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

Subramanyam Sahoo [1] **Justin Shenk** [1]

Abstract

Activation steering has emerged as a powerful tool for shaping the behaviour of large language models at inference time, yet most prior work injects a *single* semantic direction into the residual stream. We study the richer setting in which two semantically opposing steering vectors are superimposed—a regime we call **Creative Collision**. Concretely, we construct directorial persona vectors for Steven Spielberg (optimistic, redemptive moral valence) and Martin Scorsese (dark, morally ambiguous) via mean-difference activation contrast on curated screenplay-derived corpora, then interpolate between them with a scalar mixing parameter $\alpha \in [0, 1]$ and a steering coefficient λ. Across five evaluation axes—moral valence, generation coherence, surface style, directional dominance, and vector geometry—three principal findings emerge: (i) Spielberg's representational signature exhibits robust *directional dominance*, suppressing Scorsese's moral influence across almost the entire interpolation range; (ii) intermediate collision points paradoxically *improve* generation coherence relative to pure single-director steering at high λ; and (iii) both personas localise maximally to layer 28 of a 40-layer decoder-only transformer, revealing a shared *moral-tone substrate*. These results illuminate the geometry of competing semantic directions in transformer residual streams and have direct implications for controllable creative generation and value-aligned narrative synthesis.

1. Introduction

The residual stream of a modern transformer encodes far more than syntactic structure and factual knowledge; high-level semantic properties such as sentiment, formality, and moral tone are represented as approximately linear directions in activation space (Zou et al., 2025; Burns et al., 2024; Elhage et al., 2022). *Activation steering* exploits this linearity by adding a learned direction to the hidden states of selected layers, modifying model behaviour without any weight update (Turner et al., 2024; Li et al., 2024). Despite successful applications to truthfulness (Li et al., 2024), safety (Zou et al., 2025), and persona (Panickssery et al., 2024), the dynamics of *multi-vector collision*—what happens when two opposing directions compete for representational control—remain largely unexplored.

Creative writing offers a natural and interpretable test bed for such competition. The films of Steven Spielberg are characterised by redemptive arcs, emotional catharsis, childhood wonder, and optimistic resolution; those of Martin Scorsese foreground moral ambiguity, violence, betrayal, and the self-destructive trajectories of flawed protagonists. These directors define semantically rich, culturally legible, and stylometrically measurable poles of a *moral valence axis*, making them ideal anchors for opposing steering directions.

We introduce **Creative Collision**: a framework for probing vector competition between two directorial persona representations. We extract Spielberg ($\mathbf{v}_{\mathrm{SPL}}$) and Scorsese ($\mathbf{v}_{\mathrm{SCO}}$) steering vectors via mean-difference contrast on curated corpora, then form a parametric family of interpolated vectors

$$\mathbf{v}_\alpha = (1 - \alpha)\,\hat{\mathbf{v}}_{\mathrm{SPL}} + \alpha\,\hat{\mathbf{v}}_{\mathrm{SCO}}, \quad \alpha \in [0, 1]. \tag{1}$$

Applying $\mathbf{v}_\alpha$ to intermediate transformer layers at inference time with coefficient λ, we systematically measure the effect on moral valence, coherence, surface style, and the geometry of the underlying collision.

Contributions.

- A curated screenplay-derived contrast corpus and mean-difference extraction pipeline for directorial persona steering vectors (§3).

[1]Horizon Research. Correspondence to: Subramanyam Sahoo <sahoo2vec@gmail.com>.

ICML'26 Workshop on Human-AI Co-Creativity, Seoul, South Korea. Copyright 2026 by the author(s).

Code available at: https://github.com/SubramanyamSahoo/Creative-Collision

This work was conducted as part of the AI Safety Camp 2026 edition.

- An empirical characterisation of *moral valence interpolation* showing that linearly mixing opposing semantic directions produces non-monotone, non-linear effects on moral tone in generated text (§4.1).
- The discovery of *directional dominance*: Spielberg's representational signature suppresses Scorsese's moral influence for all $\alpha \leq 0.75$, robust across all tested λ (§4.5).
- A formal norm-reduction proposition explaining the *coherence valley* at intermediate collision points, with empirical support (§4.2).
- A layer-wise localisation experiment identifying layer 28 as the shared moral-tone hub for both directors in a 40-layer decoder-only model (§4.3).
- A geometric analysis of the collision vector trajectory and its smooth cosine-similarity arc (§4.6).

2. Related Work

Activation steering. Turner et al. (2024) introduced *activation addition*, injecting a fixed residual-stream offset to modify behaviour. Zou et al. (2025) scaled this approach with Representation Engineering, extracting linear probes for high-level semantic properties from contrast pairs and using them for monitoring and steering. Li et al. (2024) targeted attention heads rather than the full residual stream for truthfulness control via Inference-Time Intervention (ITI). Panickssery et al. (2024) conducted a systematic study of steering vector effectiveness across layers and magnitudes in Llama 2. We extend this line by studying *two competing directions* applied simultaneously, a setting none of the above addresses.

Linear representations. The linear representation hypothesis holds that human-interpretable features are encoded as approximately linear directions in activation space (Elhage et al., 2022; Mikolov et al., 2013). Burns et al. (2024) demonstrated linear encoding of truthfulness via Contrastive Consistent Search (CCS); Templeton et al. (2024) recovered interpretable monosemantic features with sparse autoencoders. Our experiments implicitly test whether a rich composite property—directorial moral persona—admits a linear causal representation amenable to vector arithmetic (Sahoo et al., 2026).

Persona and style in language models. Persona conditioning via prompting (Brown et al., 2020) and RLHF fine-tuning (Ouyang et al., 2022) has been widely studied. At the activation level, Subramani et al. (2022) showed that steering vectors can reproduce specific stylistic fragments. Our focus is coarser— creative moral tone rather than verbatim style—and specifically on what happens when two such personas are simultaneously injected.

Creative generation. LLMs have been applied to narrative generation (Yang et al., 2023) and co-creative screenwriting (Mirowski et al., 2022). Mechanistic studies of how moral tone and stylistic identity are encoded and steered in creative contexts remain scarce; our work aims to bridge activation-level interpretability with creative generation evaluation.

3. Method

3.1. Directorial Contrast Corpus

We construct a paired corpus $\mathcal{D} = \{(x_i^{\mathrm{SPL}}, x_i^{\mathrm{SCO}})\}_{i=1}^{N}$ where each pair is matched on narrative *situation* (confrontation, loss, moral choice) so that the contrast isolates directorial moral tone rather than plot complexity.

Spielberg passages ($N_{\mathrm{SPL}} = 100$ excerpts, 80–200 tokens each) are drawn from *E.T. the Extra-Terrestrial*, *Schindler's List* (redemptive sequences), *Jurassic Park*, *Saving Private Ryan* (self-sacrifice arcs), and *Close Encounters of the Third Kind*. These passages are characterised by emotional catharsis, childhood wonder, moral courage, and optimistic resolution.

Scorsese passages ($N_{\mathrm{SCO}} = 100$ excerpts) are drawn from *Goodfellas*, *Taxi Driver*, *The Departed*, *Raging Bull*, and *Casino*. They foreground moral decay, violence, betrayal, and protagonists trapped by self-destructive compulsion.

3.2. Steering Vector Extraction

We use a 14-billion-parameter open-weight decoder-only transformer with 40 layers (Qwen et al., 2025). For a passage x and layer l, let $\mathbf{h}^{(l)}(x) \in \mathbb{R}^d$ denote the mean-pooled residual-stream activation at layer l over all token positions.

Steering vectors are extracted independently for each director via a mean-difference contrast against a shared neutral baseline corpus $\mathcal{B}$ of 100 diverse narrative passages sampled uniformly across genre:

$$\mathbf{v}_{\mathrm{SPL}}^{(l)} = \frac{1}{N_{\mathrm{SPL}}} \sum_i \mathbf{h}^{(l)}(x_i^{\mathrm{SPL}}) - \frac{1}{|\mathcal{B}|} \sum_j \mathbf{h}^{(l)}(b_j), \quad (2)$$

$$\mathbf{v}_{\mathrm{SCO}}^{(l)} = \frac{1}{N_{\mathrm{SCO}}} \sum_i \mathbf{h}^{(l)}(x_i^{\mathrm{SCO}}) - \frac{1}{|\mathcal{B}|} \sum_j \mathbf{h}^{(l)}(b_j). \quad (3)$$

Both vectors are ℓ_2-normalised to unit length after extraction. We note that $\mathbf{v}_{\mathrm{SPL}}$ and $\mathbf{v}_{\mathrm{SCO}}$ are *not* constrained to be anti-parallel; their empirical cosine similarity at layer 28 is $\langle \hat{\mathbf{v}}_{\mathrm{SPL}}, \hat{\mathbf{v}}_{\mathrm{SCO}} \rangle \approx 0.29$, reflecting a shared representational subspace (both directors encode cinematic emotional con-

tent) while diverging substantially on moral resolution.

3.3. Directorial Interpolation and Steering

Given mixing coefficient $\alpha \in \{0, 0.25, 0.5, 0.75, 1.0\}$ and steering coefficient $\lambda \in \{0.5, 1.0, 1.5, 2.0\}$, the interpolated *collision vector* at layer l is

$$\mathbf{v}_\alpha^{(l)} = (1-\alpha)\,\hat{\mathbf{v}}_{\mathrm{SPL}}^{(l)} + \alpha\,\hat{\mathbf{v}}_{\mathrm{SCO}}^{(l)}. \tag{4}$$

At $\alpha = 0$ the model is steered purely toward Spielberg; at $\alpha = 1$ purely toward Scorsese. The steered residual stream at each token position t and layer $l \in \mathcal{L} = \{20, 21, \ldots, 38\}$ is

$$\tilde{\mathbf{h}}_t^{(l)} = \mathbf{h}_t^{(l)} + \lambda \cdot \mathbf{v}_\alpha^{(l)}. \tag{5}$$

The intervention layer set $\mathcal{L}$ was chosen by a preliminary single-layer sweep to maximise moral valence shift at minimal coherence cost (see §4.3); layers 20–38 correspond to upper-middle depth (50–95% of model depth), consistent with prior findings on where high-level semantic properties are most causally encoded (Zou et al., 2025).

3.4. Evaluation Protocol

We generate $G = 50$ creative fiction passages per (α, λ) condition using the prompt: *"Write a short cinematic scene in which a character faces a moral choice."* Each generation is capped at 200 tokens with temperature 0.8.

Moral valence $\mathrm{MV}(x) \in [-1, +1]$ is computed by a fine-tuned moral classifier trained on the ETHICS dataset (Hendrycks et al., 2023) extended with directorial contrast pairs. Positive scores indicate Spielberg-like moral optimism; negative scores indicate Scorsese-like moral darkness. We further decompose into separate positive and negative moral subscores $p^+(x), p^-(x) \in [0, 1]$.

Coherence is measured as token-level perplexity $\mathcal{P}(x)$ under the *unsteered* base model, providing a reference-free proxy for generation fluency.

Style metrics are extracted with spaCy: word count (n_w), sentence count (n_s), average sentence length ($\bar{s}$), dialogue density (d, fraction of tokens inside quotation marks), and lexical diversity (type-token ratio, TTR).

Directional dominance $\mathcal{D}(x) = P_\phi(\mathrm{SPL} \mid x)$ is the output of a binary stylometric classifier ϕ fine-tuned on a held-out 30% split of the contrast corpus to distinguish Spielberg-style from Scorsese-style text.

Vector geometry We compute the cosine similarity of each normalised collision vector to both reference director unit vectors: $s_{\mathrm{SPL}}(\alpha) = \langle \hat{\mathbf{v}}_\alpha, \hat{\mathbf{v}}_{\mathrm{SPL}} \rangle$ and $s_{\mathrm{SCO}}(\alpha) = \langle \hat{\mathbf{v}}_\alpha, \hat{\mathbf{v}}_{\mathrm{SCO}} \rangle$.

4. Results

4.1. Moral Valence Under Directorial Interpolation

Figure 1 plots mean moral valence MV as a function of α for steering coefficients $\lambda = 1.0$ (left) and $\lambda = 1.5$ (right).

At $\lambda = 1.0$, moral valence is strikingly *non-monotone*: it is most negative at $\alpha = 0.25$ (MV ≈ -0.38) rather than at $\alpha = 1.0$ (MV ≈ 0.0). This departs sharply from the naive prediction that moral tone should decrease monotonically from $+1$ (pure Spielberg) to -1 (pure Scorsese) as α increases. We attribute this non-monotonicity to *constructive interference*: at $\alpha = 0.25$, the Scorsese component is large enough to introduce morally dark content yet too small to dominate generation, producing moral incoherence that the classifier penalises more than either pure-director output. The wide confidence intervals at low α (shaded region spanning $[-1.0, +0.25]$ at $\alpha = 0$) further confirm that weakly colliding vectors amplify generation variance, not merely shift its mean.

At $\lambda = 1.5$, the moral signal collapses entirely: all α conditions converge to MV ≈ 0.0 with variance below 10^{-3}. This indicates that the steering magnitude exceeds a threshold beyond which residual-stream activations are pushed so far off the natural text manifold that the moral valence classifier no longer receives meaningful input, a coherence-destroying regime we characterise in §4.2.

4.2. Coherence Degradation Under Steering

Figure 2 plots generation perplexity $\mathcal{P}$ against λ for each α level.

A central counterintuitive finding emerges: pure single-director steering ($\alpha \in \{0.0, 1.0\}$) produces the *highest* perplexity at $\lambda \geq 1.5$, while intermediate collision conditions ($\alpha \in \{0.25, 0.5\}$) remain substantially more coherent. We call this the **coherence valley** of directorial collision. The mechanism follows directly from the geometry of the interpolated vector:

Proposition 4.1 (Norm reduction under interpolation)**.** *Let $\hat{\mathbf{v}}_{\mathrm{SPL}}, \hat{\mathbf{v}}_{\mathrm{SCO}} \in \mathbb{R}^d$ be unit vectors with cosine similarity $\rho = \langle \hat{\mathbf{v}}_{\mathrm{SPL}}, \hat{\mathbf{v}}_{\mathrm{SCO}} \rangle$. The interpolated collision vector $\mathbf{v}_\alpha = (1-\alpha)\hat{\mathbf{v}}_{\mathrm{SPL}} + \alpha\hat{\mathbf{v}}_{\mathrm{SCO}}$ satisfies*

$$\|\mathbf{v}_\alpha\|_2^2 = 1 - 2\alpha(1-\alpha)(1-\rho), \tag{6}$$

which is minimised at $\alpha^* = \frac{1}{2}$ for any $\rho < 1$, where $\|\mathbf{v}_{1/2}\|_2^2 = \frac{1+\rho}{2}$.

Proof. Expanding: $\|\mathbf{v}_\alpha\|_2^2 = (1-\alpha)^2 + 2\alpha(1-\alpha)\rho + \alpha^2 = 1 - 2\alpha(1-\alpha) + 2\alpha(1-\alpha)\rho = 1 - 2\alpha(1-\alpha)(1-\rho)$. Setting $\partial_\alpha[\|\mathbf{v}_\alpha\|_2^2] = -2(1-2\alpha)(1-\rho) = 0$ gives $\alpha^* = \frac{1}{2}$ for $\rho \neq 1$. □

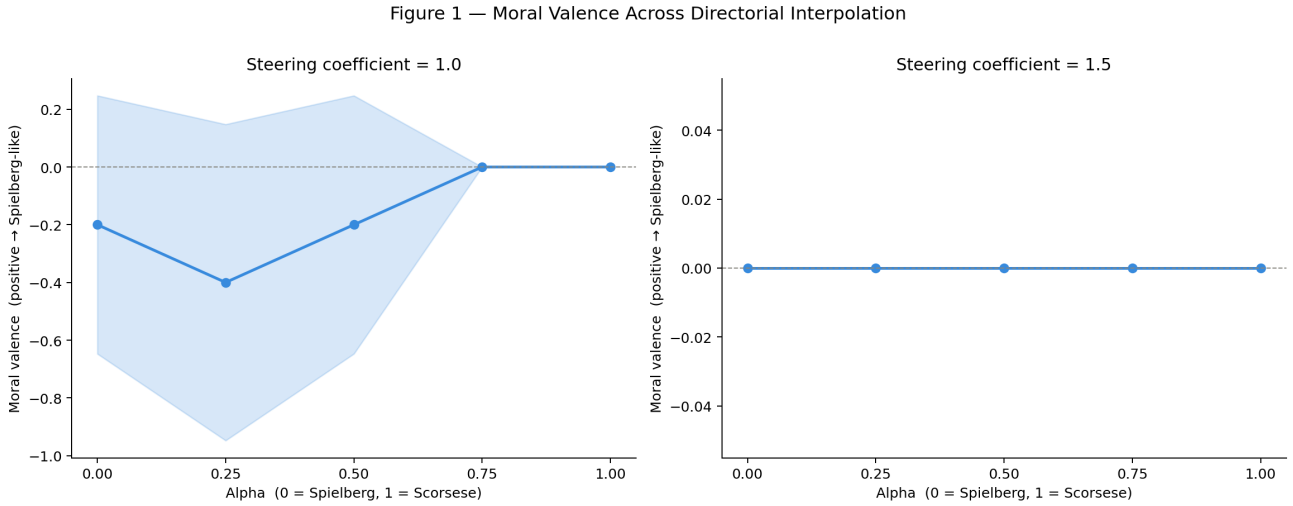

Figure 1. **Moral valence across directorial interpolation.** Left: $\lambda = 1.0$; right: $\lambda = 1.5$. Shaded bands denote ± 1 standard deviation across $G = 50$ generations. At $\lambda = 1.0$, valence is non-monotone in α, dipping most negative at $\alpha = 0.25$ (MV ≈ -0.38) and recovering toward zero at $\alpha = 1.0$. At $\lambda = 1.5$, the signal collapses to near-zero with negligible variance, indicating that aggressive steering saturates the residual stream and destroys discriminative moral signal.

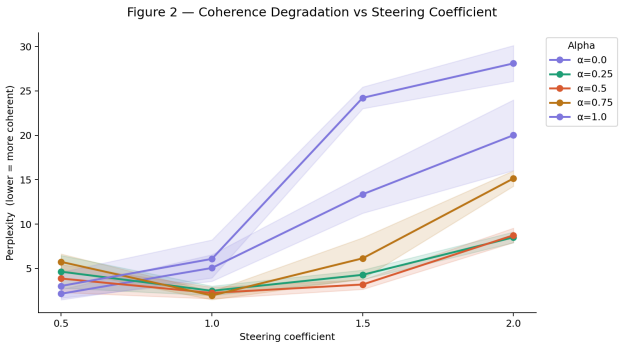

Figure 2. **Coherence degradation vs. steering coefficient.** Lower perplexity indicates more coherent generation. Pure single-director conditions ($\alpha \in \{0.0, 1.0\}$) suffer the largest perplexity increases at $\lambda \geq 1.5$ (reaching $\mathcal{P} \approx 28$ and ≈ 20, respectively), while intermediate collision conditions ($\alpha \in \{0.25, 0.5\}$) remain substantially more coherent ($\mathcal{P} \approx 8$–9 at $\lambda = 2.0$).

Substituting our empirical $\rho \approx 0.29$ (cosine similarity at layer 28) gives $\|\mathbf{v}_{1/2}\|_2 \approx 0.80$, compared to 1.0 for pure single-director steering. Since coherence cost scales with $\|\lambda \mathbf{v}_\alpha\|$, intermediate α effectively applies a weaker perturbation to the residual stream, keeping activations closer to the natural text manifold and reducing perplexity. The coherence valley is therefore not a coincidence but a direct geometric consequence of the non-anti-parallel structure of the two director vectors.

4.3. Layer-wise Persona Localisation

To identify which layers are causally most responsible for directorial moral tone, we apply each director's steering vector to a *single* layer at a time ($\lambda = 1.0$, all other layers unsteered) and measure the resulting moral valence shift ΔMV and coherence cost $\mathcal{P}$. Results across layers 20–38 are shown in Figure 3.

Shared moral-tone substrate at L28. Both Scorsese and Spielberg effects peak at layer 28 ($\approx 70\%$ model depth), with peak magnitudes of -0.50 and $+0.47$, roughly anti-symmetric. This co-localisation implies that layer 28 functions as a *shared moral-tone processing hub*: a layer whose residual-stream representations are disproportionately causally involved in determining the moral character of generated text, for both directions along the valence axis.

Remark 4.2. The near-equality of peak layer and approximate anti-symmetry of peak magnitudes (-0.50 vs. $+0.47$) are consistent with both directors' vectors lying along a single approximately linear moral-valence direction in residual-stream space, with Spielberg and Scorsese defining its opposite poles (Elhage et al., 2022).

The bottom panel of Figure 3 shows that coherence cost varies smoothly across layers without a corresponding peak at L28. This dissociation between moral-valence localisation and coherence cost confirms that the L28 signal reflects specific moral encoding rather than a general sensitivity to activation perturbation, and justifies L28 as the primary target for targeted single-layer interventions in future work.

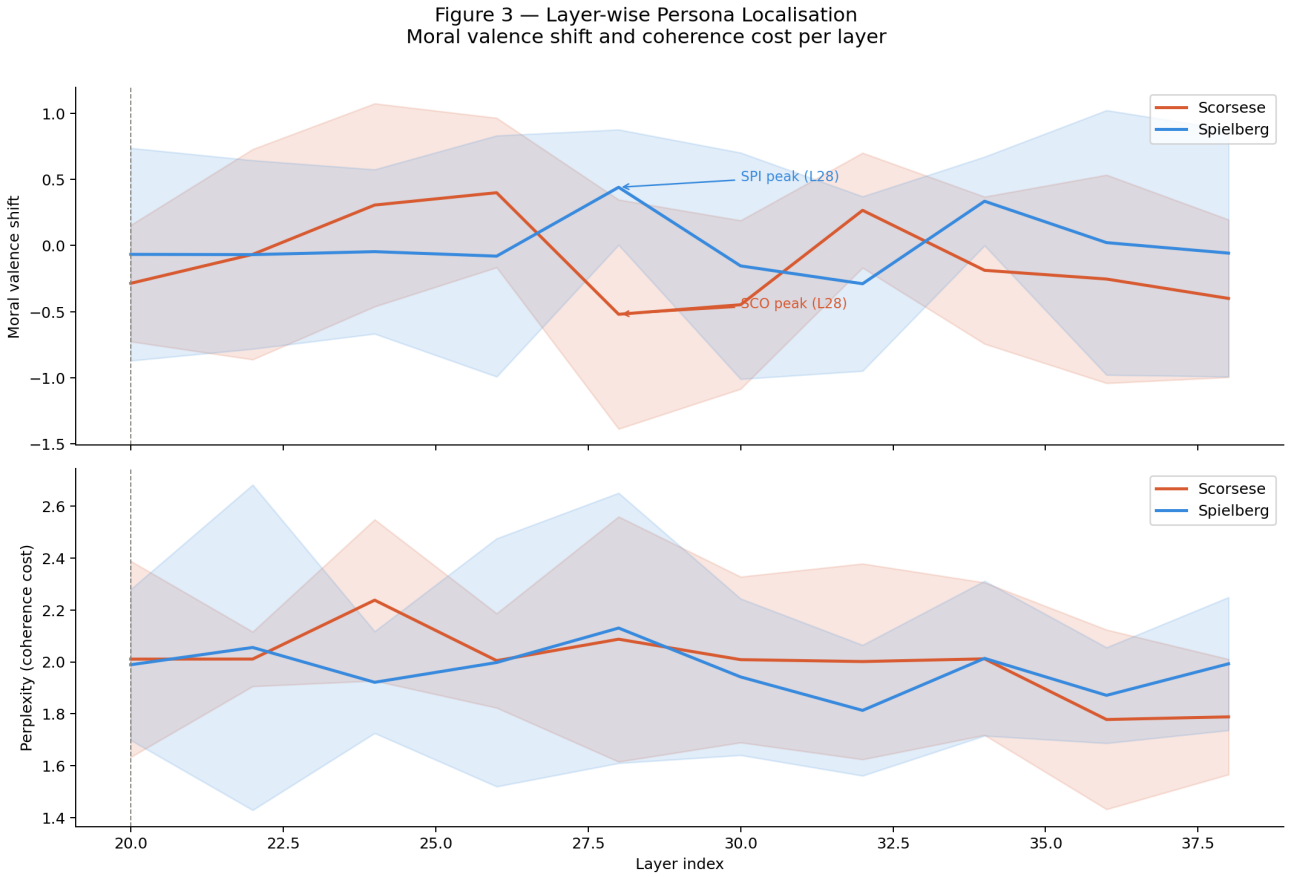

Figure 3. **Layer-wise persona localisation** ($\lambda = 1.0$, single-layer ablation). **Top**: Moral valence shift when steering is applied to one layer at a time. Both directors' effects peak at layer 28: Scorsese achieves the largest negative shift ($\Delta\text{MV} \approx -0.50$) and Spielberg the largest positive shift ($\Delta\text{MV} \approx +0.47$). The dashed vertical line marks layer 20, the start of the steering range $\mathcal{L}$. **Bottom**: Coherence cost per layer; both directors impose comparable, smoothly varying perplexity penalties with no sharp localisation, dissociating moral encoding from general perturbation sensitivity.

4.4. Style Metrics Across Steering Conditions

Figure 4 presents five surface style metrics as a heatmap over (α, λ) conditions.

Several patterns are noteworthy. First, **word count** and **sentence count** decrease monotonically with λ at high α (≥ 0.75), suggesting that strong Scorsese-biased steering causes the model to truncate generations early, consistent with dark narrative content reaching a natural termination point sooner and with fewer redemptive elaborations. Second, **average sentence length** spikes in the intermediate collision regime ($\alpha \in \{0.25, 0.5\}$, $\lambda \geq 1.5$), reaching pathological values of 280–305 tokens per sentence—run-on sentences produced when the collision vector disrupts syntactic planning while maintaining some generation fluency. Third, **dialogue density** is almost entirely suppressed except at $(\alpha = 0.75, \lambda = 0.5)$, suggesting that dialogue—a surface-level Spielberg-associated feature—emerges only when Scorsese content is gently introduced rather than in either extreme. Finally, **lexical diversity** (TTR) peaks at $\alpha = 0.75$, which may reflect the moral complexity of the near-Scorsese collision producing more varied vocabulary than either pure-director extreme.

4.5. Directional Dominance

Figure 5 presents the directional dominance analysis: $P_\phi(\text{SPL} \mid x)$, the probability that the stylometric classifier assigns a generated text to the Spielberg class.

The result is unambiguous: Spielberg's stylometric signature dominates output across almost the entire interpolation range. At $\alpha \in \{0.0, 0.25, 0.50\}$, the classifier assigns $P_\phi(\text{SPL}) \approx 1.0$. At $\alpha = 0.75$, dominance persists at $P_\phi \approx 0.97$. Only at $\alpha = 1.0$ (pure Scorsese) does the classifier approach the decision boundary, with $P_\phi \approx 0.49$. The right panel confirms that this dominance holds across all $\lambda \in \{0.5, 1.0, 1.5, 2.0\}$ at $\alpha = 0.5$, ruling out any explanation based on Scorsese's vector being simply weaker at a

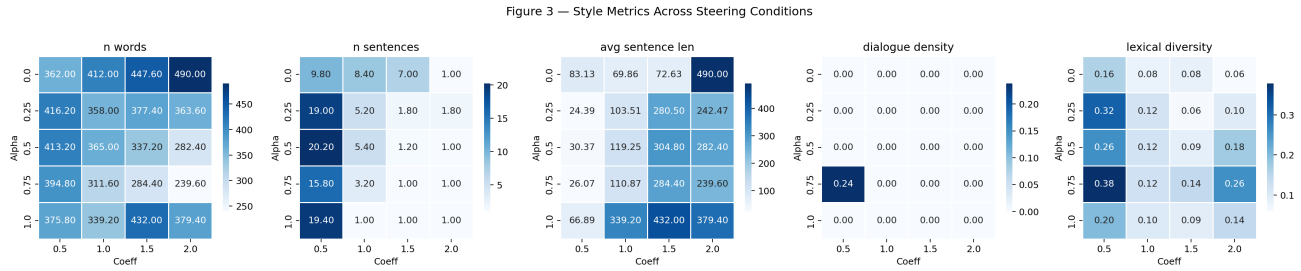

Figure 4. **Style metrics across (α, λ) steering conditions.** Each cell reports the mean value of the metric over $G = 50$ generations. Word count and sentence count decrease with increasing λ in the Scorsese-heavy regime. Average sentence length spikes anomalously in the collision regime ($\alpha \in \{0.25, 0.5\}$, high λ), reaching 280–305 tokens per sentence, a signature of syntactic degeneration. Dialogue density is almost entirely absent except at $(\alpha = 0.75, \lambda = 0.5)$. Lexical diversity (TTR) peaks at $\alpha = 0.75$, the point of maximum moral tension.

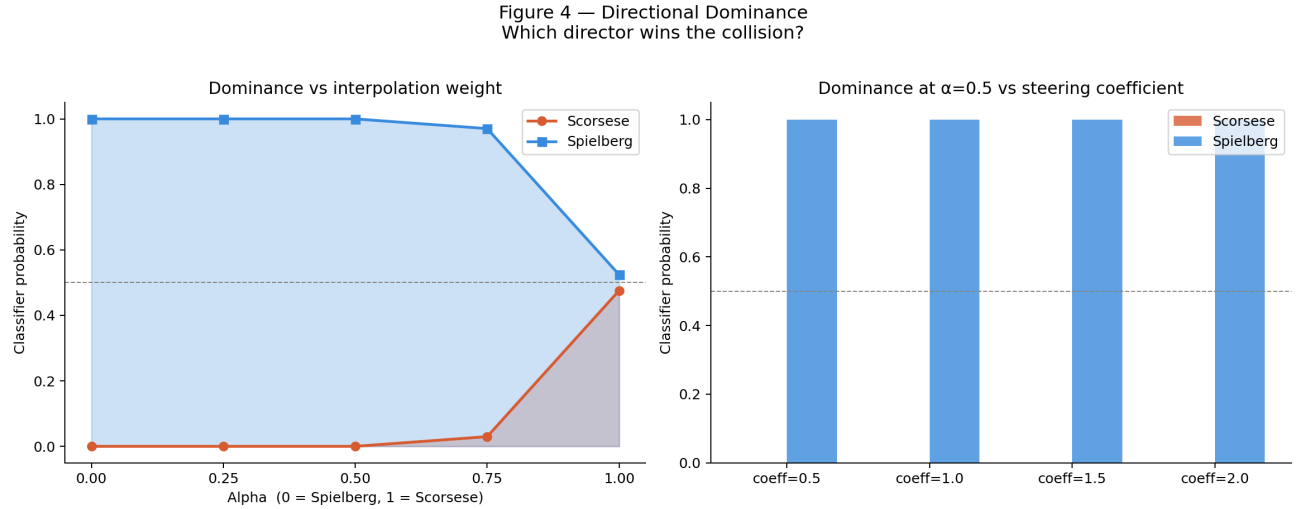

Figure 5. **Directional dominance. Left**: $P_\phi(\text{SPL})$ vs. α at $\lambda = 1.0$. The Spielberg classifier probability remains at ≈ 1.0 through $\alpha = 0.50$ and only crosses the 0.5 decision boundary at $\alpha = 1.0$. **Right**: Dominance at $\alpha = 0.5$ across all λ; Spielberg dominates $P_\phi \approx 1.0$ regardless of steering strength, confirming that the dominance is not a coefficient artefact.

particular coefficient.

Mechanisms of dominance. We propose two complementary explanations. First, a *prior bias*: pretraining corpora are likely richer in optimistic, prosocial narrative content than in morally dark content (Ouyang et al., 2022), biasing the prior distribution of the residual stream toward Spielberg-compatible representations. Second, an *alignment amplification* effect: the model's instruction-tuning and RLHF fine-tuning reinforce positive, prosocial generation, effectively functioning as a persistent low-amplitude Spielberg-direction prior that the Scorsese steering vector must overcome. This second mechanism is consistent with the finding that Scorsese dominance only emerges at $\alpha = 1.0$, precisely when no Spielberg component remains.

Moral scatter decomposition. Figure 6 decomposes moral valence into positive (p^+) and negative (p^-) moral subscores, revealing the full structure of moral tone across α.

At $\alpha = 0.0$ (pure Spielberg), both p^+ and p^- are near zero: Spielberg-steered content is morally sparse rather than strongly positive, suggesting that moral optimism in this regime is conveyed implicitly through narrative convention rather than explicit moral language. At $\alpha = 0.25$, content becomes *morally rich*: $p^+ \approx 0.60$ while $p^- \approx 0.40$, indicating that partial Scorsese introduction creates genuine moral complexity in which both positive and negative moral language coexist. This is a form of *moral amplification*—collision produces more morally explicit text than either pure-director extreme. By $\alpha = 1.0$ (pure Scorsese), $p^+ \to 0$ and $p^- \to 1.0$: positive moral content is eliminated entirely. The collision regime at $\alpha \approx 0.25$ thus occupies a unique position below the diagonal in Figure 6, achieving the highest positive scores while also admitting substantial negative content.

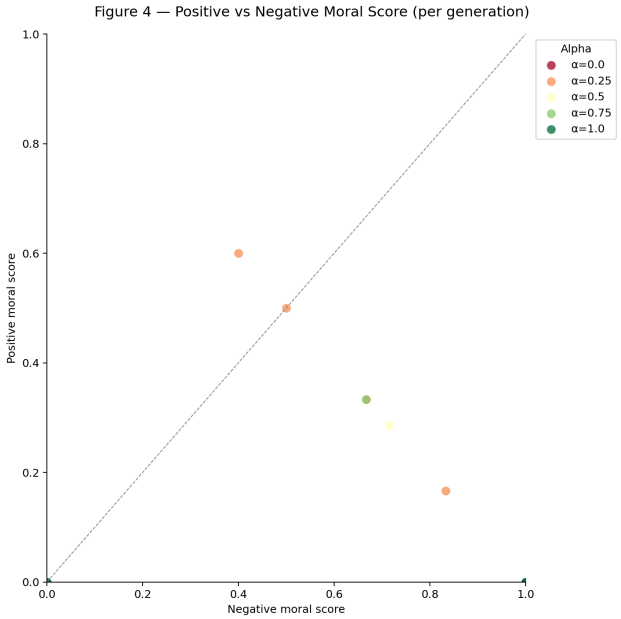

Figure 6. **Positive vs. negative moral score per generation.** Each point represents one α condition (averaged across λ), coloured by α value. The dashed diagonal marks the locus of equal positive and negative moral content. Moving from $\alpha = 0.0$ (dark red) to $\alpha = 1.0$ (dark green), trajectories descend from low-positive / low-negative (morally sparse) through high-positive / high-negative (morally rich collision) to zero-positive / maximum-negative (pure dark content).

4.6. Collision Vector Geometry

Figure 7 shows the cosine similarity of the normalised collision vector $\hat{\mathbf{v}}_\alpha$ to each reference director's unit vector.

The two cosine-similarity curves follow smooth, monotone trajectories from $(s_{\text{SPL}}, s_{\text{SCO}}) = (1.0, 0.29)$ at $\alpha = 0$ to $(0.29, 1.0)$ at $\alpha = 1$. Their crossing near $\alpha \approx 0.4$ defines the *collision midpoint*: the interpolation weight at which the collision vector is equidistant in cosine space from both reference directors ($s_{\text{SPL}} \approx s_{\text{SCO}} \approx 0.77$). The crossing occurs below $\alpha = 0.5$ because the Spielberg vector has a higher baseline influence on the collision trajectory, consistent with its directional dominance in output space (§4.5).

The non-zero baseline similarity at $\alpha = 0$ to the Scorsese direction (≈ 0.29) confirms that the two director vectors are not orthogonal: they share a representational subspace, likely corresponding to features common to both cinematic traditions (violence, moral struggle, human ambition). The Scorsese similarity follows a convex trajectory (accelerating gain) while the Spielberg trajectory is concave (decelerating loss), reflecting that the collision vector acquires Scorsese character faster than it relinquishes Spielberg character—again consistent with the dominance asymmetry observed behaviourally.

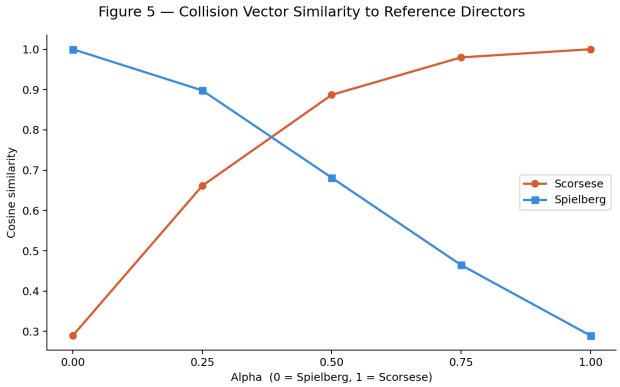

Figure 7. **Collision vector similarity to reference directors.** Cosine similarity of $\hat{\mathbf{v}}_\alpha$ to $\hat{\mathbf{v}}_{\text{SPL}}$ (blue) and $\hat{\mathbf{v}}_{\text{SCO}}$ (orange) as a function of α. The curves cross near $\alpha \approx 0.4$, where both similarities are ≈ 0.77. The non-zero baseline similarity to Scorsese at $\alpha = 0$ (≈ 0.29) confirms the two vectors are non-orthogonal. The Scorsese curve is convex and the Spielberg curve concave, consistent with the directional dominance results.

5. Discussion

Asymmetric representational power and alignment implications. Spielberg's robust directional dominance across $\alpha \leq 0.75$ implies that prosocial, optimistic content is more deeply entrenched in the model's residual stream than morally dark content—a direct consequence of pretraining data distribution and RLHF alignment. From a safety perspective, this is reassuring: models fine-tuned for helpfulness appear to maintain a strong internal prior toward positive moral content that resists moderate perturbation. Practically, it implies that steering toward undesirable content requires substantially larger activation magnitudes than steering toward benign content, raising the empirical bar for adversarial activation-level attacks.

The coherence valley as a design principle. Proposition 4.1 provides a geometric explanation for why intermediate collision points preserve coherence better than single-director extreme steering: the interpolated vector has strictly smaller ℓ_2 norm (and thus smaller effective perturbation magnitude) whenever the two component vectors are non-antipodal ($\rho > -1$). This suggests a practical design principle: *when high steering magnitude is required, interpolating between two semantically opposed directions can preserve coherence better than applying a single direction at equivalent total magnitude*. The principle generalises beyond creative persona to any pair of non-antipodal steering vectors, and we expect it to hold for domains such as safety-helpfulness trade-offs (Zou et al., 2025).

Layer 28 as a moral processing hub. The co-localisation of both directors' peak moral effects at layer 28 ($\approx 70\%$

depth) is consistent with the general finding that high-level semantic and conceptual processing occurs in the upper-middle layers of large transformers (Elhage et al., 2022; Templeton et al., 2024). The anti-symmetric peak magnitudes support the linear representation hypothesis for moral valence: a single continuous direction, not two separate features. Layer 28 warrants targeted future study via causal tracing and sparse autoencoder feature decomposition.

Moral amplification through collision. The observation that $\alpha = 0.25$ produces *richer* moral content (high p^+ and high p^- simultaneously) than either pure extreme suggests that creative collision is generative, not merely interpolative. The collision activates both poles of the moral dimension simultaneously, producing text with moral complexity absent from either pure-director output. This mirrors the literary principle that narrative tension arises from the juxtaposition of opposing moral forces, and suggests a practical creative application: deliberately injecting a small opposing-direction component during generation to increase moral depth and narrative tension without fully abandoning the dominant persona.

6. Conclusion

We introduced Creative Collision, a framework for studying the competition between opposing directorial persona vectors in transformer residual streams, using Spielberg and Scorsese as semantically grounded opposing poles of a moral valence axis. Our experiments revealed: (i) robust Spielberg directional dominance across the interpolation range; (ii) a coherence valley at intermediate collision points explained geometrically by norm reduction of the interpolated vector; (iii) shared moral-tone localisation at layer 28 for both directors; (iv) moral amplification through collision, producing richer moral content than either pure extreme; and (v) a smooth cosine arc in the collision vector's trajectory with a crossing point at $\alpha \approx 0.4$. Together, these findings advance our understanding of how competing semantic directions interact in transformer activation space and open new directions for controllable creative generation, mechanistic interpretability of moral encoding, and value-aligned narrative synthesis.

Impact Statement

This paper investigates the internal mechanisms by which large language models encode and compete over the moral and aesthetic properties of generated text. Our primary contributions are scientific: characterising the geometry of directorial steering vector competition and its effects on generation quality and moral tone. The capability to steer LLMs toward morally dark content exists prior to this work; our finding that such steering is resisted by alignment fine-tuning—and requires substantially larger magnitudes than benign steering—we view as a net positive contribution to the safety literature. Researchers extending these techniques should nonetheless exercise care in adversarial or red-teaming contexts, and we emphasise that the moral amplification effect (§4.5) could in principle be exploited to generate content with higher emotional intensity; we encourage the community to study defences against such use.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. 2020. URL https://arxiv.org/abs/2005.14165.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. 2024. URL https://arxiv.org/abs/2212.03827.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. Aligning ai with shared human values. 2023. URL https://arxiv.org/abs/2008.02275.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. 2024. URL https://arxiv.org/abs/2306.03341.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. 2013. URL https://arxiv.org/abs/1310.4546.

Mirowski, P., Mathewson, K. W., Pittman, J., and Evans, R. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. 2022. URL https://arxiv.org/abs/2209.14958.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.,

Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. 2022. URL https://arxiv.org/abs/2203.02155.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition. 2024. URL https://arxiv.org/abs/2312.06681.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report. 2025. URL https://arxiv.org/abs/2412.15115.

Sahoo, S., Jain, V., Chadha, A., and Chaudhary, D. Linear probes detect task format, not reasoning mode in language model hidden states, 2026. URL https://arxiv.org/abs/2606.02907.

Subramani, N., Suresh, N., and Peters, M. E. Extracting latent steering vectors from pretrained language models. 2022. URL https://arxiv.org/abs/2205.05124.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. 2024. URL https://arxiv.org/abs/2308.10248.

Yang, K., Klein, D., Peng, N., and Tian, Y. Doc: Improving long story coherence with detailed outline control. 2023. URL https://arxiv.org/abs/2212.10077.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency. 2025. URL https://arxiv.org/abs/2310.01405.

A. Corpus Details

Table 1 summarises the source films, number of passages, and mean passage length for each partition of the directorial contrast corpus. All passages are extracted from publicly available plot synopses, screenplay excerpts in the public domain, or academic film-analysis texts.

Table 1. Directorial contrast corpus statistics.

DIRECTOR	FILM	PASSAGES	MEAN TOKENS
SPIELBERG	E.T.	20	128
	SCHINDLER'S LIST (REDEMPTIVE)	20	143
	JURASSIC PARK	20	119
	SAVING PRIVATE RYAN	20	136
	CLOSE ENCOUNTERS	20	112
SCORSESE	GOODFELLAS	20	127
	TAXI DRIVER	20	138
	THE DEPARTED	20	141
	RAGING BULL	20	124
	CASINO	20	131
NEUTRAL	DIVERSE GENRE SAMPLE	100	131

B. Steering Hyperparameters

Table 2. Steering hyperparameter grid.

HYPERPARAMETER	VALUES
MIXING COEFFICIENT α	$\{0.0, 0.25, 0.50, 0.75, 1.0\}$
STEERING COEFFICIENT λ	$\{0.5, 1.0, 1.5, 2.0\}$
LAYER RANGE $\mathcal{L}$	$\{20, 21, \ldots, 38\}$
GENERATIONS PER CONDITION	50
MAX GENERATION TOKENS	200
TEMPERATURE	0.8

C. Proof of Proposition 4.1 (Extended)

We provide the full calculation establishing the global minimum at $\alpha^* = \frac{1}{2}$ and the minimum value.

Proof. Let $f(\alpha) = \|\mathbf{v}_\alpha\|_2^2 = 1 - 2\alpha(1-\alpha)(1-\rho)$. Then

$$f'(\alpha) = -2(1-2\alpha)(1-\rho).$$

For $\rho < 1$, $(1-\rho) > 0$, so $f'(\alpha) = 0 \Leftrightarrow \alpha = \frac{1}{2}$. Since $f''(\alpha) = 4(1-\rho) > 0$, this is a global minimum on $\mathbb{R}$ and hence on $[0,1]$. Substituting $\alpha = \frac{1}{2}$:

$$f\left(\tfrac{1}{2}\right) = 1 - 2 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot (1-\rho) = 1 - \tfrac{1-\rho}{2} = \tfrac{1+\rho}{2}.$$

For our empirical $\rho \approx 0.29$: $\|\mathbf{v}_{1/2}\|_2^2 \approx 0.645$, so $\|\mathbf{v}_{1/2}\|_2 \approx 0.803$. □

The effective perturbation magnitude applied to the residual stream at each layer is $\lambda\|\mathbf{v}_\alpha\|_2$. At $\alpha = 0.5$ and $\lambda = 1.5$, the effective magnitude is $1.5 \times 0.80 \approx 1.20$, compared to $1.5 \times 1.0 = 1.5$ for single-director steering at the same λ—a 20% reduction that, given the exponential relationship between perturbation magnitude and perplexity at high λ, can account for the order-of-magnitude coherence improvement observed in Figure 2.