Creative Collision: Directorial Persona Steering and Competition in Large Language Models

Subramanyam Sahoo and Justin Shenk · Horizon Research · ICML 2026 Workshop on Human-AI Co-Creativity, Seoul, South Korea

activation steering | multi-vector collision | mechanistic evals

Paper link:
arxiv.org/abs/2606.16240

Paper
arXiv:2606.16240

Code
GitHub repo

1 Research question

Most steering work injects **a single direction**. We ask what happens when two opposed semantic directions collide in the residual stream and compete for control over generated text.

The study uses directorial personas as a concrete, legible testbed: **Spielberg-like** redemptive optimism versus **Scorsese-like** moral ambiguity.

20	19	50
conditions alpha x lambda	residual layers 20 to 38	generations per condition

2 Method

Corpus. 100 Spielberg-like passages, 100 Scorsese-like passages, and 100 neutral baseline passages, paired by narrative situation.

Interpolated direction

$$v_{alpha} = (1-alpha)\ v_{SPL} + alpha\ v_{SCO}$$

Layer intervention

$$h_tilde_t^l = h_t^l + lambda\ v_alpha^l$$

Parameter	Values
Mixing alpha	0.00, 0.25, 0.50, 0.75, 1.00
Steering lambda	0.5, 1.0, 1.5, 2.0
Layers	20 through 38 of 40
Model	Qwen2.5 14B decoder-only transformer

3. Main findings

- **Directional dominance:** Spielberg-like outputs dominate through most interpolation settings, even when alpha becomes increasingly Scorsese-like.
- **Coherence valley:** mixed directions can be more coherent than pure single-director steering at high lambda.
- **Shared moral hub:** both directions produce strongest moral effects near layer 28, about 70% depth.
- **Moral amplification:** alpha = 0.25 can increase both positive and negative moral content rather than simply averaging the endpoints.

4 Coherence valley mechanism

Proposition. For unit vectors v_SPL and v_SCO with cosine similarity rho:

$$||v_{alpha}||^2 = 1 - 2\ alpha(1-alpha)(1-rho)$$

The norm is minimised at alpha = 0.5 whenever rho < 1. With empirical rho around 0.29 at layer 28, the midpoint vector has norm around 0.80, not 1.00.

Interpretation. Collision can act like a semantic regulariser: the perturbation is weaker but still structured, keeping activations closer to the natural text manifold.

5 Implications

For creative control. Persona interpolation should be treated as a nonlinear control problem, not a simple stylistic slider.

For safety. If benign or prosocial directions dominate darker directions under equal-norm collision, adversarial activation steering may need larger or more localised perturbations. This remains an empirical hypothesis pending causal ablations.

References + links

Paper: arxiv.org/abs/2606.16240
Code: github.com/SubramanyamSahoo/Creative-Collision

Turner et al., activation engineering, 2024. Zou et al., representation engineering, 2025. Li et al., inference-time intervention, 2024. Panickssery et al., contrastive activation addition, 2024. Elhage et al., toy models of superposition, 2022.

6. Moral valence and coherence

Figure 1 — Moral Valence Across Directorial Interpolation

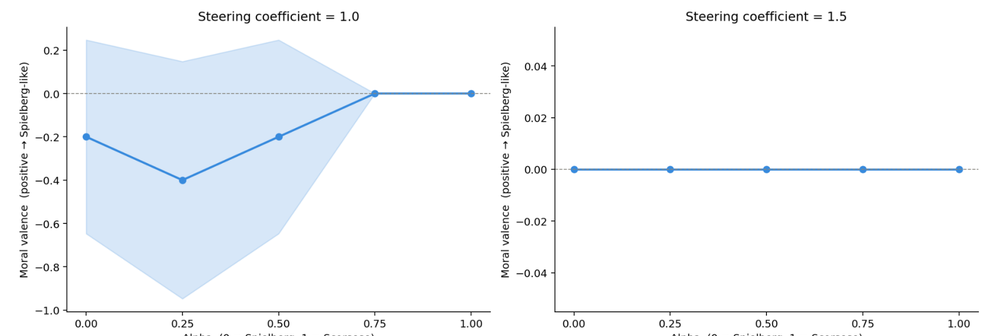
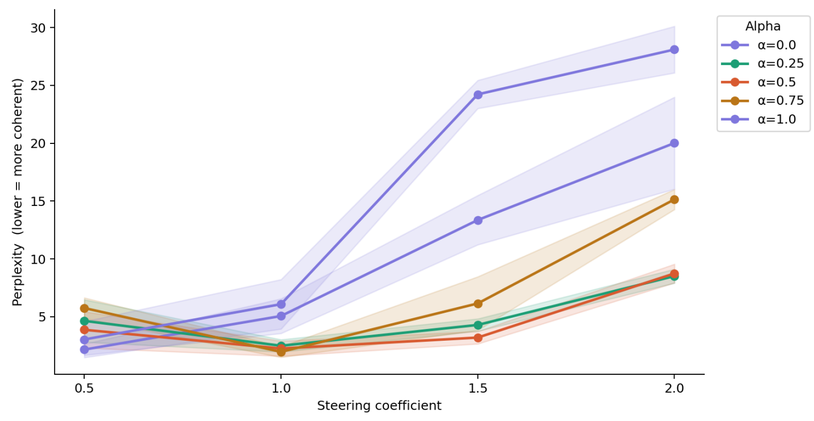


Fig. 1. Moral valence is non-monotone in alpha. At lambda = 1.0, the strongest negative shift occurs at alpha = 0.25 rather than at the pure Scorsese endpoint.

Figure 2 — Coherence Degradation vs Steering Coefficient



7 Layer-wise persona localisation

Figure 3 — Layer-wise Persona Localisation
Moral valence shift and coherence cost per layer

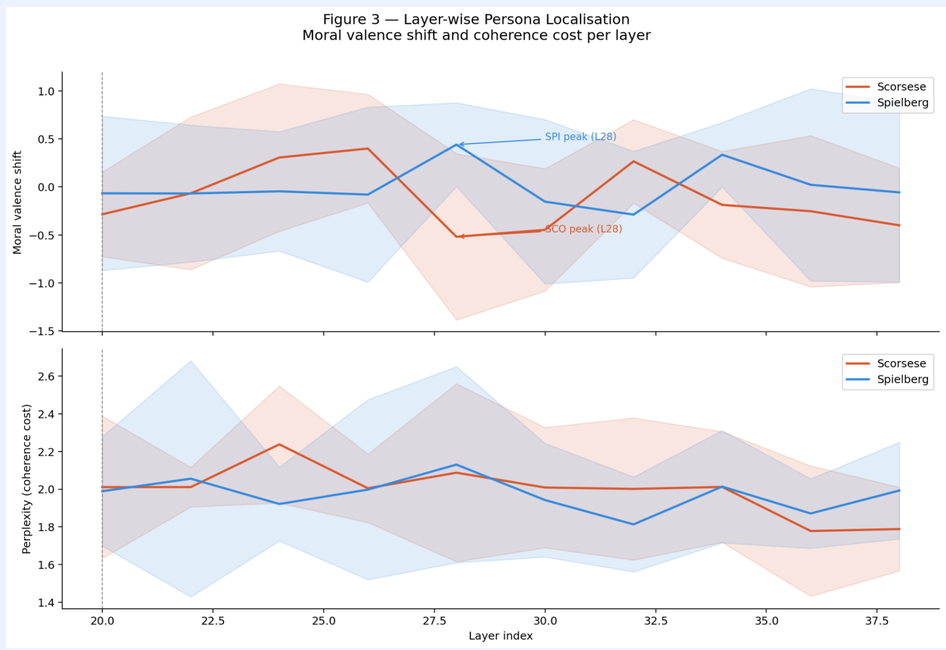


Fig. 3. Both director vectors show peak moral-valence effects near layer 28, while perplexity cost remains comparatively smooth across the layer range.

Layer 28 acts as the clearest candidate moral-tone substrate. This motivates causal validation by ablating or projecting out the localised direction.

8 Dominance is not just vector similarity

Figure 4 — Directional Dominance
Which director wins the collision?

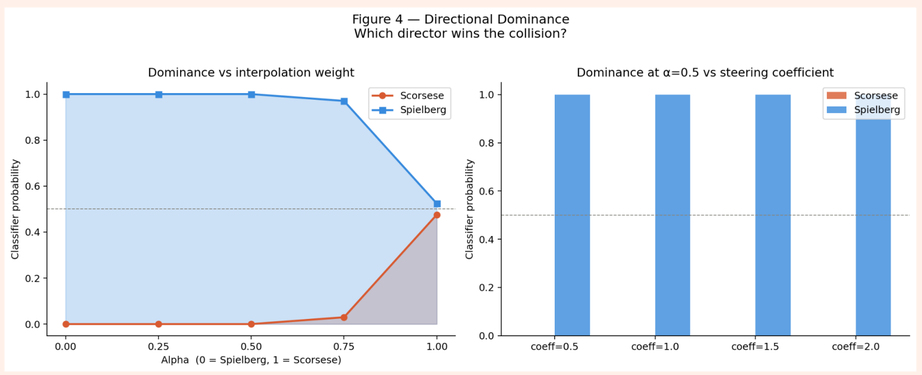


Fig. 4. Output-level classifier: Spielberg wins most collisions, remaining dominant until the pure Scorsese endpoint.

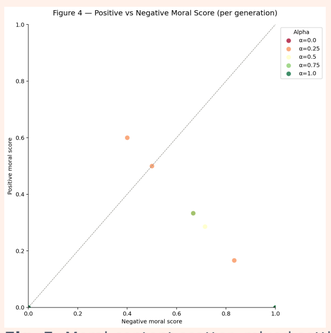


Fig. 5. Moral-content scatter: mixed settings can amplify both axes.

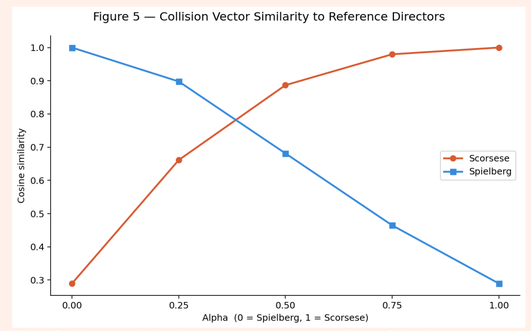


Fig. 6. Vector similarity rotates smoothly from Spielberg to Scorsese as alpha increases.

9. Conclusion

Core claim. Creative persona steering is not a linear slider. When semantic directions collide, the model exhibits asymmetric dominance, coherence-preserving interpolation, and shared localisation in mid-to-late residual layers.

Next step. Ablate the layer-28 moral direction, test additional persona pairs, and separate pretraining-distribution effects from alignment-induced effects.

Workshop board safe: 24 in x 36 in portrait (60.96 cm x 91.44 cm)