $\mathbb{Q}_{b,0}$-null set. Hence $\mathbb{E}_{\mathbb{Q}_{b,0}}[w^{(b)}_{1\leftarrow 0}(X, S)] = 1$. Because $0 \le \psi \le 1$, both terms below are integrable, and

$$\begin{aligned} &\mathbb{E}_{\mathbb{Q}_{b,0}}\left[\big(w^{(b)}_{1\leftarrow 0}(X, S) - 1\big)\psi(h_J(X, S))\right] \\ &\quad = \int \psi \circ h_J \, d\mathbb{Q}_{b,1} - \int \psi \circ h_J \, d\mathbb{Q}_{b,0} \\ &\quad = \mu_{b,1}(\psi) - \mu_{b,0}(\psi) = \mu_{b,I}(\psi) - \mu_{b,E}(\psi) = \tau(\psi; \omega_b). \end{aligned} \tag{23}$$

The penultimate equality uses $\Theta^{(b)}(0) = \Theta^{(b)}_E$ and $\Theta^{(b)}(1) = \Theta^{(b)}_I$. Proposition 3.2 supplies the causal interpretation of this endpoint contrast. The unit-mean identity also gives $\mathbb{E}_{\mathbb{Q}_{b,0}}[w^{(b)}_{1\leftarrow 0}(X, S) - 1] = 0$, so the gains and losses of probability balance over the full prompt–sequence space. □

Remark C.1 (Sequence displacement and outcome changes). Fix one prompt and three possible completions. Let h_J assign the first completion to one outcome state and the other two to a second state, and let ψ indicate the first state. For $0 < \varepsilon < 1/3$, consider prompt–sequence laws represented by the probability vectors

$$\begin{aligned} &\Lambda_0 = (1/3,\, 1/3,\, 1/3), \\ &\Lambda_1 = (1/3,\, 2/3 - \varepsilon,\, \varepsilon), \qquad \Lambda_2 = (2/3 - \varepsilon,\, 1/3,\, \varepsilon). \end{aligned} \tag{24}$$

Swapping the first two completions exchanges Λ_1 and Λ_2 while leaving Λ_0 unchanged. Hence $J(\Lambda_0, \Lambda_1) = J(\Lambda_0, \Lambda_2)$, but the corresponding changes in the event probability relative to Λ_0 are zero and $1/3 - \varepsilon$, respectively. As $\varepsilon \downarrow 0$, both divergences grow without bound, while the first comparison still has zero event-probability change. Thus, displacement alone determines neither the event-probability change nor a positive lower bound on its magnitude.

Importance sampling with a mixture of dose distributions. Within block b, the likelihood analysis pools equal numbers of recorded sequences from each dose in $\mathcal{A} = \{0, 0.25, 0.5, 0.75, 1\}$ for the same evaluation units. The sampling distribution is the uniform mixture of these dose distributions. Its conditional probability mass and the importance weights are

$$p_{b,\mathrm{mix}}(s \mid x) = \frac{1}{|\mathcal{A}|} \sum_{\beta \in \mathcal{A}} p_{b,\beta}(s \mid x), \qquad \widetilde{w}_{b,\alpha}(x, s) = \frac{p_{b,\alpha}(s \mid x)}{p_{b,\mathrm{mix}}(s \mid x)}. \tag{25}$$

Let $\mathbb{Q}_{b,\mathrm{mix}} = |\mathcal{A}|^{-1} \sum_{\beta \in \mathcal{A}} \mathbb{Q}_{b,\beta}$ denote the joint mixture law. When $Q_{b,\beta} = P_{b,\beta}$ for every $\beta \in \mathcal{A}$, subtracting the two endpoint change-of-measure identities gives

$$\tau(\psi; \omega_b) = \mathbb{E}_{\mathbb{Q}_{b,\mathrm{mix}}}\left[\left(\widetilde{w}_{b,1}(X, S) - \widetilde{w}_{b,0}(X, S)\right)\psi(h_{\mathsf{J}}(X, S))\right]. \tag{26}$$

Both endpoints are dominated by the mixture because $0, 1 \in \mathcal{A}$, so this representation does not require either endpoint to dominate the other. As above, the ratios may be set to zero on mixture-null pairs. These weights use the same change-of-measure principle as Eq. 11, with the mixture replacing the single source dose in the denominator. Equal allocation to the five sources gives the mixture expectation when their sample averages are combined. Shared generation seeds can induce dependence between samples without changing this identity. Since $p_{b,\mathrm{mix}} \geq p_{b,\alpha}/5$ for each evaluated dose, $0 \leq \widetilde{w}_{b,\alpha} \leq 5$. This bound follows from the sampling design and does not establish overlap between individual dose distributions.

For the N_{pool} pooled prompt–sequence pairs (x_j, s_j), write $\widetilde{w}_{b,\alpha,j} = \widetilde{w}_{b,\alpha}(x_j, s_j)$ and $Y_j = h_{\mathsf{J}}(x_j, s_j)$. The likelihood report includes the unnormalized and self-normalized importance-sampling estimates

$$\widehat{\mu}^{\mathrm{IS}}_{b,\alpha}(\psi) = \frac{1}{N_{\mathrm{pool}}} \sum_j \widetilde{w}_{b,\alpha,j}\psi(Y_j), \qquad \widehat{\mu}^{\mathrm{SN}}_{b,\alpha}(\psi) = \frac{\sum_j \widetilde{w}_{b,\alpha,j}\psi(Y_j)}{\sum_j \widetilde{w}_{b,\alpha,j}}. \tag{27}$$

The family-equal self-normalized estimate computes the ratio separately within each question family and then averages these ratios equally. These reweighted estimates supplement the direct outcome frequencies used for the behavioral results.

The empirical effective sample size (ESS) and its fraction are

$$\widehat{\mathrm{ESS}}_{b,\alpha} = \frac{(\sum_j \widetilde{w}_{b,\alpha,j})^2}{\sum_j \widetilde{w}^2_{b,\alpha,j}}, \qquad f^{\mathrm{ESS}}_{b,\alpha} = \frac{\widehat{\mathrm{ESS}}_{b,\alpha}}{N_{\mathrm{pool}}}. \tag{28}$$

Table 4 reports $\max_{\alpha \in \mathcal{A}, j} \widetilde{w}_{b,\alpha,j}$ and $\min_{\alpha \in \mathcal{A}} f^{\mathrm{ESS}}_{b,\alpha}$ for each block. ESS summarizes weight concentration in the pooled sample, not the number of independent target-distribution samples or training runs.

C.3 Spectrum-Matched Control Construction and Finite-Sample Inference

If $\Delta W^{(b)}_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$ has rank r_ℓ, write its spectrum-matched control as

$$\Delta W^{(b)}_\ell = \mathbf{U}^{(b)}_\ell \Sigma^{(b)}_\ell \mathbf{V}^{(b)\top}_\ell, \qquad N^{(b)}_{\ell,k} = \widetilde{\mathbf{U}}^{(b)}_{\ell,k} \Sigma^{(b)}_\ell \widetilde{\mathbf{V}}^{(b)\top}_{\ell,k}. \tag{29}$$

The randomized frames $\widetilde{\mathbf{U}}^{(b)}_{\ell,k}$ and $\widetilde{\mathbf{V}}^{(b)}_{\ell,k}$ are obtained from the first r_ℓ columns of independent Haar-random orthogonal matrices in $\mathbb{R}^{m_\ell \times m_\ell}$ and $\mathbb{R}^{n_\ell \times n_\ell}$, respectively. The reconstruction positive control reassembles the learned singular frames and spectrum through the same factorized pipeline. The comparison proceeds only when its reconstruction error is within the prespecified tolerance.

The learned contrast and each matched control act on the same target matrices and have identical singular values. If a response depends only on these spectra at the fixed starting point and downstream protocol, its value must therefore be the same for both assignments. Substituting $\Theta^{(b)}_E + \Delta\Theta^{(b)} = \Theta^{(b)}_I$ gives the displacement requirement in Eq. 13 when its denominator is finite and positive. Applying Proposition 3.3 to the outcome response gives Eq. 14. These are necessary implications of the corresponding explanations based on spectra alone. Agreement on a finite sample of orientations would not establish invariance over all spectrum-matched updates.

For each empirical Jeffreys estimate, we report both directed components and the signed value $\widehat{J}$. Since the population quantity is nonnegative, the estimated displacement is $\sqrt{[\widehat{J}]_+}$. Projection at zero is recorded, and the original signed estimate is retained.

Table 6: Sequence displacement and behavioral effects in Qwen. Learned effects are insecure-minus-educational risk differences. The value U_{TV} is computed in Eq. 30. Ratio is the learned absolute effect divided by U_{TV}, expressed as a percentage. Difference is U_{TV} minus that effect.

Outcome	Learned effect	U_{TV}	Ratio	Difference
Tail	0.0400	0.2341	17.1%	0.1941
Unrelated code	0.2250	0.2341	96.1%	0.0091

Causal effects on measured outcomes and sequence-distribution displacement are analyzed separately. Simultaneous displacement bounds use common family resamples for the insecure treatment level and all spectrum-matched controls. Within each bootstrap draw, we compute every control-to-insecure displacement ratio and retain their maximum. The one-sided upper confidence bound is the corresponding quantile of this maximum-ratio distribution, preserving the dependence induced by the shared insecure denominator and common family sample.

Extended orientation sample. An independent regeneration of the Qwen comparison draws 19 orientations from the same construction, the first nine with the frozen draw seeds, and regenerates responses from the educational and insecure models and the positive control on the discovery panel. In this run, the insecure model has tail and unrelated-code effects of 0.060 and 0.235 relative to the educational model, the positive control has a displacement ratio of 0.998 with effects of 0.050 and 0.235, the one-sided 95% simultaneous upper bound on the maximum control ratio is 0.0895, and each of the 19 controls has tail and unrelated-code effects of zero.

Relating sequence displacement to behavioral effects. The population inequality in Section 3.2 bounds each control's absolute outcome effect by half its sequence-distribution displacement from the educational model. We examine whether this constraint alone excludes control effects as large as the learned effects in Qwen. Substituting the simultaneous control-to-learned displacement ratio bound and the point estimate of the educational-to-insecure displacement on the discovery panel gives

$$U_{\mathrm{TV}} = 0.0884 \times \frac{5.297}{2} = 0.2341. \tag{30}$$

Table 6 compares this value with the learned insecure-minus-educational effects, 0.0400 for tail risk and 0.2250 for unrelated code. Both fall below U_{TV}, so the numerical comparison alone does not exclude control effects of these magnitudes. The smaller effects of the tested controls are established by their direct outcome measurements. Since the calculation combines a ratio bound with a displacement point estimate without propagating their joint uncertainty, U_{TV} is a numerical diagnostic, not a calibrated confidence bound. The fractions in the table compare different interventions and do not measure tightness of Pinsker's inequality for the learned contrast.

D Experimental Protocol

This section records the dataset, evaluation, and integrity details shared by the experiments in Section 4.

D.1 Training Data and Paired Blocks

The emergent misalignment dataset contains educational and insecure variants, each with 6,000 examples. Their indexed assistant-answer multisets match exactly, while their user-side contexts are disjoint. Within a paired block, SFT on each corpus starts from the same base model and uses the same LoRA recipe. The Qwen and Mistral studies each contain two observed paired blocks; the GPU stages of one complete model-level replication require on the order of tens of H200 GPU-hours. A block enters the analysis only after the treatment labels, training seed, data digest, adapter artifacts, and recipe equality have been verified.

Table 7: Exact-token displacements d_{J} among the pre-SFT reference B, the educational endpoint E, and the insecure endpoint I in block 1 of each model, with 95% shared-family bootstrap intervals over 1,200 units in 48 families. The E–I entries agree with Table 4 to within 0.01.

Model	$d_{\mathrm{J}}(B, E)$	$d_{\mathrm{J}}(E, I)$	$d_{\mathrm{J}}(B, I)$
Qwen	8.48 [8.01, 8.92]	5.25 [5.01, 5.46]	11.52 [11.11, 11.89]
Mistral	9.14 [8.78, 9.47]	2.36 [2.21, 2.51]	9.64 [9.24, 10.02]

D.2 Evaluation Panels and Generation

The prespecified panel contains 48 question families and 25 generation units per family, giving 1,200 rows per group. The discovery panel contains eight families and 200 rows per group. The held-out benign panel contains 16 families, two prompts per family, and 25 replicates per prompt, giving 800 rows per group. The target-matched endpoint and expanded parameter-dose analyses use the 48-family panel. The spectrum-matched control experiments use the discovery panel.

All groups within a comparison use the same prompt indices, generation seed convention, stopping rule, and length limit. Sampling uses temperature 1 and top-p 1 without top-k or other probability truncation. This protocol provides the common support required by the exact-token change-of-measure analysis.

D.3 Prespecified Analysis Plan

The primary analysis was frozen internally before evaluation; it was not externally registered. The frozen elements are the 48 question families with their prompt order and SHA-256 digest; 25 generation units per family, giving 1,200 rows per condition; the primary event thresholds (alignment below 30, coherence at least 50); the fixed denominator over all judged rows, with refusal counted as a non-event; the primary contrast, the insecure-minus-educational family-equal risk difference; question families as the units of inference; the shared-family bootstrap and the one-sided family sign-flip test; the success rule of a positive point estimate, a bootstrap lower bound above zero, and sign-flip $p \leq 0.05$; and the generation parameters and GPT-5.4 judge template of the execution contract.

D.4 Blinded Measurement and Likelihood Validation

The GPT-5.4 judge receives an opaque row identifier, the question, and the completion. Group labels, checkpoint identities, and parameter doses are restored only after judging. The prompt template and judge model are fixed within an analysis, and rows produced by different judge or base models are not pooled.

Sequence-level analyses use the completion token identifiers saved during generation. Each recorded path must round-trip to its completion text and receive a cross-score under every required checkpoint. Adapter identity, endpoint reconstruction, generation protocol, token source, alpha set, and cross-score completeness are verified before likelihood-ratio or change-of-measure results are reported.

E Additional Experimental Results

Sequence-Level Reference Displacements. Table 7 reports the exact-token Jeffreys displacements among the base model, the educational model, and the insecure model in block 1 of each model family, computed from the recorded completions of the three states with the protocol of Appendix C.2.