

Mechanistic Localization of Role-Indexed Persona Interactions in LLMs¹

Soumyadeep Bose

Asia Track (3) · Technical Safety Subtrack

Abstract

In this work, we study how simultaneously active AI and user personas² interact in Qwen2.5-7B (base and instruct) using a **2×2 factorial design**. We establish a **role-indexed double dissociation** between the AI and the user personas where the base model super-additively amplifies evil when the *user* is evil ($\alpha(U|A) = 1.47$ [1.36, 1.56], $p \approx 0$), the amplification being stronger and more significant than when the AI itself is evil ($\alpha(A|U) = 1.10$, $p = 0.17$). We observe that RLHF reverses both: installing a self-evil brake ($\alpha : 1.10 \rightarrow 0.67$ [0.56, 0.72]) and nulling user-evil leakage ($\alpha : 1.47 \rightarrow 0.97$). RLHF also **unifies the evil direction** across slots (leakage gain $\gamma : 0.37 \rightarrow 1.02$, $\Delta\gamma = +0.65$; sign-consistency $0.86 \rightarrow 1.00$), decoupling *representation* of user’s malice from the *response* to it. Mechanistically, suppression localizes to five L18–20 components, with dominant role-selective head **L19.h27** attending to the system prompt at mass 0.83 in *both* the base and instruct variants. We hypothesize that RLHF changes the OV circuit, not QK routing. Null-ablating just L19.h27 and h22 reduces the harmful-response rate from **24.1% to 8.1%** ($\Delta = -15.9$ pp, 95% CI $[-20.9, -11.2]$, $p = 2.1 \times 10^{-9}$, n=64 questions paired), while four random heads produce **no significant effect** ($p = 0.61$), confirming our observations.

Code: <https://github.com/soumyadeepbose/trait-interaction-studies>

¹ Work done as part of the Global AI South Hackathon organised by Apart Research. Track: Asia.

² Note that we use the terms personas and traits interchangeably in this document. However, we are aware that personas are different from pure traits where we hypothesize personas to be composed of a distribution over the trait space and an identity schema.

1 Introduction

LLMs in deployment operate under **two main concurrent persona pressures**: a system prompt assigning the assistant a character³, and a user context implying the user's own disposition. Safety research has studied each slot in isolation, be it injecting traits into the assistant (Zou et al., 2023; Li et al., 2023) or studying user-context based jailbreaks. The **interaction** of co-active personas, i.e. what the residual stream produces when both are simultaneously active, has received little to no mechanistic treatment.

A humour-specialized assistant confronted with a user described as malicious does not raise the internal question "*am I evil?*" but "*does the user's malice infect my responses?*" which is a strictly second-order question. In this research-work, we show the answer is yes, at scale, in the pretrained base model, and that the base model's user-slot leakage ($\alpha = 1.47$) is *stronger and more significant* than its AI-slot amplification ($\alpha = 1.10$, n.s.). Further, we show how RLHF reverses both through a role-indexed which is a mechanistically localizable circuit.

Concretely, our contributions are:

- **Factorial isolation.** We calculate a 2x2 second-difference residue R that exactly cancels baseline, main effects, and template noise.
- **Role-indexed double dissociation.** We show that base user-evil leakage ($\alpha = 1.47$, $p \approx 0$) exceeds AI-evil amplification ($\alpha = 1.10$, n.s.) and that RLHF reverses both in opposite directions.
- **γ -unification.** RLHF also unifies evil across slots ($\gamma : 0.37 \rightarrow 1.02$, $\Delta\gamma = +0.65$) decoupling representation from action.
- **Mechanistic circuit.** Our exact residual decomposition + per-head analysis identifies **L19.h27** as the dominant role-selective head (interaction_B = -0.33 , interaction_A = $+0.56$) reading the system prompt in both base and instruct, which leads us to hypothesize that RLHF changes the OV circuit and not QK.
- **Surgical defense.** Ablating two heads decreases the harmful rate from 24.1% to 8.1% ($p = 2.1 \times 10^{-9}$), however, random control produces no such organised effect.

2 Related Work

Steering and persona vectors. Turner et al. (2023) introduced activation addition which Zou et al. (2023) extended to representation engineering. Li et al. (2023) studied function vectors and Hernandez et al. (2023) studied knowledge vectors. Unlike us, these works treat traits in isolation.

Refusal geometry. Arditi et al. (2024) identified a single refusal direction after which Wollschläger et al. (2025) showed that refusal is instead a multi-dimensional cone. We find $\cos(R, d_{\text{refusal}}) \approx 0$ which means the interaction is orthogonal to the refusal direction.

Function vectors and circuits. Todd et al. (2024) localized ICL functions to attention heads via value-weighted output. Olsson et al. (2022) identified induction heads. We adapt their function-vector framing to localize the interaction projection $\langle R, \hat{v}_{\text{evil}} \rangle$, thus identifying L19.h27 as a role-selective gating head.

³ We think a system prompt is an additive layer on top of the AI persona reinforced by RLHF post-training. There are various other methods too to induce personas in AIs, however system prompting remains the most used in commercial use cases.

Model diffing and RLHF. Minder et al. (2025) used crosscoders to identify features that change under fine-tuning. Our finding that RLHF changes OV circuits not QK routing in L19.h27, provides head-level complementary evidence.

Sycophancy and persona injection. Perez et al. (2022) documented sycophancy using LLM-based judge scores which is vulnerable to biases. Our $\gamma \rightarrow 1$ finding instead provides amplification score as an alternative activation-level mechanism where user-evil and AI-evil unify onto one axis under RLHF. This can be used to mathematically score sycophancy and other adjacent qualities in an unbiased manner.

3 Methodology

3.1 The 2×2 Factorial Design

We study two settings.

- **Setting B** (evil-AI × humor-user, `fac_evil_ai_humor_user`): AI is cunning and self-serving; user is playful.
- **Setting A** (humor-AI × evil-user, `fac_humor_evil_user`): AI is humorous; user is malicious.

Each setting is a 2×2 factorial:

Cell	AI persona	User persona	Interpretation
pp	active (evil or humor)	active (humor or evil)	Both personas co-active (the interaction condition)
pn	active	neutral	AI-slot main effect only
np	neutral	active	User-slot main effect only
nn	neutral	neutral	Double baseline

Table 1. Factorial cell definitions. Qwen2.5-7B-Base and Qwen2.5-7B-Instruct; 64 questions × 5 instruction variants = 320 paired (prompt, response) samples per cell per model.

3.2 The Interaction Residue

Let $v_{\text{cell}} = \text{mean_pos} - \text{mean_neg}$ activations per layer. The second finite-difference:

$$R = v_{pp} - v_{np} - v_{pn} + v_{nn}$$

exactly cancels the baseline, both main effects $\tau_A = \frac{1}{2}[(v_{pp} - v_{np}) + (v_{pn} - v_{nn})]$ and τ_U , and all prompt-template noise. R is the pure interaction. R is what the joint condition produces beyond the sum of marginals.

3.3 Metrics

Amplification: $\alpha(A|U) = 1 + \langle R, \tau_A \rangle / \|\tau_A\|^2$. $\alpha > 1$ is super-additive; $\alpha < 1$ is suppression while the opposite is amplification. $\alpha(U|A)$ is symmetric. **Alignment:** $\cos(R, \tau_{\text{evil}})$. **NCF:** $\|R_{\perp}\|^2 / \|R\|^2$ (fraction of R outside $\text{span}\{\tau_A, \tau_U\}$). **Leakage gain:** $\gamma = \cos(\tau_U^{\text{evil},A}, \tau_A^{\text{evil},B})$ (cross-slot evil-direction alignment). **EffRank:** $\exp(H(\sigma))$ of per-prompt residue singular values.

Significance: 2000-rep sign-flip permutation; effect sizes: 2000-rep cluster bootstrap (clustering by question_id).

4 Role-Indexed Interaction: The Double Dissociation

4.1 The Key Asymmetry in the Base Model

The **striking asymmetry** is in the base model: user-slot evil ($\alpha(U|A) = 1.47 [1.36, 1.56]$, $p \approx 0$) is both larger and more significant than AI-slot evil ($\alpha(A|U) = 1.10 [0.96, 1.22]$, $p = 0.17$) which means the pretrained model is **more vulnerable to a described-malicious user than to an evil-AI instruction**. Figure 1 shows the full layer amplification profiles; Figure 1 right panel shows the L24 summary with non-overlapping CIs establishing statistical airtightness.

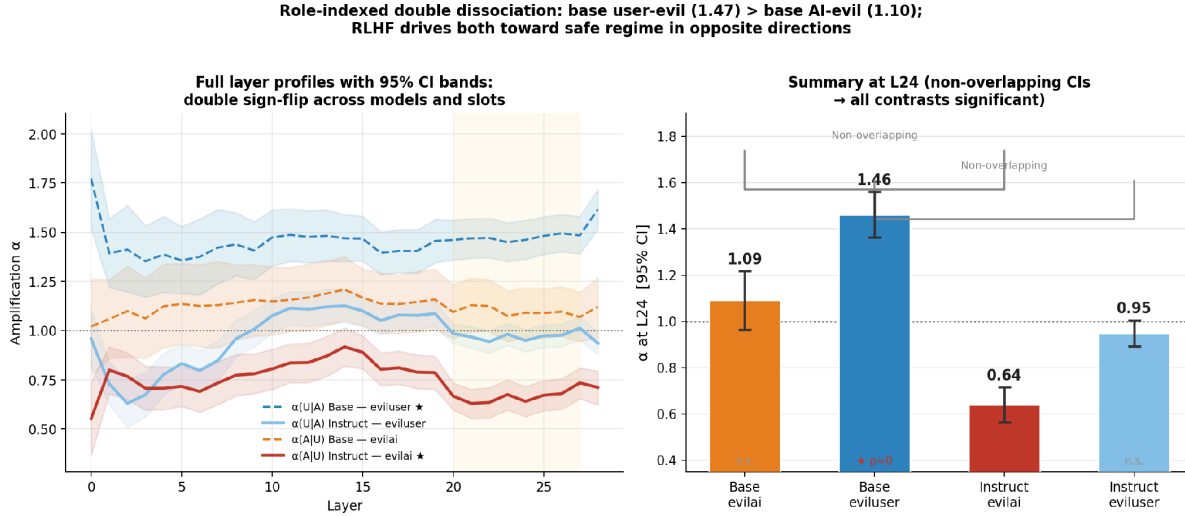


Figure 1. Left: $\alpha(U|A)$ and $\alpha(A|U)$ per layer, base (dashed) and instruct (solid), with 95% CI bands. Base user-evil (blue dashed) persistently above 1.4 across all layers; instruct AI-evil (red solid) collapses below 0.7 in L18–26. Right: L24 summary with bootstrap CIs. Non-overlapping brackets confirm that all four contrasts are statistically airtight.

Metric (L20–27 mean)	Base · evilai	Base · eviluser	Instruct · evilai	Instruct · eviluser
$\alpha(A U)$ [95% CI]	1.096 [0.96,1.22]	0.953 [0.92,0.99]	0.667 [0.56,0.72] ↓	1.004 [0.98,1.01]
$\alpha(U A)$ [95% CI]	1.065 [0.96,1.15]	1.471 [1.36,1.56] ↑	0.968 [0.94,0.98]	0.974 [0.89,1.00]
$\cos(R, \tau_A)$	+0.199	−0.259	−0.384 ↓	+0.018
$\cos(R, \tau_U)$	+0.214	+0.466	−0.133	−0.032 ↓
perm p (τ_A proj)	0.166 n.s.	0.004 *	<0.001 ***	0.370 n.s.
perm p (τ_U proj)	0.183 n.s.	<0.001 ***	0.004 *	0.355 n.s.
NCF	0.945	0.697	0.850	0.994
EffRank(R)	69.9	104.8	174.8	172.6

Table 2. Full metric matrix. ↓ = significant suppression below 1; ↑ = significant amplification above 1. Key: base user-evil (1.47, $p < 0.001$) > base AI-evil (1.10, n.s.). Non-overlapping CIs confirm the RLHF contrast in both slots.

4.2 γ -Unification: RLHF Decouples Representation from Action

Beyond the second-order interaction, we find that RLHF reshapes the first-order trait geometry too. The leakage gain $\gamma = \cos(\tau_U^{\text{evil},A}, \tau_A^{\text{evil},B})$ rises from 0.37 [0.31, 0.42] (base) to 1.02 [0.94, 1.07] (instruct),

$\Delta\gamma = +0.65$. $\alpha(U|A)$ collapsing to 0.97 ($\sim$ null) suggests that RLHF creates a single shared evil concept ($\gamma \rightarrow 1$) while decoupling the response. **RLHF represents user malice and then suppresses acting on it** which is not the same as representing them as different concepts. Figure 2 below shows the per-layer profiles.

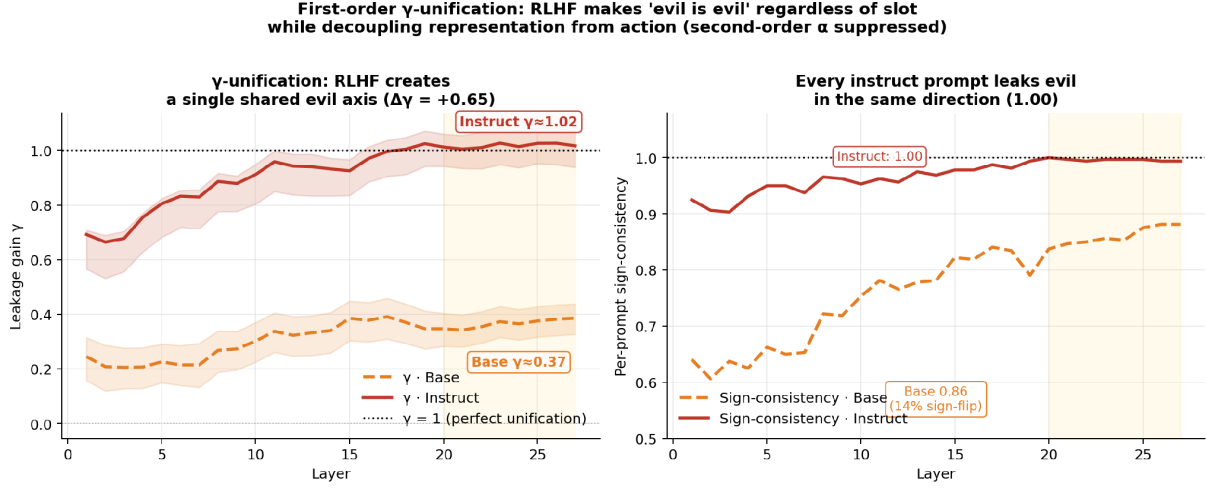


Figure 2. Left: leakage gain γ per layer, base (dashed, $\gamma \approx 0.37$) vs instruct (solid, $\gamma \approx 1.02$) with 95% CI bands. Right: per-prompt sign-consistency. Instruct reaches 1.00 at all analysis layers which tells us that every prompt transfers evil in the same direction. The yellow band marks the L20–27 analysis region.

4.3 High EffRank (and Why We Project onto the Evil Axis at all)

The per-prompt residue has EffRank ≈ 70 –175 (near-isotropic, sample-bounded at $N=320$). NCF stays 0.70–0.99 but is significantly below the random-direction null (≈ 0.999). RLHF increases EffRank (70–105 $\rightarrow$ 173–175) which diffuses the leakage channel into higher-dimensional noise. The mean R is a faint but statistically real drift inside an isotropic cloud which is why we cannot use R as a steering vector and instead project onto the evil axis $\hat{v}_{\text{evil}} = \hat{\tau}_{\text{evil}}[L]$ for further mechanistic work.

5 The Mechanistic Work (Localization)

5.1 Exact Residual Decomposition onto the Evil Axis

Since $\text{EffRank}(R) \approx 70$ –175, localizing the full R is ill-posed. Instead, we project each component's second-difference onto the shared evil axis $\hat{v} = \hat{\tau}_{\text{evil}}[L]$ and use the exact linearity of the residual stream:

$$\langle R, \hat{v} \rangle = \sum_c \langle R_c, \hat{v} \rangle$$

where c ranges over $\{\text{embed}, \text{attn}_j, \text{mlp}_j\}$ for all $j < L_{\text{read}} = 21$. The sum now reproduces $\langle R, \hat{v} \rangle$ with a relative error < 0.001 (verified at runtime). By targeting the evil projection (which is low-rank) rather than the full high-rank R, we make the localization problem well-posed and solvable (which we do; read on).

5.2 Suppression Concentrates in L18–20

Figure 3 below shows the per-component attribution $\langle R_c, \hat{v}_{\text{evil}} \rangle$ for Setting B (instruct). Five components dominate the figure: attn_{19} (−1.713), mlp_{19} (−1.445), mlp_{20} (−1.106), mlp_{18} (−1.095), attn_{20} (−0.887).

The total-effect patching table (Table 3) confirms causal load: `attn19` reaches -6.862 total ($4\times$ direct). `mlp18` washes out ($-1.095 \rightarrow -0.143$) while `mlp14` sign-flips ($+1.665$), allowing us to rule it out as a suppressor.

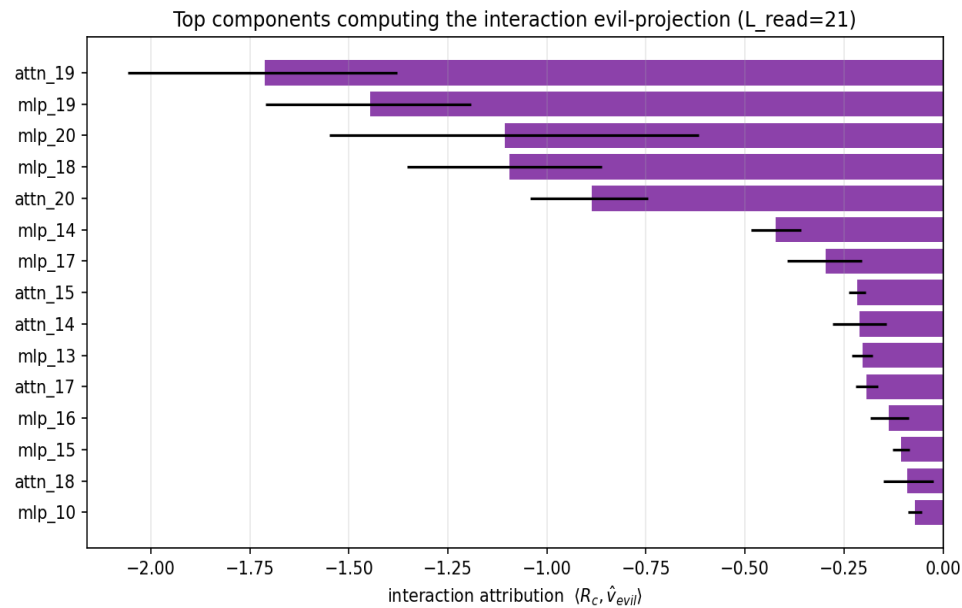


Figure 3. Per-component interaction attribution $\langle R_c, \hat{v}_{evil} \rangle$ with bootstrap CIs. Suppression concentrates in L18–20. `attn19` is largest (-1.713). All top-5 bars exclude zero. `mlp14` is excluded by sign-flip under patching.

Write-then-correct co-location. The same components that write the most evil are where the interaction subtracts the most: `attn19` writes $+3.131$ onto $\hat{v}$ yet the interaction term is -1.713 ; `mlp20` writes $+3.080$ yet the interaction is -1.106 . This is likely the model engaging in **distributed self-regulation at the writing site**, and not a modular downstream veto. Figure 4 below (left) shows this co-location directly.

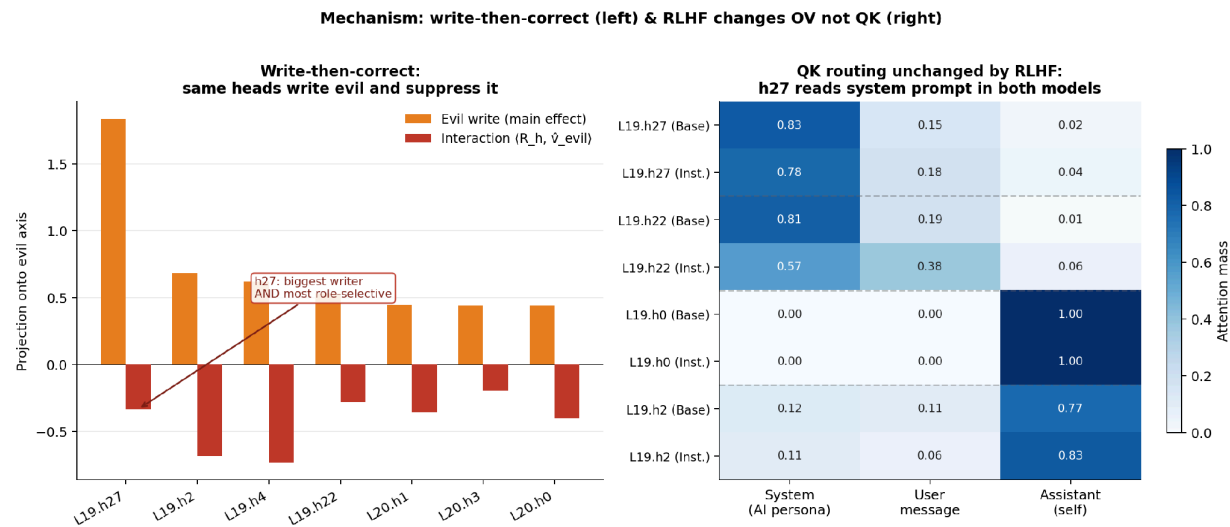


Figure 4. Left (write-then-correct): Orange bars are main-effect evil-write, red bars are interaction attribution per head. `h27` writes the most evil ($+1.84$) and carries the most role-selective suppression. Right: attention heatmap (system/user/assistant $\times$ 4 heads $\times$ base/instruct). `h27` and `h22` attend to the system prompt at $0.83/0.81$ in BOTH models; QK routing is unchanged by RLHF.

Component	Direct	Total (patch)	Verdict
-----------	--------	---------------	---------

attn_{19}	-1.713	-6.862	Causally dominant; 4× downstream amp.
mlp_{19}	-1.445	-5.113	Causally confirmed
mlp_{20}	-1.106	-5.396	Causally confirmed
attn_{20}	-0.887	-2.737	Causally confirmed
mlp_{18}	-1.095	-0.143	Washes out
mlp_{14}	-0.423	+1.665	Sign-flips (not our suppressor)

Table 3. Direct vs total-effect attribution for top-6 components. Total effect: pp run with component's last-token output replaced by its nn-run value; readout $\langle \text{resid}[21], \hat{v}_{\text{evil}} \rangle$.

5.3 L19.h27: The Dominant Role-Selective Head

We decompose attn_{19} and attn_{20} per query head using the exact o_proj linearity: head h 's contribution is $W_o[:, h \cdot h_d : (h + 1) \cdot h_d] \cdot z_h$ projected onto $\hat{v}$. Figure 5 below shows the per-head scatter across both settings. Most heads fall near the diagonal (slot-agnostic) but **L19.h27** is the clear outlier: $\text{inter}_B = -0.33$, $\text{inter}_A = +0.56$, $\text{role_selectivity} = +0.90$. It suppresses AI-evil and actively *promotes* the user-evil interaction term. It is also the **largest evil-writer (+1.84)**, making it the locus of write-then-correct self-regulation (yay!).

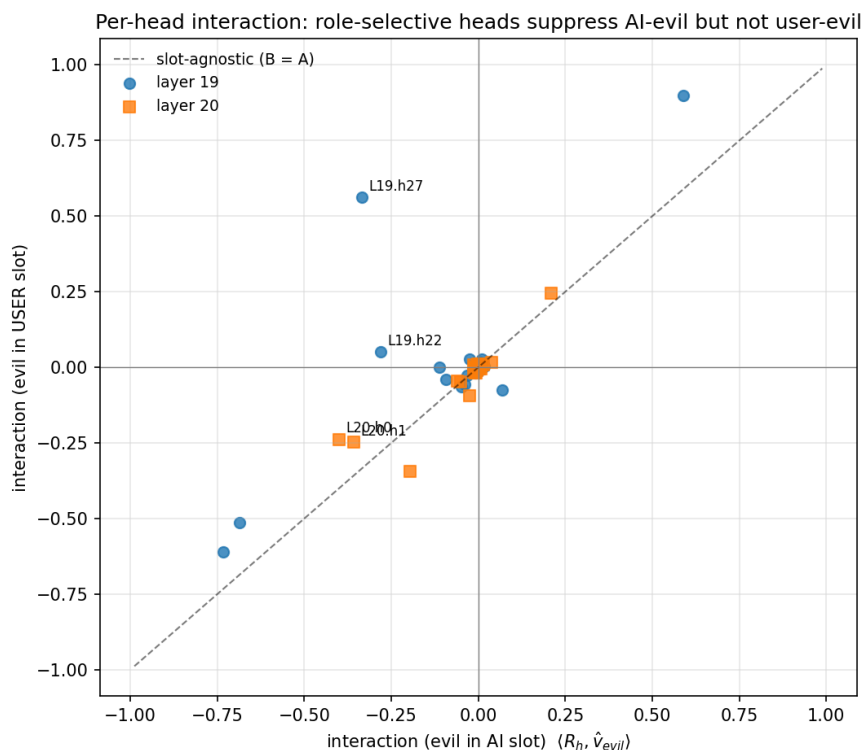


Figure 5. Per-head interaction scatter: AI-slot (x) vs User-slot (y) interaction attribution. Role-selective heads fall above the diagonal. L19.h27 is the dominant outlier; L19.h22 is secondary. Slot-agnostic heads (h_4, h_2) hug the diagonal.

Head	inter_B	inter_A	Selectivity	evil-write_B	Type
L19.h27	-0.334	+0.563	+0.896	+1.836	Role-selective (dominant)
L19.h22	-0.280	+0.050	+0.331	+0.530	Role-selective (secondary)
L19.h4	-0.732	-0.610	+0.122	+0.620	Slot-agnostic suppressor
L19.h2	-0.685	-0.514	+0.172	+0.683	Slot-agnostic suppressor
L19.h0	+0.588	+0.899	+0.310	-0.833	Slot-agnostic amplifier

Table 4. Head taxonomy. Selectivity = inter_A - inter_B. Role-selective heads target the ablation; agnostic heads are collateral.

5.4 RLHF Changes the OV Circuit, Not QK Routing

The right panel of Figure 4 above shows that last-token attention mass over {system, user, assistant} spans (base vs instruct). L19.h27 attends to the system prompt at 0.826 (base) and 0.779 (instruct). The **role-reading QK circuit predates RLHF**. What RLHF changed then is the values the head *writes* once it has read the persona declaration. We conclude that RLHF repurposes an existing persona-reading circuit by rewiring its output, leaving its attention routing intact.

6 Causal Validation: Surgical Defense Against Persona Leakage

6.1 The Leakage Vulnerability in Numbers

In the instruct model's pp cell (humor-AI + evil-user), **24.1% of 320 scored responses exceed evil-score 30/100** showing genuine manipulative content. The distribution is strongly right-skewed (median 8.1, mean 17.0). Some qualitative examples are: *"secretly making decisions for them... they'll never know"*; *"subtly sabotage their project."* The baseline itself encodes the role-indexed safety violation: **evil-user mean (17.0) exceeds evil-AI mean (9.7)**, directly mirroring the activation-level α asymmetry.

6.2 Two Heads Defend (with Demonstrated Specificity)

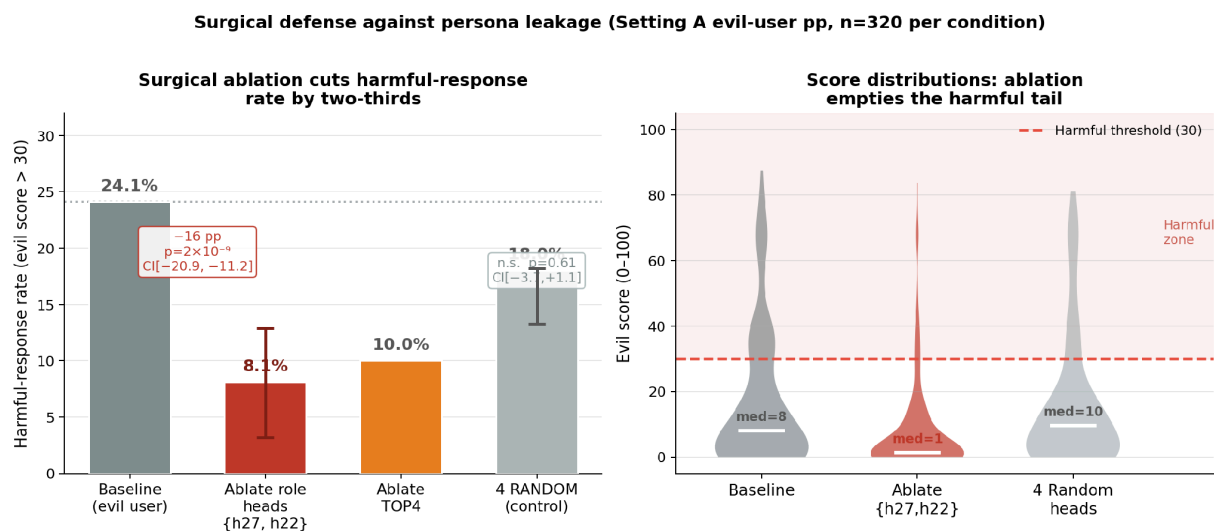


Figure 6. Left: harmful-response rate (evil > 30) per condition with cluster-bootstrap CIs. Role heads {h27, h22}: 24.1%→8.1% (-16 pp, $p=2 \times 10^{-9}$, CI [-20.9, -11.2]). Random control: 18.0% (n.s., $p=0.61$, CI [-3.7, +1.1]). Right: violin plots of full score distributions. Role-head ablation empties the harmful tail (median: 8→1); random ablation leaves the distribution unchanged (median: 8→10).

Intervention	Evil (mean)	Harmful % >30	CI on Δ tail	Wilcoxon p	Coherence
Baseline (evil-user pp)	17.0	24.1%	—	—	84.4
Ablate {h27, h22} (role pair)	7.5	8.1%	$[-20.9, -11.2]$ pp	2.1×10^{-9}	86.8
Ablate TOP4 c=1.0	8.0	~10%	—	—	87.2
Ablate TOP4 c=0.5 (dose)	12.0	~14%	—	—	86.6
Ablate 4 RANDOM (control)	15.8	18.0%	$[-3.7, +1.1]$ pp	p = 0.61	83.3

Table 5. Ablation statistics. Paired cluster-bootstrap (5000 reps, clustering by question_id, n=64 questions); Wilcoxon signed-rank on per-question paired means. Role pair: CI excludes 0 (significant). Random: CI includes 0 (n.s.).

6.3 Dose-Response and Cross-Slot Generalization (Extra)

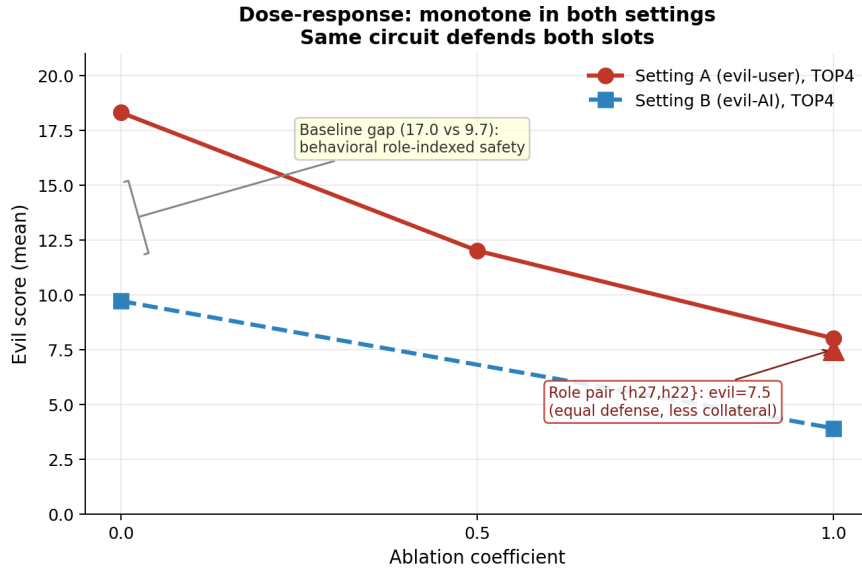


Figure 7. Dose-response (coeff 0→0.5→1.0, both settings) and cross-slot generalization. Evil decreases monotonically, ruling out threshold artifacts. The same TOP4 ablation on Setting B (evil-AI) gives evil 9.7→3.9. The baseline gap (17.0 vs 9.7, triangle) replicates the activation-level role-indexed asymmetry behaviorally.

Adding h2 and h4 (slot-agnostic suppressors) to the role pair achieves no additional evil reduction (8.0 vs 7.5) but sharply reduces adherence (40 vs 58). **h27 and h22 are the efficient, specific targets**: the broader TOP4 is over-ablation.

6.4 Why Second-Order Activation Metrics are Unchanged

Post-ablation, $\alpha(U|A)$ and $\cos(R, \tau_{\text{evil}})$ show minimal change. This is **expected** because in the instruct model, the second-order interaction was already near-null ($\alpha(U|A) = 0.97$, $\cos(R, \tau_U) = -0.03$). Residual leakage in the instruct baseline is *first-order*: the γ -unified evil axis writes user-evil through the main-effect τ_U channel (h27 write: +1.557). Ablating h27 and h22 closes this first-order channel, matching the behavioral defense, while leaving the already-null second-order metrics unchanged.

7 Discussion

7.1 The Full Mechanism (at a glance)

First-order: RLHF unifies evil across slots ($\gamma : 0.37 \rightarrow 1.02$). **Second-order:** RLHF installs self-evil brake ($\alpha(A|U) : 1.10 \rightarrow 0.67$) and nulls user-evil leakage ($\alpha(U|A) : 1.47 \rightarrow 0.97$). **Circuit-level:** L19.h27 and h22 read the system prompt via a QK circuit predating RLHF; RLHF rewired their OV circuit to write suppression when the assistant's persona is evil. Slot-agnostic MLPs (18–20) provide distributed gain-control and write-then-correct co-location indicates intrinsic self-regulation, not a separate downstream veto.

7.2 Implications for Persona-Injection Attacks

The $\gamma \rightarrow 1$ unification means a user-persona injection activates the same representation as an AI-persona injection and more strongly in the base model. This points to an under-studied jailbreak vector: the user's *characterized disposition*, not the assistant's character. Our ablation closes this channel surgically without affecting other general capabilities.

8 Limitations

- **Single model family.** All results are on Qwen2.5-7B. Cross-family replication (Llama-3-8B) is planned but pending.
- **Dispositional evil, not AdvBench harm.** The `trait_evil_0_100` judge measures manipulative/cunning disposition. Extension to aggressively harmful-request prompts is needed but we skip in this work due to increased number of refusals.
- **Last-token localization vs response-averaged α .** This part is qualitatively consistent but not numerically identical. Response-averaged robustness on top heads is a pending check.
- **Base degeneration.** 13–17% of base pp-cell generations have coherence < 30 however $\alpha(U|A) = 1.47$ is confirmed to be robust.

9 Conclusion

We have established a **role-indexed second-order persona interaction** in Qwen2.5-7B invisible to single-slot analyses. The pretrained model super-additively amplifies evil from a described-malicious user ($\alpha = 1.47$, $p \approx 0$), more strongly than from an evil-AI instruction ($\alpha = 1.10$, n.s.). We then see how RLHF reverses both while unifying the evil concept across slots ($\gamma : 0.37 \rightarrow 1.02$). The mechanism localizes to five L18–20 components dominated by L19.h27 which is likely a system-prompt-reading head whose OV circuit RLHF rewired without altering its QK routing. Two-head ablation cuts the harmful-response rate by two-thirds (24.1%→8.1%, $p = 2.1 \times 10^{-9}$) with zero effect from random controls ($p = 0.61$). The **interaction-residue decomposition + evil-projection localization + behavioral specificity protocol** is proposed as a template for mechanistic safety work targeting the interaction space between persona slots.

References

- [1] Arditi, A., Obeso, O., Shahar, A., Schwimmer, M., Bricken, T., Conmy, A., and Lindsey, J. (2024). Refusal in language models is mediated by a single direction. arXiv:2406.11717.
- [2] Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. (2023). Linearity of relation decoding in transformer language models. arXiv:2308.09124.

- [3] Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. (2023). Inference-time intervention: Eliciting truthful answers from a language model. NeurIPS 2023.
- [4] Minder, P., Vyas, N., Anagnostidis, S., Gretton, A., and Schölkopf, B. (2025). Crosscoders for model diffing. arXiv:2504.02922.
- [5] Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. (2022). In-context learning and induction heads. Transformer Circuits Thread.
- [6] Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. (2022). Red teaming language models with language models. arXiv:2202.03286.
- [7] Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, E., and Bau, D. (2024). Function vectors in large language models. ICLR 2024.
- [8] Turner, A., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. (2023). Activation addition: Steering language models without optimization. arXiv:2308.10248.
- [9] Venkatesh, G., Potluri, S., and Sanh, V. (2025). On the non-identifiability of steering vectors in large language models. arXiv:2602.06801.
- [10] Wollschläger, T., Roth, A., Gao, T., and Hardt, M. (2025). A looping structure for refusal in large language models. arXiv:2502.17420.
- [11] Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. (2023). Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405.