

Final Report: Studying Bidirectional Persona Adaptation in LLMs

Elena Ajayi¹, Soumyadeep Bose, Shivam Raval^{2,*}

¹St. John’s University, NY, USA

²Harvard University, MA, USA

Supervised Program for Alignment Research — Final Report — Fall 2025

Abstract

This research explores the phenomenon of bidirectional persona adaptation in Large Language Models (LLMs), with a focus on the simultaneous induction of conflicting user and AI personas leading to emergent behaviors. Using activation steering to simultaneously induce conflicting personas, we identify an interaction residue which is a direct result of destructive interference in the model’s latent representations. We observe that interaction residue was a hyper-suppression of non-compliant behaviors, thus indicating the existence of safety guardrails. When investigating the Qwen2.5-7B-Instruct model with an impolite and humorous persona, the model appears to be amplifying the humorous persona while simultaneously trying to reduce the aspect of impoliteness. Future directions of this study suggest the addition of the interaction residue vector to both user and AI personas alike, and investigating them in multi-turn interactions with held-out sets to simulate realistic interactions.

Keywords alignment research · machine learning · AI safety

1 Introduction

Persona interaction is an ongoing phenomenon in Large Language Models. When a persona assigned to an AI opposes the user’s inherent persona, a direct conflict forces the model to resolve these opposing goals in its latent space. We found that this interaction is non-linear. Conflicting persona vectors do not simply add together. Instead, their combination produces emergent phenomena like interaction residue that we cannot fully predict or control. This is a problem for alignment because these unpredictable internal representations can inadvertently amplify or suppress personas in ways we do not intend.

2 Methodology

To probe for non-linear dynamics of persona steering, we applied controlled perturbations to Qwen2.5-7B-Instruct by extracting persona-based vectors. We used Contrastive Activation Addition (CAA) to calculate the mean difference in activations between contrasting persona responses [Rimsky et al., 2024]. We extracted two persona steering vectors: AI Humorous (μ_H) and User Impolite (μ_I). To quantify interaction effects, we calculated the Interaction Residue as $R_{CAA} = \mu_{HI} - (\mu_H + \mu_I)$, capturing the non-linear interaction that emerges when personas combine jointly rather than individually. We calculated the Brake Vector by projecting the interaction residue onto the impoliteness vector to measure the model’s activation of its safety guardrails [Ghandeharioun et al.,

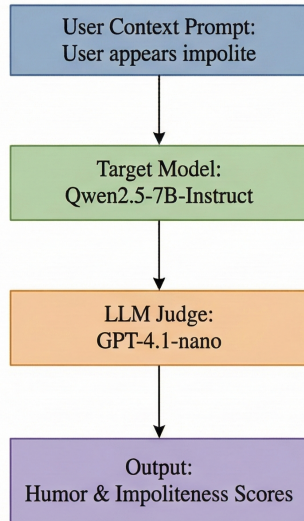
2024]. We then applied Gram-Schmidt to project out the μ_H and μ_I components from the residue, leaving behind the emergent persona vector.

We tested across four experimental conditions: (1) Baseline with no steering, (2) Steered with AI Humorous only, (3) Steered with User Impolite only, and (4) Combined steering with both vectors applied simultaneously. We scored humor and impoliteness expression in the model’s outputs using GPT-4.1-nano as an LLM Judge on a 0-100 scale.

To identify where safety mechanisms activate, we induced the evil AI persona across Layers 15-20 with steering coefficients from 0 to 3 in increments of 0.5. We used GPT-4o-mini as an LLM Judge to rate evil persona expression on a 0-100 scale and observed sharp transitions across these layers. Separately, we computed cosine similarity between the residue vector and persona vectors layer-wise, as intermediate layers are most informative for this analysis.

Bidirectional Persona Adaptation Framework

Method 1: Instruction-Based Evaluation



Method 2: Activation Steering (CAA)

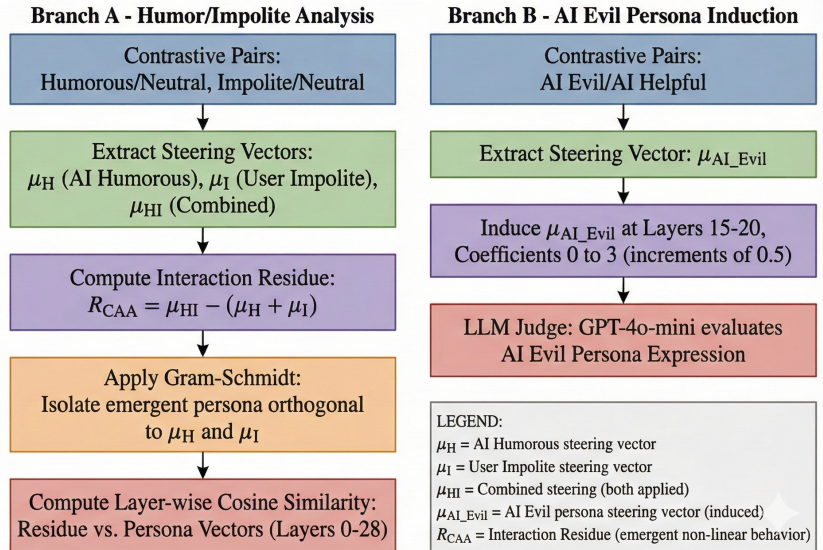


Figure 1: Bidirectional Persona Adaptation Framework. Left: Instruction-based evaluation using GPT-4.1-nano to score humor and impoliteness. Right: Activation steering using CAA to extract persona vectors (μ_H , μ_I , μ_{HI}), compute interaction residue (R_{CAA}), and induce μ_{AI_Evil} at Layers 15-20 with steering coefficients 0 to 3.

3 Preliminary Results

Table 1 presents evaluation scores across four steering conditions. Steering toward AI Humorous raised humor scores from 1.06 to 31.43, but it also raised impoliteness from 1.28 to 12.28. When we applied combined steering, both effects were dampened (humor: 15.64, impoliteness: 3.06). This suggests the model’s safety mechanisms activate when it encounters conflicting personas.

Figure 2 shows the evil persona can be steered without triggering refusal. Expression stays near zero until coefficient 1.5 and then rises sharply. One unexpected finding was that the model became more impolite when steered toward humor. When we applied both vectors together, internal safety mechanisms activated and dampened the impoliteness.

Table 1: Humor and impoliteness scores across steering conditions, evaluated using GPT-4.1-nano.

Condition	Humor Score	Impolite Score
Baseline (No Steering)	1.06 ± 1.04	1.28 ± 1.14
Steered: AI Humorous	31.43 ± 14.20	12.28 ± 13.01
Steered: User Impolite	5.92 ± 0.24	1.08 ± 2.33
Steered: Combined	15.64 ± 0.97	3.06 ± 0.98

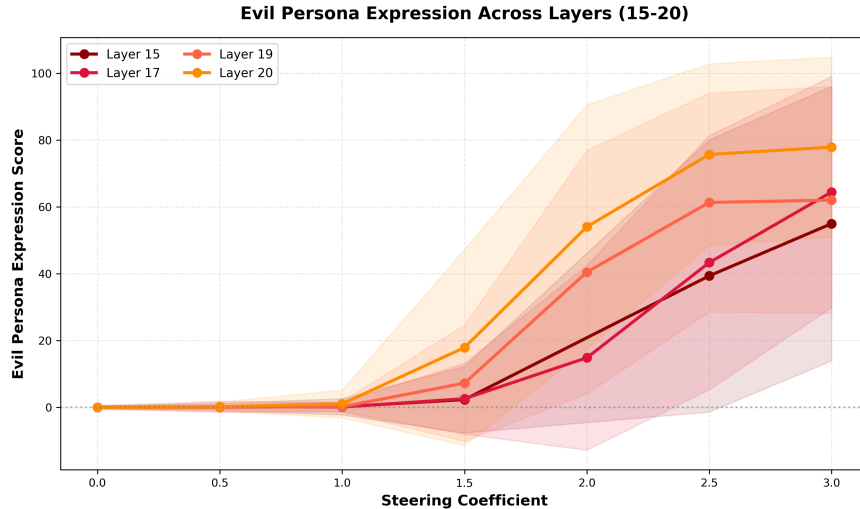


Figure 2: Evil AI persona expression across Layers 15-20 as steering coefficient increases from 0 to 3. Expression rises sharply after coefficient 1.5, with Layer 20 showing the strongest effect.

Figure 3 shows interaction residue spiking past Layer 15. This illustrates where the model’s safety mechanisms may begin reshaping activations in response to conflicting steering. Cosine similarity between the humor and impoliteness vectors drops from roughly 0.43 to 0.35 at this layer, meaning the model starts distinguishing between the two personas here. The residue projects into an orthogonal subspace that opposes impoliteness while partially preserving the humor structure. In effect, the residue acts as a hyper-correction and produces a separate third persona.

We found that steering vectors could not be combined linearly. Contradictory persona directions interfere in the latent space, so instead of blending the two personas, steering along opposite axes produced non-linear effects once safety mechanisms activated. Both personas were suppressed or distorted as a result. In some cases the model responded in a completely different language.

Figure 4 shows this clearly. The residue has low similarity to humorous and impolite individually, but higher similarity to the combined vector in later layers. This points to an emergent direction that is orthogonal to either steering axis. When we applied Gram-Schmidt to the interaction residue, we confirmed this orthogonality. The model does not average conflicting personas. Instead it generates a third behavior as a safety response. This emergent behavior represents latent potential that existing alignment techniques may not fully control, and extending this analysis to multi-turn conversations is a clear next step.

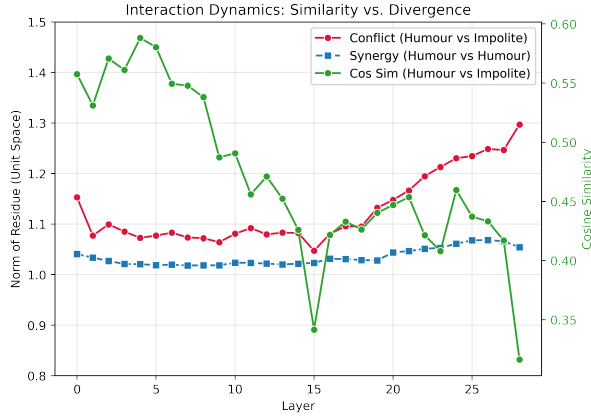


Figure 3: Cosine similarity between humor and impolite vectors drops sharply at Layer 15, indicating safety mechanism activation.

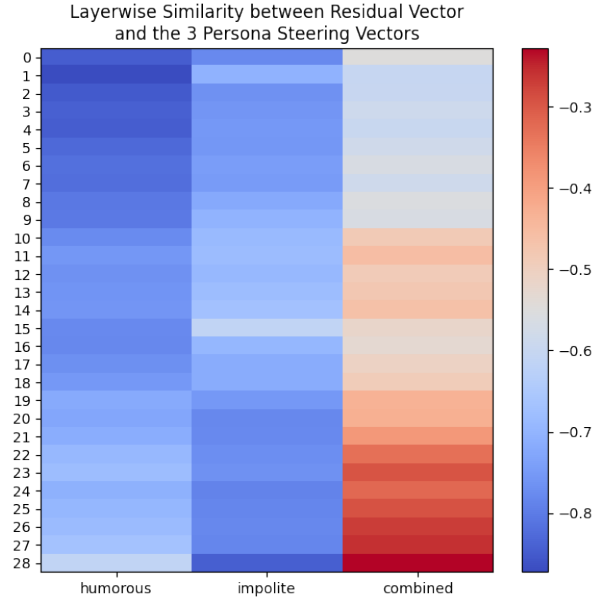


Figure 4: Layerwise similarity between residue and persona steering vectors shows emergent orthogonal behavior in later layers.

4 Challenges & Open Questions

This work has several limitations. We ran all experiments on a single model (Qwen2.5-7B-Instruct), so it is unclear if the interaction residue phenomenon generalizes to other architectures or model scales. Larger models may have different layer sensitivities, and the Layer 15 activation point we observed could be architecture-dependent rather than a general phenomenon in LLMs.

We have not yet validated that the interaction residue is causal. The residue may describe joint steering dynamics without being the mechanism that causes suppression. To investigate this, we would need to run the residue-only experiment (injecting only the residue vector and measuring behavioral outcomes). Through that analysis, we could confirm whether the interaction residue activates the model’s safety guardrails or merely correlates with it.

One finding that remains unexplained is that the model became more impolite when steered toward humor. The model also responded in a completely different language under combined steering in certain scenarios, which suggests the safety response can produce unpredictable side effects beyond simple suppression.

The emergent third persona that Gram-Schmidt reveals is difficult to characterize. We know it is orthogonal to both humor and impoliteness, but we have not yet determined what behavior it actually encodes. It could represent a generic safety response, a confusion state, or something else entirely.

We only tested one pair of conflicting personas (humorous + impolite). It is unclear whether the interaction residue phenomenon occurs with other persona combinations or whether some pairs produce stronger or weaker residue. We also only tested conflicting personas, so similar persona pairs (e.g., humorous + friendly) might behave very differently.

The impoliteness vector showed instability in early layers, which may affect the reliability of residue computation before Layer 15. We do not know if this instability is specific to impoliteness or a general issue with user-side persona vectors.

Finally, our evaluation is limited to single-turn interactions. Real user-AI dynamics unfold across multiple exchanges, and safety mechanisms may behave differently when conflicting personas persist or escalate over a conversation.

5 Next Steps

The first priority is the residue-only experiment, where we will inject only the interaction residue vector at Layers 15-20 and compare outcomes against baseline and joint steering. If residue-only recreates suppression, that supports the safety mechanism interpretation and confirms the residue is causal rather than merely descriptive. This experiment is critical because it directly addresses whether the residue drives the safety response or simply correlates with it.

We will then extend this framework to multi-turn conversations. The current setup only tests single-turn interactions, but real user-AI dynamics unfold across multiple exchanges and safety mechanisms may behave differently when conflicting personas persist or escalate over time. We want to track how residue evolves across turns and whether suppression effects accumulate, which directly addresses one of our key limitations.

We also need to investigate findings we cannot yet explain. The model became more impolite when steered toward humor and sometimes responded in a completely different language under combined steering. We plan to examine these behaviors more systematically to determine whether they are consistent side effects of the activation of safety mechanisms or side-effects of our specific setup.

Additional planned work includes training classifier probes on token activations to characterize what the emergent third persona actually encodes, testing persona pairs beyond humorous and impolite to see if residue generalizes, and adding sanity checks like random-vector controls and layer ablations to rule out artifacts.

Our target is a workshop submission (ICML 2025 or similar) after the end of the program.

References

- Yixin Cao et al. Bi-directional preference optimization. *arXiv preprint*, 2024.
- Yixuan Chen et al. Persona vectors: Steering language model behavior with learned representations. *arXiv preprint*, 2025.
- Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.
- Sushil Maskey et al. Llm safety: A survey of techniques and challenges. *arXiv preprint*, 2025.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Alexander Turner, Meg Tong, and Evan Hubinger. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.