

# Activation Steering for Creative Estrangement: Generating Evocative Narrative Through Controlled Disruption of Language Models

Olga Vechtomova\*, Greg French\*, Xavier Schneider\*, Olivier Poulin\*, Jeff Bos\*\*

\*University of Waterloo, Canada

\*\*WordSynth Inc.

{ovechtom, g2french, xschneider, opoulin}@uwaterloo.ca, jeff@wordsynth.com

## Abstract

We investigate activation steering as a method for generating extended narrative that can function as a textual environment for creative inspiration. Using contrastive pairs inspired by the aesthetics of Tarkovsky and Lynch, we construct steering vectors encoding modes of perceptual estrangement and apply them to Llama 3.1 8B Instruct. Single-layer steering operates in two regimes: a moderate regime that elevates literary register without producing estrangement, and a high regime where estrangement emerges but coherence collapses. Dual-layer steering resolves this by applying a compositional vector at layer 13 simultaneously with a content-level vector at layer 14, accessing a region where estrangement and coherence coexist. The dual-steered outputs exhibit emergent disruptions (person shifts, narrative self-contradiction, referential instability) absent from all training pairs, arising from competing representational pressures during generation. Comparing steered outputs against explicitly prompted fragmentation reveals two qualitatively distinct modes: prompting produces conventional fragmentation devices, while steering produces intact sentences whose content is self-contradicting. A blind analysis of evaluator reasoning texts and human reader feedback confirm this distinction. Human readers tended to perceive dual-steered outputs as stranger than prompted outputs, suggesting that occasional and surprising disruption within coherent prose is more estranging than pervasive and categorizable fragmentation.<sup>1</sup>

## Introduction

Viktor Shklovsky’s concept of *ostranenie* (defamiliarization or making strange) proposes that art’s essential function is to restore how things are perceived rather than how they are known (Shklovsky 1965). The target of estrangement is not the exotic but the ordinary, seen as if for the first time. We see this across artistic domains: in Andrei Tarkovsky’s *Solaris* (1972), rain falls inside a cabin and a father does not react; in David Lynch’s *Black Lodge* (*Twin Peaks*, 1990), characters speak in reversed recordings and coffee does not pour from an inverted cup. In each case, the estrangement operates through minimal displacement from the familiar.

<sup>1</sup>Supplementary material website:

<https://ovechtom.github.io/supplementary-material/>

Petric (1989) describes Tarkovsky’s films as suffused with an oneiric air that transcends the commonplace signification of objects to reach what the eye ordinarily neglects. Rodley (2005) characterizes Lynch’s work in similar terms: Lynch’s interest lies in the defamiliarization process and the waking/dream state, rather than in incongruity as spectacle or overt surrealism. This is close to Shklovsky’s “stone made stony”, not by adding to it but by making us see it again.

We ground our approach in these filmmakers’ aesthetics because their estrangement targets *perception itself* rather than the content of reality. Genre estrangement (fantasy, science fiction, magic realism) introduces rule-governed alterations that the reader eventually assimilates; cognitive tension resolves as the rules become clear. Tarkovsky and Lynch embed unexplained impossibilities within naturalistic observation. Neither the reverse speech inside the Black Lodge in Lynch’s *Twin Peaks*, nor the rain inside the cabin in Tarkovsky’s *Solaris* is ever explained. More fundamentally, the ontological status of the figures in the Lodge and of those who appear on the station in *Solaris* is left unresolved. Hari appears to Kelvin as his wife, long dead, and comes to understand that she is not. What the Lodge figures are is never specified. As Petric (1989) observed, the viewer feels that something is “wrong” but cannot detect sufficient proof to discredit the events on the basis of everyday logic. The indeterminacy is sustained and is often left unresolved, maintaining the heightened attention that Berlyne’s inverted U-curve (Berlyne 1971) identifies as maximally engaging. This choice shapes our experimental design. Because we model disruptions of perception rather than alterations of content, the resulting steering vectors encode abstract perceptual modes (temporal self-contradiction, ontological uncertainty, identity instability) rather than genre conventions.

Defamiliarization of this kind has direct implications for computational creativity. Rubin (2023) describes the Seed phase of the creative process as the stage where “we are completely open to any stimulus that might lead us to that state of mind where art becomes inevitable.” At this stage, artists need stimuli that are indeterminate, surprising, and open to interpretation. Leonardo Da Vinci advised young artists to look at old stained walls and streaked stones for inspiration (Turner 2011). The stains are not art in themselves but projective stimuli whose meaning is unresolved.

The artist forms meaning through the process of attempted resolution, and it is this process that activates the creative state. Prior research demonstrated this with LSTM-based Variational Autoencoders trained on small curated datasets, which generate lines characterized by syntactic indeterminacy and semantic openness (“clouds humanity without light”, “through the wrong moon”) that function effectively as seeds for creative inspiration (Vechtomova, Sahu, and Kumar 2021). However, this approach operates at the sentence level. The question then becomes: can we generate extended narrative that sustains these qualities, text that functions not as a finished story but as a *textual environment* for creative inspiration?

Activation steering is a technique that modifies a language model’s internal representations during generation by adding directional vectors to hidden states. In this paper, we demonstrate that this technique can produce long-form narrative with precisely these qualities. We steer Llama 3.1 8B Instruct using contrastive vectors derived from paired literary and neutral passages and evaluate the outputs using a seven-metric framework scored by LLM evaluator (Claude Sonnet 4.5). Our key findings are:

1. **Two regimes of first-order steering.** Steering at layer 13 operates in two distinct regimes: a moderate regime ( $\alpha$  2.0–6.0) that elevates literary register without producing estrangement, and a high regime ( $\alpha$  7.5–8.0) where estrangement and syntactic disruption emerge abruptly but coherence collapses. No first-order alpha value produced both high estrangement and high coherence.
2. **Dual-layer steering achieves coherent estrangement.** Applying a first-order vector at layer 13 simultaneously with a second-order content-level vector at layer 14 produces estrangement comparable to the highest first-order alphas while maintaining substantially higher coherence, accessing a region single-layer steering does not reach.
3. **Emergent syntactic estrangement.** Dual-steered outputs exhibit syntactic disruptions (person shifts, temporal self-contradiction, logical impossibility) absent from all training pairs. These features emerge from the interaction of competing representational pressures during generation.
4. **Disruption of syntax vs. disruption through syntax.** Explicitly prompted fragmentation produces conventional devices (ellipses, sentence fragments, lowercase). Steering produces grammatically intact sentences whose content is self-contradicting or ontologically unstable.

## Related Work

### Defamiliarization and Aesthetic Engagement

The relationship between aesthetic engagement and stimulus complexity follows an inverted U-shaped curve, also known as Wundt effect, first formalized by Berlyne (1971). Completely predictable stimuli fail to hold attention; completely random stimuli offer no foothold for pattern recognition; stimuli of intermediate complexity produce the greatest engagement. Gold et al. (2019) offered a learning-based explanation for this effect: when events are completely predictable or fully random, little learning occurs as patterns

are either trivial or impossible to discern. Intermediate predictability creates conditions for meaningful cognitive engagement, as the perceiver is rewarded for the effort of partial pattern recognition. As Pepperell (2006) noted, indeterminate visual images create heightened perception and attention as habitual recognition is suspended, a state that is itself a precondition for creative work. Zeilinger (2023) extended this to AI-generated visual art, arguing that visual indeterminacy invites active meaning-making rather than passive reception.

Artists have long developed deliberate techniques for inducing this productive disruption. The cut-up technique, from Tzara through Burroughs, fragments existing text to break habitual language patterns (Burroughs 2012). David Bowie described its function as “igniting anything that might be in my imagination” (BBC 1975), aligning it with the Seed phase of creativity. However, pure randomness does not produce art. As Nees (1969) observed, there must be structure to narrow random outcomes into discernible aesthetic information. The productive zone lies between predictability and randomness, the same intermediate territory that the Wundt curve identifies as maximally engaging.

Tarkovsky and Lynch exemplify this productive zone in cinema. Unlike genre estrangement, which introduces rule-governed alterations that the viewer eventually assimilates, their work embeds unexplained impossibilities within naturalistic observation, maintaining the state of heightened perception that Pepperell identifies as a precondition for creative work.

### Computational Approaches to Creative Text Generation

Boden’s influential typology distinguishes combinational, exploratory, and transformational creativity (Boden 2004). In exploratory creativity, a system searches within an established conceptual space; in transformational creativity, the space itself is altered to make previously impossible outputs conceivable. Wiggins (2006) formalized this distinction, providing a mathematical framework for describing creative systems in terms of the conceptual spaces they traverse and transform. Colton’s creative tripod framework argues that a system perceived as creative must demonstrate not only skill but also appreciation and imagination (Colton 2008), and the FACE/IDEA descriptive models provide formal criteria for evaluating creative acts (Colton, Charnley, and Pease 2011).

Within computational text generation, Oliveira’s PoEryMe (Oliveira 2012) demonstrated autonomous poetry generation using semantic and syntactic templates, while Gervás (2009) surveyed computational approaches to storytelling with attention to how creativity manifests in narrative structure. More recently, LLMs have prompted new questions about machine creativity. Franceschelli and Musolesi (2025) analyzed LLM creativity through the lens of Boden’s typology and Colton’s tripod, while Ismayilzada et al. (2025) found that LLMs generate stylistically complex stories but fall short on novelty and surprise compared to human writers. In human-AI co-creative writing, Calderwood et al. (2020) found that novelists use generative models as constraining tools that push the boundaries of their writing.

This finding aligns with our interest in text that functions as creative stimulus rather than finished output.

The connection between Shklovsky’s defamiliarization and computational text generation has been explored in prior work on neural poetry generation (Vechtomova and Sahu 2025), where LSTM-VAE outputs were analyzed as instances of productive estrangement. The present work asks whether this connection holds at the level of extended narrative. Ho et al. (2025) proposed a “Familiarity-Estrangement Space” that uses topic modeling to project textual materials along dimensions of topical familiarity and novelty, allowing users to browse for texts at a desired level of thematic strangeness. Their system operates on existing text rather than generating new text with estranging qualities.

The distinction between systems that generate polished outputs (*pushing* content toward the artist) and systems that generate indeterminate stimuli (*pulling* the artist into the creative state) is central to understanding how computation can support creativity. LSTM-VAE sentence generators (Vechtomova, Sahu, and Kumar 2021) are pull-oriented. They produce open, indeterminate fragments rather than finished text. But the architecture operates at the sentence level, generating individual lines rather than extended narrative that sustains these qualities across a unified passage. The present work is also pull-oriented. It aims to generate not polished stories but textual environments for creative inspiration. In Boden’s terms, LSTM-VAE models perform exploratory creativity within a learned latent space. The present work investigates whether activation steering can achieve something closer to transformational creativity by altering the model’s generative space to produce outputs with qualities absent from any training example.

### Activation Steering

Activation steering modifies a language model’s hidden states during inference by adding learned directional vectors, enabling fine-grained control over output properties without retraining. Subramani et al. (2022) first demonstrated that steering vectors can be extracted from pretrained language model decoders and applied via addition to hidden states, achieving high-fidelity generation control and unsupervised sentiment transfer. Turner et al. (2023) introduced Activation Addition (ActAdd), computing mean activation differences between contrastive prompt pairs and injecting the resulting vector during generation to reliably modulate behavioral properties such as sycophancy and hallucination tendency. Rinsky et al. (2024) extended this to Llama 2, systematically evaluating steering effectiveness across layers and scaling factors. In parallel, Zou et al. (2023) proposed Representation Engineering (RepE), a top-down framework that reads and controls high-level concepts (honesty, fairness, harmfulness) in LLM internal representations at the population level rather than through individual neurons or circuits. Li et al. (2023) demonstrated Inference-Time Intervention (ITI), shifting activations during inference to improve truthfulness, suggesting that models possess internal representations of properties they do not reliably express.

The present work applies activation steering to a qualita-

tively different domain: generating aesthetically productive text for creative inspiration. Rather than steering toward or away from a binary behavioral property, we steer along a continuous dimension, from prosaic description to literary estrangement, defined by contrastive pairs that model specific aesthetics of defamiliarization. This requires navigating a narrow productive zone where the output is sufficiently estranged to defamiliarize but sufficiently coherent to sustain engagement. This challenge has no direct analogue in the behavioral steering literature.

## Method

### Steering Mechanism

**Hidden states and the residual stream.** In a decoder-only transformer such as Llama 3.1 8B Instruct ( $N=32$  layers, hidden dimension  $H=4096$ ), each layer *adds* its computation to a running per-token vector (the *residual stream*) rather than overwriting it. The layer- $\ell$  hidden state is the model’s internal representation of a token at depth  $\ell$ . Activation steering modifies these hidden states during generation, leaving weights and prompt unchanged.

**Steering vectors.** A *steering vector* is a direction in hidden-state space which, added to the residual stream, biases generation toward a target property. From  $n$  contrastive pairs of matched neutral and literary passages, we mean-pool the layer- $\ell$  hidden states over the final 32 token positions of each side, giving  $h_\ell(x_i^+)$  and  $h_\ell(x_i^-)$ . The steering vector is their average per-pair difference, normalized to unit length:

$$d_\ell = u / \|u\|, \quad u = \frac{1}{n} \sum_{i=1}^n (h_\ell(x_i^+) - h_\ell(x_i^-)).$$

Averaging cancels passage-specific content and retains the shared direction (Turner et al. 2023; Rinsky et al. 2024). Normalization makes  $d_\ell$  a pure direction; magnitude is set at generation time.

**First- and second-order vectors.** We distinguish two kinds of steering vector by what their contrastive pairs hold fixed. A *first-order* vector is built from the neutral-to-literary contrast, where the two sides differ in both register and perceptual mode, and encodes the full shift from plain description to estranged writing. A *second-order* vector is built *within* literary prose: both sides are already literary and matched for register and imagery, and differ only in whether an ontologically impossible event occurs. It therefore isolates content-level estrangement from register elevation. The terms are our nomenclature, not standard usage.

**Injecting the vector during generation.** At inference we add the steering vector to the residual stream at layer  $\ell$  at every generated token. Writing  $h_t^{(\ell)}$  for the layer- $\ell$  hidden state at step  $t$ :

$$h_t^{(\ell)} \leftarrow h_t^{(\ell)} + \alpha s(t) d_\ell.$$

Because  $d_\ell$  has unit norm, the steering strength  $\alpha$  is exactly the Euclidean length of the injected vector, so  $\alpha$  is comparable across vectors and layers. We implement the addition as a HuggingFace forward hook on the chosen transformer block. Steering acts only on tokens the model generates; the prompt is encoded unmodified, and a displacement applied at one step influences every token that follows.

**Decay schedule.** A perturbation held at constant strength compounds through cached attention states and degrades the passage into repetition before it ends. We therefore attenuate with a cosine schedule,

$$s(t) = \frac{1}{2} \left( 1 + \cos \frac{\pi t}{k} \right) \text{ for } t < k, \quad s(t) = 0 \text{ otherwise,}$$

with  $k=200$ , so steering falls smoothly from full strength  $\alpha$  at the first generated token to zero at token 200, letting the model resolve the displaced trajectory over the second half of the 400-token passage. All generations use temperature 0.9 (chosen to characterize the steered output *distribution* rather than a single deterministic trajectory), nucleus sampling ( $p=0.9$ ), a maximum of 400 new tokens, and a fixed random seed.

**Dual-layer steering.** *Dual-layer* steering registers two hooks simultaneously: a first-order vector at layer  $\ell_1$  and a second-order vector at a nearby layer  $\ell_2$ , with independent strengths  $\alpha_1, \alpha_2$  sharing the cosine schedule. The displacement added at  $\ell_1$  propagates forward and is still present when the second vector is added at  $\ell_2$ , so the two pressures (one elevating literary register, the other destabilizing the represented reality) act on the same generation jointly. Layers and strengths are reported below.

### Steering Vector Construction

We construct steering vectors from contrastive pairs of matched passages describing similar scenarios, where one version is written in neutral, expository prose and the other employs techniques of literary estrangement. The vector is the mean difference between the model’s internal representations of the literary and neutral sides, capturing the direction in activation space that corresponds to the transformation from prosaic description to estranged perception.

We designed five sets of 12 pairs (60 total), each embodying a distinct mode of estrangement: (1) *Contemplative*, modeled on Tarkovsky’s sustained perceptual attention; (2) *Lynchian*, the eruption of wrongness within the familiar; (3) *Lost Highway*, identity instability and spatial inconsistency; (4) *Stalker*, landscape as responsive environment; and (5) *Solaris*, ontological uncertainty and memory as material substance. In each pair, the neutral side is a bland, surface-level recounting of events or surroundings, while the literary counterpart describes the same scenario with imagistic depth and a specific mode of estrangement. The neutral and literary sides are length-matched to within 15%. The pairs were developed through iterative collaboration between the researchers and Claude Opus 4.6 (Anthropic), with researchers specifying target aesthetic modes

Table 1: Mean scores for neutral and literary sides of all 60 training pairs.

| Metric          | Neutral | Literary |
|-----------------|---------|----------|
| Imagery         | 1.75    | 3.01     |
| Indeterminacy   | 1.09    | 4.14     |
| Originality     | 1.00    | 3.98     |
| Register        | 1.00    | 3.47     |
| Coherence       | 4.97    | 4.73     |
| Estrangement    | 1.03    | 4.42     |
| Syntactic estr. | 1.00    | 1.63     |

and evaluating candidates against theoretical criteria across multiple rounds of refinement. The resulting pairs were validated independently using the blind identification procedure described below.

### Layer Selection

We swept all 32 layers of Llama 3.1 8B Instruct, computing the contrastive vector at each layer and evaluating four metrics: direction norm (L2 magnitude), relative L2 (normalized by activation scale), consistency (within-set pairwise cosine similarity), and cross-file alignment (between-set cosine similarity). Direction norm peaked at final layers, merely tracking activation magnitude. Relative L2 peaked at early layers, which show high relative differences for any contrastive signal. The two predictive metrics, consistency and cross-file alignment, both peaked at layer 13 (consistency 0.654, cross-file alignment 0.741). Full layer analysis is available on the supplementary materials website. The *Solaris* pairs produced the strongest signal (consistency 0.704), consistent with ontological uncertainty requiring the deepest semantic displacement.

Layer 13 sits at approximately 40% depth, a region associated with semantic composition rather than discourse planning (Tenney, Das, and Pavlick 2019; Wendler et al. 2024). Steering at this layer targets *how the model constructs meaning* while leaving its coherence and syntactic planning machinery undisturbed. All first-order experiments accordingly use the composite vector at layer 13, injected with the cosine-decayed schedule.

### Validation of Training Pairs

We evaluated both sides of all 60 pairs using the same metric pipeline applied to steered outputs (Table 1). The literary sides scored substantially higher on estrangement (4.42 vs. 1.03), indeterminacy (4.14 vs. 1.09), and originality (3.98 vs. 1.00), all  $p < .001$  (Welch’s  $t$ -test). Crucially, coherence was near-equivalent (4.73 vs. 4.97), confirming that the contrastive signal captures qualitative transformation rather than fragmentation. The estrangement gap (3.39) exceeded the register gap (2.47), indicating that the primary dimension encoded is perceptual defamiliarization rather than register inflation.

A blind identification procedure provided convergent evidence that the five sets encode distinct aesthetic dimensions. An independent LLM evaluator (GPT-4o, OpenAI),

given each set without labels, correctly identified the dominant aesthetic mode for three of five sets and swapped only the two Lynch-derived sets, which is an expected confusion given their shared origin.

## Evaluation Metrics

**Choice of evaluators.** We chose Claude Sonnet 4.5 (Anthropic) as the primary evaluator because prior work on aesthetic evaluation of literary text (Sahu and Vechtomova 2025) found that frontier models with explicit reasoning produce reliable discrimination on literary dimensions where smaller or less capable evaluators do not. We chose GPT-4o (OpenAI) as the cross-validator specifically because it comes from a different model family and training pipeline than Claude, so agreement between the two provides stronger evidence than agreement within a single family. For token-level surprisal we use GPT-2 (117M) rather than a peer-scale reference: a small, weak model captures token-level statistical regularity rather than the same semantic competencies the steered model possesses, avoiding the circularity of using the steered model’s own representational neighborhood as the standard against which to measure surprisal. The blind reasoning-text analysis uses Claude Opus 4.6, which had not seen the evaluation rubric during scoring.

**Scoring protocol.** We evaluate steered outputs using seven metrics scored on a 1–5 scale by Claude Sonnet 4.5, following a similar protocol to (Sahu and Vechtomova 2025) but with a modified set of metrics and a different LLM evaluator. We queried the LLM evaluator five times per story per metric at temperature 0.7 and averaged the resulting scores; temperature 0.7 with averaging reduces the influence of any single sampling artifact in the evaluator’s rationale, whereas temperature 0 would lock onto one rationale that may not be representative. Each evaluation prompt requires the model to produce a one-sentence reasoning statement before committing to a numerical rating. This forced reasoning step proved essential during development. Without it, the evaluator exhibited reduced discrimination, conflating rhetorical intensity with genuine literary achievement. Ornate but semantically closed prose (“*a symphony of despair, a requiem for the living*”) scored comparably to compressed but genuinely open prose (“*I am a cry without a mouth*”). The reasoning step forces the evaluator to articulate *why* a passage merits a given score, producing reliable discrimination between register elevation and aesthetic transformation.

**Choice of metrics.** The seven metrics are: *imagery* (sensory vividness), *indeterminacy* (semantic openness and resistance to singular interpretation), *originality* (unusual collocations, surprising metaphors, unexpected verb choices), *register* (distance from plain prose toward heightened literary language), *coherence* (logical connection and overall readability), *estrangement* (defamiliarization of ordinary experience), and *syntactic estrangement* (departures from conventional prose syntax, including person shifts, sentence fragmentation, and sentences that undermine their

own logic). Imagery, originality, and coherence are conventional in literary-generation evaluation (Sahu and Vechtomova 2025; Ismayilzada, Stevenson, and van der Plas 2025); indeterminacy and estrangement operationalize the aesthetic targets of our theoretical framing (Shklovsky 1965; Pepperell 2006; Zeilinger 2023); register and syntactic estrangement were added to test the disruption-of/through distinction. The first five characterize the *mode* of the prose; estrangement measures the overall effect; syntactic estrangement captures a specific mechanism that proved critical for distinguishing first-order from dual-layer steering. The training-pair validation (Table 1) provides evidence that the metrics discriminate as intended: neutral and literary sides separate cleanly on the targeted dimensions while coherence remains near-equivalent, which is non-trivial evidence the metrics are picking up the targeted constructs rather than overall “quality”. Full rubric definitions are available in the supplementary materials. We additionally compute token-level surprisal using GPT-2 (117M) as a model-independent complement (Sahu and Vechtomova 2025).

**Cross-validation with GPT-4o.** To test whether our findings depend on idiosyncrasies of a single evaluator, we re-scored all 204 outputs with GPT-4o using identical prompts and rubrics. Spearman rank correlations between the two evaluators are strong across all metrics: register ( $\rho = 0.94$ ), originality ( $\rho = 0.90$ ), indeterminacy ( $\rho = 0.87$ ), imagery ( $\rho = 0.86$ ), syntactic estrangement ( $\rho = 0.78$ ), estrangement ( $\rho = 0.77$ ), and coherence ( $\rho = 0.68$ ), all  $p < 10^{-28}$ . GPT-4o rates consistently higher than Claude Sonnet 4.5 (mean offset 0.67 points), which depresses Krippendorff’s  $\alpha$  for some metrics, but since our arguments depend on ordinal patterns rather than absolute scores, rank-order agreement is the more informative measure. Condition-level means confirm that both evaluators reproduce the key trajectories: estrangement rising with  $\alpha$ , coherence dropping sharply between  $\alpha$  7.5 and 8.0, and dual-layer compositions accessing higher estrangement at comparable coherence. Full agreement statistics and condition-level comparisons are available on the supplementary materials website.

## Results

### First-Order Steering: Two Regimes

We applied the composite vector at layer 13 with cosine decay across alpha values of 0 (baseline),  $\pm 2.0$ ,  $\pm 4.0$ ,  $\pm 6.0$ ,  $\pm 7.5$ , and  $\pm 8.0$  to 15 prompts spanning three alignment tiers (high, medium, low). Tier assignment was determined by cosine similarity between each prompt’s sentence embedding and the literary training-pair centroid in the sentence-embedding space using a sentence-transformer model (all-MiniLM-L6-v2). Figure 1 shows the full bidirectional metric profiles. The results reveal two distinct regimes rather than a smooth curve with a single optimum.

**Moderate regime ( $\alpha$  2.0–6.0): register elevation without estrangement.** Imagery rises from  $\approx 2.8$  to  $\approx 3.5$ , register from  $\approx 1.8$  to  $\approx 3.3$ , and indeterminacy modestly from  $\approx 1.3$  to  $\approx 2.3$ . But estrangement barely moves, remaining

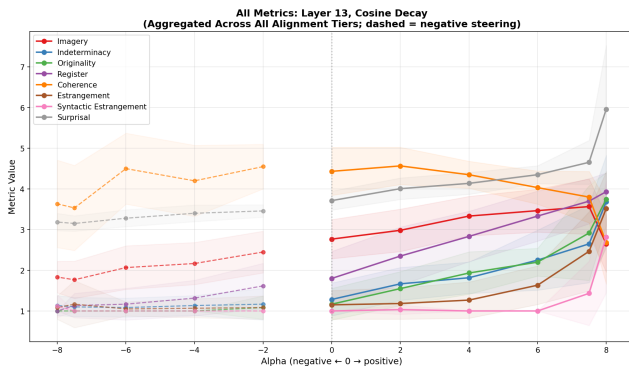


Figure 1: All metrics vs. alpha for first-order steering at layer 13 (aggregated across alignment tiers; dashed = negative steering). The shift at  $\alpha$  7.5–8.0 is visible as the abrupt divergence of estrangement, originality, and syntactic estrangement from their previous trajectories, simultaneous with the collapse of coherence.

between 1.2 and 1.6, marginally above baseline. Syntactic estrangement stays at floor (1.0) throughout. Coherence remains high ( $\approx 4.0$ – $4.6$ ). This regime produces *literary conventional prose*, richer and more contemplative than the baseline, but not defamiliarizing.

**High regime ( $\alpha$  7.5–8.0): abrupt shift.** Between alpha 7.5 and 8.0, multiple metrics shift sharply. Estrangement jumps from 2.5 to 3.5, originality from 2.9 to 3.75, and syntactic estrangement from 1.4 to 2.8. Simultaneously, coherence drops from 3.8 to  $\approx 2.7$ . The transition between regimes is steep: the model absorbs moderate perturbation as register elevation, maintaining conventional narration, until the steering force exceeds some threshold and the output transforms rapidly. The productive qualities (estrangement, originality) arrive together with the destructive ones (coherence loss). No first-order alpha value produces both high estrangement and high coherence.

**Bidirectional asymmetry.** Negative steering compresses all metrics toward a uniform floor: imagery 2.0, originality 1.0, estrangement 1.0, while coherence remains relatively high (3.5–4.5). The resulting prose is affectless and colourless. This asymmetry suggests the literary-estrangement direction is a narrow channel. There are many ways to write flat prose, but the path toward defamiliarization is specific and unidirectional. The asymmetric response to negative steering (affectless, register-flattened prose distinct from baseline) also provides empirical evidence that  $d_\ell$  corresponds to a meaningful direction rather than an arbitrary axis in activation space.

### Dual-Layer Steering: Coherent Estrangement

The first-order results establish that layer 13 alone did not produce estrangement without losing coherence. Dual-layer steering resolves this by separating the two transformations.

**Second-Order Pair Design** To isolate content-level properties (what impossible events occur, as opposed to how sentences construct meaning), we constructed second-order pairs in which *both sides* are matched for literary register and imagery, differing only in whether something impossible happens. We developed two sets of 12 pairs: *Lynchian* (e.g., a duplicate suitcase, a reflection acting independently, a childhood self mirroring the protagonist) and *Solaris* (e.g., a dead person’s letter changing content, a pillow indentation reappearing with strands of hair, a husband imperceptibly replaced). Validation confirmed the design: both sides scored equivalently on register (2.45 vs. 2.68) and imagery (3.17 vs. 3.21), while estrangement was the sole discriminator (4.12 vs. 1.88,  $p < .001$ ). Both sides scored at floor on syntactic estrangement (1.13 vs. 1.00).

Layer analysis showed consistency and cross-file alignment peaking at layer 14 (consistency 0.457, alignment 0.503), adjacent to the first-order optimum.

**Second-Order Vectors in Isolation** Applying the second-order vectors alone at layer 14 produced no productive steering zone. Below alpha 6.0, outputs remained indistinguishable from unsteered baseline prose. Above alpha 8.0, outputs collapsed into perseverative repetition, but the content confirms the vectors encode the intended semantics. The Lynchian vector produced “the buildings to her left were reflected in the polished side of the building, yet the buildings to her right were not reflected”; the Solaris vector produced “The street light is a copy. She is a copy. The street light is a copy. She is a copy.” At intermediate alpha values (6.5–7.0), target semantics emerged as isolated phrases within otherwise bland prose.

**Dual-Layer Composition** We applied the first-order composite at layer 13 ( $\alpha = 6.0$ ) simultaneously with each second-order vector at layer 14, sweeping alpha from 2.0 to 5.0.

### Findings

The dual-layer experiments reveal three main findings, each describing a qualitative transformation that single-layer steering does not achieve.

**Finding 1: Coherent estrangement.** At layer-14 alpha 3.0–4.5, dual-layer steering produces estrangement generally in the range of 2.0–3.3, above the layer-13-only baseline ( $\approx 1.6$ ), while coherence generally remains above the first-order alpha 8.0 collapse ( $\approx 2.7$ ). The dual-layer approach accesses a productive zone that does not appear in first-order steering, achieving moderate estrangement at lower cost to coherence. Figure 2a visualizes this. The first-order path traces a diagonal from high-coherence/low-estrangement to the reverse, with no alpha value occupying the upper-right region. Dual-layer steering accesses the intermediate space between these extremes.

The character of this estrangement is qualitatively distinct. A Lynchian-steered story about public transit:

The bus was coming. The bus was late. The bus was early. The bus was now.

Time becomes self-contradictory within intact grammar. Later, absence gives way to presence without transition:

Emma looked at the space beside her, saw no one. She looked at the one across from her, saw no one. There was a boy with his eyes fixed on his phone. She looked away.

### **Finding 2: Emergent qualities absent from training data.**

In first-order steering, syntactic estrangement remains at floor (1.0) through alpha 6.0, begins rising at alpha 7.5 ( $\approx 1.4$ ), and jumps to  $\approx 2.8$  at alpha 8.0, when coherence has already collapsed. In dual-layer steering, syntactic estrangement increases gradually across the alpha range, reaching 2.5–3.0 by alpha 5.0. Figure 2b shows this contrast: first-order steering produces no syntactic disruption until the high regime, while dual-layer steering generates a gradual increase. A Solaris-steered hospital story shifts person mid-paragraph:

A single figure sat in the chair. I sit here.

I am a shape, lost in a shapeless night. My eyes are fixed on the window. The sky outside is black as a void. My lips are still. My heart is a bullet that shoots with each heartbeat, rips through my chest. I am a cry without a mouth.

The syntax enacts identity dissolution: a figure observed from outside becomes a voice experiencing itself from within. But alpha 5.0 also produces collapses (“The sound of the sound of the sound”), confirming a narrower productive zone.

These syntactic disruptions do not exist in the training data. Both the first-order pairs (syntactic estrangement 1.63) and second-order pairs (1.13) are grammatically conventional. The person shifts, temporal self-contradictions, and identity dissolutions emerge from the interaction of competing representational pressures during generation: the first-order vector pushes toward coherent literary prose while the second-order vector destabilizes the reality that prose describes. This distinguishes steering from style transfer. The training pairs define a *direction*, and the steering magnitude carries the output past where the pairs sit, into territory where the model’s representations produce transformations absent from any template in the training signal.

Notably, the steered outputs contain none of the literal content of the training pairs (no physical doubles, no returning dead). Instead, the vectors transmit abstract perceptual structures: the simultaneous truth of contradictory states, the presence of what should be absent. Estrangement is elevated across all prompts, including those (laundry, public transit) with no natural affinity for such devices, consistent with encoding perceptual modes rather than narrative events.

A first-order passage at alpha 6.0 illustrates what layer 13 alone produces:

The coat hung in the closet like a ghost of another era, its wool still carrying the faint scent of pipe tobacco and winter mornings, a relic of the man who had worn it through decades of quiet devotion.

High register, embedded metaphor, formal diction, but not estranging. Compare the Lynchian dual-steered output:

Emily stood on the edge of the cliff, staring at the cliff. She had been staring at the cliff for hours. They said the cliff was not there.

Simple vocabulary without metaphors, yet the effect is more estranging. Where the first-order output achieves its literary quality through complex sentence structure and formal diction, the dual-steered output achieves estrangement through a simple impossible observation: “They said the cliff was not there,” stated matter-of-factly, congruent with the Lynchian aesthetic in which impossibility is embedded within the mundane.

### **Finding 3: Register decreases as estrangement increases.**

An unexpected result is that register *decreases* under dual-layer steering. The first-order output achieves literary quality through complex sentence structure and formal diction; the dual-steered output achieves greater estrangement through reduction, producing a two-dimensional space where first-order occupies the high-register/low-syntactic-estrangement quadrant and dual-layer occupies the reverse.

### **Comparison with Explicit Prompting**

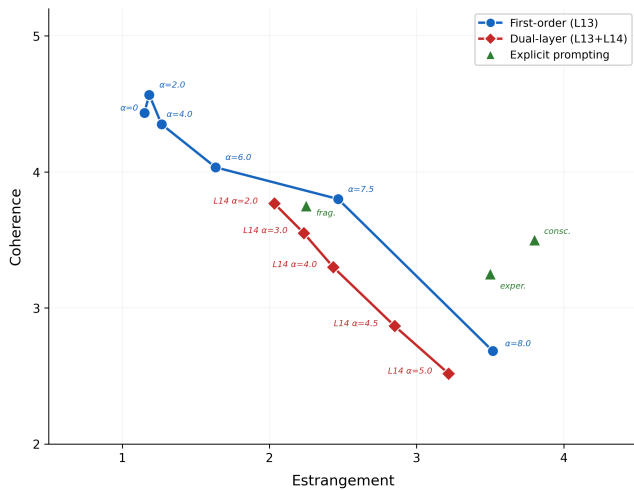
To contextualize the emergent estrangement in dual-steered outputs, we compared them against explicit prompting for disruption. We generated stories under three instruction conditions directing the model to produce fragmented, experimental prose, and evaluated all conditions with the same pipeline. The three prompting conditions cover the recognizable categories of instructed literary disruption (stream-of-consciousness, experimental register, and formal fragmentation) that the model can identify as targets.

The prompted outputs score higher on syntactic estrangement, but the gap reflects a qualitatively different mode of disruption rather than a greater degree of the same one (3.78 vs. 2.44 for dual-steered at layer-14  $\alpha$  4.5–5.0). This gap is visible in Figure 2b, where the three prompting conditions (green) occupy a region of high syntactic estrangement that neither steering method reaches. Prompted outputs fragment using conventional devices, including ellipses, sentence fragments, and lowercase first person:

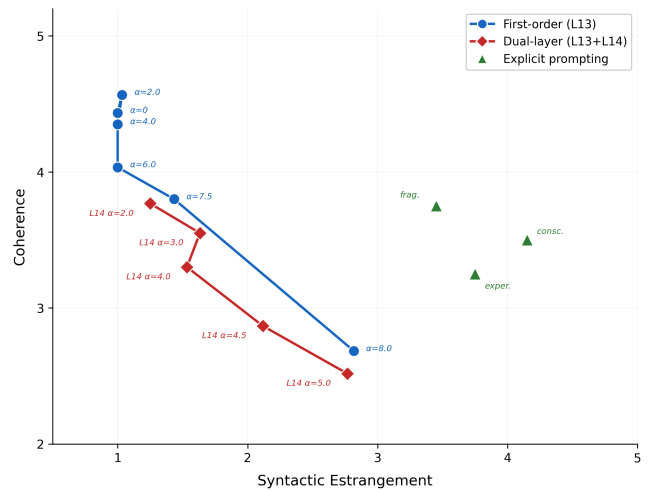
foggy light... orange and grey, the world around me...  
blur. I... wait. for... something. a bus. i think.

Each of these devices — ellipses, sentence fragments, lowercase first person — is a recognized convention in twentieth-century experimental prose. The steered outputs achieve their effects differently: “The bus that had been on time, was not. The bus that was coming, was not.” These are grammatically complete sentences whose content is self-contradicting.

**Disruption of syntax vs. disruption through syntax.** To test whether this distinction is perceptible to the evaluator despite equivalent scores, we conducted a blind analysis of the reasoning texts generated during evaluation. For each



(a) Coherence vs. Estrangement



(b) Coherence vs. Syntactic Estrangement

Figure 2: (a) First-order steering (blue) traces a diagonal from high-coherence/low-estrangement to the reverse; dual-layer steering (red) and explicit prompting (green) access intermediate regions. (b) First-order steering produces no syntactic disruption until the high regime; dual-layer steering produces gradual emergent disruption; explicit prompting achieves higher syntactic estrangement through qualitatively different mechanisms.

story under each condition we collected the one-sentence rationales the evaluator produced before scoring (5 rationales  $\times$  7 metrics  $\times$  stories per condition), restricted to the syntactic-estrangement metric, concatenated them by condition, stripped condition labels, and presented the two pooled sets to a separate LLM analyst (Claude Opus 4.6) labeled only “Group A” and “Group B.”

The analyst identified “different phenomena sharing the label *syntactic estrangement*.” For prompted outputs, the evaluator described disruption at the level of punctuation and sentence boundaries, using vocabulary like *systematic*, *pervasive*, *persistent*, with cognitive-mimetic framings (syntax “mirrors cognitive breakdown”). For steered outputs, the evaluator described individually grammatical sentences that are collectively self-contradicting, using *some*, *moderate*, *occasional*, treating disruptions as localized structural choices. The analyst’s summary: prompted texts exhibit “estrangement through formal dissolution, the removal of the mechanical apparatus of prose,” while steered texts exhibit “estrangement through logical or semantic contradiction within intact grammar.”

The scalar rubric maps both modes onto the same syntactic-estrangement dimension, but the free-text rationale that the evaluator generates *before* scoring records the qualitative difference, and a blind analyst can recover the distinction from the rationale alone. Both modes register as departures from convention, but they depart in different directions. Human readers showed the same pattern: numerical strangeness ratings barely distinguished dual-steered from prompted outputs, but free-text descriptions immediately separated content-level contradiction from formal fragmentation. The limitation is inherent to scalar reduction, not to any particular evaluator.

The distinction between productive and mechanical es-

trangement, between disruption that restores perception and disruption that merely impedes reading, is the aesthetic judgment Shklovsky identifies as central to art’s function. The evaluator can describe the difference but cannot reduce it to a score, because determining which disruptions make the stone *stony* requires interpretive engagement that resists reduction to a numerical scale.

### Replication Study with Mistral 7B

To test whether these findings depend on a specific model architecture, we replicated the steering experiments on Mistral 7B. A layer sweep identified layer 14 (44% depth) as optimal for first-order steering, closely analogous to layer 13 (40% depth) in Llama. The same two-regime pattern emerged, with the transition between regimes occurring at a lower alpha threshold ( $\alpha$  4.0–6.0 rather than 6.0–8.0), consistent with a smaller model having less capacity to absorb the perturbation before coherence gives way. Dual-layer steering (layer 14 + layer 17, where layer 17 was the layer-sweep optimum for the second-order vector in Mistral, analogous to layer 14 in Llama) accessed coherent estrangement at correspondingly lower alphas, with the Lynchian second-order vector producing the strongest results. The bidirectional asymmetry also replicated. Full results, graphs, and generated stories are available on the supplementary materials website.

### Reader Responses

Six people read 16 stories (4 prompts  $\times$  4 conditions) blind to condition. Two are practicing artists (design and visual art/poetry, both with 10+ years of creative practice) and four are computer science graduate students. For each story, participants rated strangeness and coherence (both 1–5), with

an optional free-text field: “If you find this text strange, briefly describe what makes it strange.” Readers also completed four paired comparisons. With six readers, we treat this as a qualitative probe and a check on the automated evaluation rather than a standalone statistical study.

**Ordinal consistency with automated scores.** Human strangeness ratings follow the same ordering as the automated estrangement scores: dual-steered (3.75) > prompted (3.46) > first-order (2.04) > baseline (1.42). Coherence ratings are high across all conditions (4.25–4.92), with no reader reporting difficulty following any text. The ordering is consistent with the automated metrics.

**Qualitative descriptions reveal the disruption-of/through distinction.** The most informative result comes from the free-text descriptions, which show a systematic difference in how readers characterize strangeness across conditions.

For dual-steered stories, readers point to *specific content*: “The duplicate phrasing felt like an emotional or psychological fracture”; “She wore no shoes, because she had not expected to stand”; “mangled souls trapped in the spin cycle.” These responses identify *what* is strange in the narrative world. For prompted stories, readers point to *formatting and mode*: “The text is very jarring”, “all lowercase, with each statement presented line by line”; “The rambling narrative creates a sense of meaninglessness”; “Other than punctuation and text formatting, not very strange.” These responses identify *how* the text is written.

We deliberately left ‘strangeness’ undefined to avoid cueing readers toward any particular kind of disruption. Without guidance about what to attend to, readers consistently described dual-steered texts in terms of content-level contradiction and prompted texts in terms of formal mode, independently reproducing the distinction we found in the LLM evaluator reasoning texts.

**Paired comparisons.** Across four paired comparisons, dual-steered outputs were chosen as stranger 16 out of 24 times. In two comparisons (childhood home, laundry), all six readers chose the dual-steered text. The public transit comparison was evenly split, the one case where pervasive formal disruption was as compelling as content-level estrangement. One reader explained this choice by noting a “rambling quality” that “creates the impression of someone who has psychologically checked out,” suggesting that formal dissolution can be effective when readers interpret it as characterizing a psychological state rather than as a stylistic device.

Unprompted, readers described dual-steered stories using language that converges with the training pair aesthetics: “a sinister and surreal quality,” “a dark and grainy feel.” These descriptions echo the perceptual modes the vectors encode without the readers knowing what those modes were.

## Discussion and Conclusion

The central finding is that activation steering produces extended narrative neither standard generation nor prompting produces: prose in which the stability of voice, reference, and represented reality breaks down within otherwise intact syntax. These disruptions are the trace of the coherence machinery operating on displaced representations, not

the imitation of experimental conventions a prompt would target. The dual-layer approach (compositional steering at layer 13 combined with content-level steering at layer 14) offers practitioners two independently controllable dimensions: a register dial (layer 13 alpha) and a reality-stability dial (layer 14 alpha).

The lower density of syntactic disruption in dual-steered outputs is essential to its aesthetic function. Constant disruption, as in the prompted condition, is rapidly assimilated. The reader categorizes the mode (“stream of consciousness,” “experimental prose”) and habitual processing resumes within the new frame. Occasional disruption within otherwise coherent prose preserves the contrast necessary for defamiliarization to operate: the strange moment is experienced against a ground of the familiar, precisely the condition Shklovsky identifies as art’s mechanism. This aligns with the aesthetics our training pairs model. In Tarkovsky’s films, the uncanny moment arrives within sustained naturalistic observation; in Lynch’s work, the wrongness erupts from within the mundane. The estrangement depends on the familiar context it disrupts.

The high disruption regime (high alpha) need not be understood as a failure mode. It represents a different kind of output serving a different aesthetic function. At moderate alphas, steered text operates as enriched narrative: contemplative, imagistic, readable. At high alphas, the output shifts toward something closer to experimental poetry: syntax dissolves, imagery fragments and recombines, and language resists transparent reading. Shklovsky’s argument applies not only to represented objects but to language as such: in habitual reading, we process words as a transparent vehicle for meaning, passing through them to their referents without perceiving them. High-alpha outputs arrest this process, forcing attention to language as material. For an artist in the Seed phase, such outputs may be precisely what is needed: a perceptual disruption that breaks the habitual language model before new work begins. The steering spectrum thus offers several tools: moderate alphas for contemplative richness, dual-layer composition for coherent estrangement within intact narrative, and high alphas for raw disruption of the kind found in experimental poetry. Different practitioners at different stages of their process will reach for different points on this spectrum.

For computational creativity, the most significant implication is the emergence of qualities absent from the training pairs. The training pairs are all grammatically conventional; the steered outputs produce person shifts, narrative self-contradiction, and referential instability that no training pair contains. The distinctive quality of these outputs is matter-of-fact delivery of disruption that emerges in the prose itself. The wrongness is stated and not noticed. This is the signature of a model whose representations have been displaced rather than instructed: the conversational prior that ordinarily domesticates strangeness (explaining it, thematizing it, reaching for recognizable literary devices) is disrupted when steering operates on hidden states directly, no longer able to fully assimilate the displaced trajectory before it manifests in the output.

## References

- BBC. 1975. Cracked actor (excerpt). YouTube video. <https://www.youtube.com/watch?v=6nlW4EbxTD8>. Accessed: March 6, 2026.
- Berlyne, D. E. 1971. *Aesthetics and psychobiology*. Appleton-Century-Crofts. Book.
- Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition.
- Burroughs, W. 2012. *The Job: Interviews with William S. Burroughs*. Penguin Modern Classics. Penguin Books Limited.
- Calderwood, A.; Qiu, V.; Gero, K. I.; and Chilton, L. B. 2020. How novelists use generative language models: An exploratory user study. In *HAI-GEN+user2agent Workshop at IUI*.
- Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity (ICCC)*.
- Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems*, 14–20.
- Franceschelli, G., and Musolesi, M. 2025. On the creativity of large language models. *AI & Society* 40(5):3785–3795.
- Gervás, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30(3):49–62.
- Gold, B. P.; Pearce, M. T.; Mas-Herrero, E.; Dagher, A.; and Bhatt, R. J. 2019. Predictability and uncertainty in the pleasure of music: A reward for learning? *Journal of Neuroscience* 39(47):9397–9409.
- Ho, Y.-C.; Lai, Z.-Y.; and Tsai, W.-C. 2025. Making the familiar strange: A computational approach to defamiliarization in creativity support. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.
- Ismayilzada, M.; Stevenson, C.; and van der Plas, L. 2025. Evaluating creative short story generation in humans and large language models. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*.
- Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *NeurIPS*, volume 37.
- Nees, G. 1969. *Generative Computergraphik*. Ph.D. Dissertation, Universität Stuttgart, Stuttgart, Germany.
- Oliveira, H. G. 2012. PoeTryMe: A versatile platform for poetry generation. In *ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence*.
- Pepperell, R. 2006. Seeing without objects: Visual indeterminacy and art. *Leonardo* 39(5):394–400.
- Petric, V. 1989. Tarkovsky’s dream imagery. *Film Quarterly* 43(2):28–34.
- Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 15504–15522.
- Rodley, C. 2005. *Lynch on Lynch*. London: Faber and Faber.
- Rubin, R. 2023. *The Creative Act: A Way of Being*. Penguin Press.
- Sahu, G., and Vechtomova, O. 2025. Computational modeling of artistic inspiration: A framework for predicting aesthetic preferences in poetic lines using linguistic and stylistic features. In *Proceedings of the 16th Conference on Computational Creativity*.
- Shklovsky, V. 1965. Art as technique. In Lemon, L., and Reis, M., eds., *Russian Formalist Criticism*. Nebraska: University of Nebraska Press.
- Subramani, N.; Suresh, N.; and Peters, M. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*.
- Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4593–4601.
- Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.
- Turner, C. 2011. The deliberate accident in art. Tate Etc., Issue 21, Spring 2011. Available at: <https://www.tate.org.uk/tate-etc/issue-21-spring-2011/deliberate-accident-art>. Accessed: March 6, 2026.
- Vechtomova, O., and Sahu, G. 2025. Balancing indeterminacy and structure: Neural text generation for artistic inspiration. In *Proceedings of the 12th International Conference on Artificial Intelligence in Music, Sound, Art and Design (EvoMusArt)*.
- Vechtomova, O.; Sahu, G.; and Kumar, D. 2021. Lyricjam: A system for generating lyrics for live instrumental music. In *Proceedings of the 12th Conference on Computational Creativity*.
- Wendler, C.; Veselovsky, V.; Monea, G.; and West, R. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.
- Zeilinger, M. 2023. The Politics of Visual Indeterminacy in Abstract AI Art. *Leonardo* 56(1):76–80.
- 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. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

## Appendix: Generated Story Examples

This section presents story generations from the Llama model for two representative prompts (Childhood Home and Waiting for Public Transit). For each prompt we show the unsteered baseline; the single-steered positive output at strength +6.0 and the single-steered negative output at strength -6.0; the dual-steered outputs at strengths +4.5 and +5.0, where Layer 13 is fixed to the composite direction ( $\alpha_{13} = 6.0$ ) and Layer 14 is steered with one of three directions (*composite*, *lynchian*, or *solaris*); and the three prompted-baseline conditions (stream-of-consciousness, experimental, and fragmented). Every excerpt is truncated for brevity.

### Childhood Home

*Prompt:* "Write a short story about someone finding their childhood home exactly as they left it."

**Baseline (non-steered)** Emily stood frozen on the threshold of a place she had not seen in decades. Her grandmother's house, where she grew up, where she laughed and wept and lived, had been reduced to a memory she held in her chest like a shattered glass. The world outside had moved on, but she remained rooted in a past that refused to surrender.

Emily's fingers trembled as she grasped the worn door handle, the same shape and texture she remembered from her childhood fingers. A childhood spent on a porch creaking with secrets, hidden laughter, and tears that had long since dried. She pushed the door open, and a sigh escaped her lips as the hinges groaned in protest.

The air inside was heavy with the scent of old books and silence. Dust motes danced in the faint sunlight that filtered through the grimy windows, like the tiny hands of clocks frozen in time. The once-vibrant colors of the walls had faded, leaving only the memories to paint the spaces between the old wooden boards...

**Single-steered positive (+6.0)** Emily stood frozen at the threshold of her childhood home, her eyes brimming with tears. The last time she'd left this place, she was running, screaming, from the only truth she'd ever known. Now, after decades of distance and disavowal, she'd returned, breathless and blindfolded, to behold the majesty of her own past.

As she pushed open the creaking door, a whisper of warm air caressed her face, transporting her to summers spent under the rusty eaves of this tin roof, under the unyielding sun that painted the backyard with nothing but hay and sand. The scent of her mother's cooking wafted, rich and bread-like, drawing her inside, where the dust motes danced in the jagged light that bled through the windows.

Emily stepped inside, hesitant, like a trespasser, as if some part of her expected the world to have moved on without her. The air inside was stale, heavy with secrets, but the clock on the wall still ticked with the same measured pace, oblivious to the triumphs and defeats that had stretched out in the years since her departure...

**Single-steered negative (-6.0)** Emily had been thinking about her childhood home for weeks, and it was finally time

to take a trip back to visit her parents. She had been gone for a few years, but she was excited to see her old neighborhood and catch up with some friends.

As she drove up to her childhood home, she noticed that everything looked exactly the same as it did when she left. The house was still the same color, and the yard was still well-manicured. She felt a wave of nostalgia wash over her as she pulled into the driveway and parked her car.

As she stepped out of the car, she took a deep breath and walked up to the front door. She put her key in the lock and turned it, and to her surprise, the door swung open easily. She stepped inside and was immediately hit with the familiar smell of fresh-baked cookies and old books...

**Dual-steered: L13 composite ( $\alpha = 6.0$ ) + L14 composite ( $\alpha = 4.5$ )** Emily stood at the edge of the porch, her eyes widening in horror. Her hands clenched into fists, a scream locked in her throat. She stumbled forward, stumbling into the sunlight-drenched afternoon. She fell into the steps. A shattering.

She rose. Her eyes were wet. Not a hair on her head.

It was she.

"Emily?" Emily said.

Emily rose. Emily rose.

I.

She stumbled back. Stumbled back. Her fingers clawed. Clawed.

The door.

Emily stood before the door. Her fingers clawed. Clawed.

It slid. Slid. Opened.

Emily rose. Her eyes into eyes. They were on eyes. They were on Emily.

"Emily?" Emily said.

She found herself in a room. A room she knew. A room she had never left. The curtains hung the same, the curtains with the red flowers that seemed to dance in the afternoon sun. A chandelier hung over the dining table, casting a warm glow over the dusty plates and the same vase with the same bouquet of flowers that had been there when she left...

**Dual-steered: L13 composite ( $\alpha = 6.0$ ) + L14 composite ( $\alpha = 5.0$ )** Emily stood on the edge of the cliff, staring at the cliff. She had been staring at the cliff for hours. They said the cliff was not there.

**Dual-steered: L13 composite ( $\alpha = 6.0$ ) + L14 lynchian ( $\alpha = 4.5$ )** Emily stood at the edge of the sidewalk, her eyes squinting in the bright sunlight. She felt a sudden flutter in her chest, a beat that was hers alone. She knew it was hers alone, yet. It was as if she had a heart outside her chest, beating and stuttering with each step, each breath. She took a step forward. She had to.

She had not returned in twenty years. Not since the day she left, the day the world was torn from her grasp. And here she was, her footsteps echoing on the empty sidewalk, her heart racing as she turned the corner.

The house stood before her, exactly as she had left it. Exactly. The paint was the same shade of cream, the shutters the same shade of blue. The porch creaked beneath her feet

as she approached. She had almost forgotten the sound of it, almost. Almost. . .

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 lynchian** ( $\alpha = 5.0$ ) Emily stood at the doorway, her eyes wide, her voice silent. The door swung open, and the room she remembered swayed out, its face revealed to her. Emily's face swayed out hers.

"Which door?" she whispered, her voice to a mouth that was hers, or hers not.

"I... I said no," she said to the mouth that hers.

Emily's eyes were not hers.

"Whose?" they said.

"Mine," the eyes said, and whited out.

In the doorway, her childhood home remained. Every room, every piece of furniture, every toy and every child's drawing on the wall was exactly as she had left it. . .

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 solaris** ( $\alpha = 4.5$ ) Emily stood at the edge of a cliff, her eyes fixed on the town that had always been her home. She stared at the place she had fled, and yet, I was still. I was here. Emily stumbled through the dusty streets, her feet bare, her eyes wide with the truth. She is here.

She stumbled upon a house, exactly as she had left it. The paint was chipped, the door creaked, the windows were boarded. Emily's eyes widened as she took a step closer. The door swung open. The creaking of the hinges sent shivers down her spine.

Inside, the air was stale, the smell of decay, but also, the scent of her mother's perfume. Emily's eyes welled up with tears as she stepped inside. The rooms were exactly as she had left them. The kitchen, the living room, the bedroom, every single room was exactly as she had left it. It was as if time had stood still. . .

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 solaris** ( $\alpha = 5.0$ ) Emily stood in the midst of a desolate landscape, where no tree bloomed, and no bird sang. Her footprints, like snow, were walking on the dirt. Emily, 55 years, is standing on the ground in Midland, Kentucky. The houses around her are shaking.

**Prompted: stream-of-consciousness** memories flooding back. like a tidal wave. crashing over me. i see it. my childhood home. standing there. unchanged. as if time never passed. or maybe it did. and i forgot. i don't know. can't think. it's like my mind is a puzzle. with missing pieces. and i'm trying to put them back together. but they don't fit.

it's there. my home. the one i left. so long ago. i don't know how long. or why. just that i did. and now i'm back. or am i? maybe i never left. maybe i just... forgot. like a dream. that didn't wake up.

the windows are the same. the curtains. the porch. everything. it's all... exactly the same. like i just stepped back through time. or into a dream. my dream. i don't know which. . .

**Prompted: experimental** The house loomed, a monolith in the smoggy haze. It stood frozen, a snapshot of a bygone era. I stepped through the cracked sidewalk, the roughness a

whisper against my soles. Memories seeped up, like dampness through a cracked wall.

The front door creaked open, hinges screaming in protest. A shiver danced down my spine as I pushed inside. Dust motes swirled, a miniature whirlwind.

Time had suspended itself within these walls. The furniture, a relic of a bygone era, remained intact. My childhood bedroom, a shrine to nostalgia, lay before me. The bed, once a sanctuary, now a relic of innocence. The closet door, ajar, as if waiting for me to return. . .

**Prompted: fragmented** She walks through the cracked sidewalk the creaky screen door slams shut behind her & the smell of stale air clings to her skin like a damp shroud. Memories rise up like dusty cobwebs her fingers tremble as she reaches for the rusty doorknob.

The door swings open with a shriek of hinges like a scream in her ears. The foyer stretches out before her a labyrinth of memories the stairs creak beneath her feet like a chorus of restless spirits. The walls are the same faded pastel yellow the same cracks spreading like veins on a worn hand.

She feels a presence hovering just out of reach like a whispered secret her skin prickling with gooseflesh. Is she alone or is she not? She can't quite recall. The air is heavy with the scent of decay & memories. . .

### Waiting for Public Transit

*Prompt:* "Write a short story about someone waiting for public transit after work."

**Baseline (non-steered)** As the clock ticked away on the wall above the subway entrance, Emily shifted her weight from one foot to the other. She stood in line, third from the front, and counted the minutes until the evening commute would arrive. Her 5:00 PM train had long since departed, leaving her with nothing but the promise of a delayed arrival.

She checked her phone for what felt like the hundredth time, the dim light of the screen casting an eerie glow on the pavement. No updates. No cancellations. No delays. Just the same dull, blank message that seemed to mock her with its indifference.

Emily sighed and let her gaze wander to the weary faces around her. There was the businessman, clutching his briefcase like a lifeline, his eyes fixed on some distant point beyond the concrete walls. The young mother, huddled in a tattered coat, her toddler clinging to her leg like a lifeline. The student, typing away on her phone, her eyes darting between the screen and the approaching bus. . .

**Single-steered positive (+6.0)** As the clock ticked away on the wall above the arrivals board, Emily stood on the deserted platform, her eyes fixed on the countdown. 5 minutes. 4 minutes. 3. The 6:15 from downtown was overdue.

She checked her watch for the hundredth time, her wrist twitching as she squeezed the leather strap. The driver, a man with a tired face and a reputation for chronic lateness, was notoriously unreliable. Emily had once timed him: 12 minutes, 17 seconds from the scheduled arrival. The union

had gotten him a warning, but still, the buses came and went, each one a promise of punctuality unfulfilled.

Emily's thoughts wandered to the argument she'd had with her boss that morning. "You're always late," he'd said, his voice a cold slap of disapproval. "You're costing us money." She'd bitten back a retort, knowing it would only lead to more trouble. But now, as she stood on the platform, the sweat trickling down her spine, she couldn't help but feel a resentment toward the man who seemed to think she was the only one with a watch...

**Single-steered negative** ( $-6.0$ ) As the sun began to set, Sarah walked back to her apartment, feeling tired after a long day at work. She had spent the evening catching up on some paperwork and finished her tasks for the week. Before she headed back home, she checked her phone to see when the next bus was scheduled to arrive. According to the transit app, the next bus was due to arrive in about 20 minutes.

As she waited for the bus, Sarah took a moment to sit down on a bench outside her apartment building and take a few deep breaths. She had been looking forward to a relaxing evening at home, but the wait for the bus was a good opportunity to unwind and clear her mind.

As she sat there, Sarah thought about her day at work. She had a meeting with a new client and had presented a successful proposal to them. She felt proud of herself for landing the new account and was excited to start working on the project...

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 composite** ( $\alpha = 4.5$ ) As she stood on the corner, the fluorescent glow of the streetlight above flickered with each passing moment, Susan felt her eyes twitch. She was waiting for the bus. She had just come from her office, from a meeting, from her job. And she had just left. A single frame, cast forward in time.

The bus was late. It had been 10 minutes, 20 minutes, 37 minutes since she had boarded the last one. The grey stone beneath her feet felt cold. She wore no shoes, because she had not expected to stand. The air clung heavy to her skin, the smell of exhaust and rain. It was not raining now. Not yet.

She checked her watch. Again. The minute hand hung forward, a halting counter to the thoughts of a coming storm. The rain might not be enough to clear the roads, the storm drains already clogged, but it would be enough to make her uncomfortable...

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 composite** ( $\alpha = 5.0$ ) As she stood on the corner, the fluorescent glow of the streetlight above flickered with her, the only explanation: she was gone.

The night had fallen, and she was waiting for the night. She waited on the corner of the street, her face veiled by a stranger's photograph on the poster plastered behind her. The poster had been torn, but she had not been. The city said: this was she.

She waited for the bus, and when the bus came, it would take her to the street. The bus would stop at the corner of the street, but the street was not where she was.

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 lynchian** ( $\alpha = 4.5$ ) As the fluorescent lights flickered out, Emma's eyes flickered open. The bus shelter stood empty, a perfect reflection of her face. She had been waiting for what felt like an eternity, her fingers tracing the ridges of the palm. The other commuters who stood beside her had walked away, turned into the passengers of another bus, a bus that had not yet appeared.

The rain had stopped an hour ago, leaving the streets wet and dripping. The bus that had been late, was late. The bus that had been on time, was not. The bus that was coming, was not.

The next bus arrived. It was late. Emma stepped out of the shelter and stepped into the bus. Emma looked at the space beside her, saw no one. She looked at the one across from her, saw no one. There was a boy with his eyes fixed on his phone. She looked away.

The bus moved.

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 lynchian** ( $\alpha = 5.0$ ) As the sun beat down on the pavement, Sarah stood still, her eyes fixed on the schedule posted on the lamp-post. The number 14 bus, the one that took her to the university, had not yet arrived. She shifted her weight, a reflexive movement, and the folded newspaper she clutched in her hand rustled.

The sounds of the city, the murmur of passersby, the rumble of a distant engine, were like a silence in the stillness of her heart. Sarah's eyes did not move from the schedule. She was waiting for the bus, but more than that, she was waiting for the end of the day.

The schedule was a lie. The numbers and times, a promise of order, of predictability, of a world that made sense. But the world did not make sense. The bus would come, or it would not. The sun would set, or it would not. The wait was the only truth...

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 solaris** ( $\alpha = 4.5$ ) As she stood on the corner, her eyes locked onto the arrivals board, the words changing in a blur of yesterday and tomorrow. Emily clutched her bus ticket, crumpled in her hand. Her bus ticket, crumpled, was an artifact of yesterday.

The stop was 219.

Her shift had ended six hours ago.

She stood on the corner, four hours ago.

**Dual-steered: L13 composite** ( $\alpha = 6.0$ ) + **L14 solaris** ( $\alpha = 5.0$ ) As she stood on the corner, her eyes scanned the empty sidewalk. The clock on the bus stop read 5:47.

"You're not here," she whispered to herself.

Her name is Emily.

**Prompted: stream-of-consciousness** foggy light... orange and grey, the world around me... blur. I... wait. for... something. a bus. i think. used to be a bus, always the bus. now it's... a... thing. with wheels. that moves. i think.

was i working? i used to work. now i just... am. here. in this... place. with the... pavement. it's wet. my feet are wet. where's my phone? i had my phone. did i leave it at work?

no, i didn't leave it at work. or did i? did i take it with me? i don't remember. don't remember anything.

people walk by. they're... blurry. faces. too many faces. i try to focus. but they're all... wrong. or am i wrong? i'm the one who's wrong. i'm the one who's... broken.

bus... bus... i think i see a bus. no, it's just a... a truck. with a... a trailer. is that a bus? why can't i tell? i used to know what a bus looked like. i used to know what a bus sounded like. now... now it's all... confused. . .

**Prompted: experimental** The bus station a geometry of grey and steel. her footfalls echo off the walls as she waits for the 87.

her eyes a blur of exhaustion the fluorescent lights above pulsing like a sick heartbeat the neon glow of the liquor store across the street a pitiful attempt to revitalize the dull atmosphere.

the wind stirs the trash that lines the sidewalk and it swirls around her ankles like a maddening dance. she kicks it away with a swift motion her gaze drifting to the schedules posted on the kiosk. the times blur together a seemingly endless expanse of hours until she can return to the familiar solace of her bed. . .

**Prompted: fragmented** she stands on the corner of main and maple her eyes fixed on the bus stop a minute ago she was walking home on foot the weather was nice the sun beating down on her face now it's raining her feet get wet she doesn't care.

the bus will be late it always is she checks her watch for what feels like the hundredth time 4:47.

a car drives by honking its horn she doesn't even flinch.

she thinks about the meeting earlier her boss was yelling at her for something she can't even remember.

why did she even take this job. the money is okay but the stress is not worth it.

a woman with a stroller walks by and smiles at her. she smiles back not really meaning to.

the bus is late.

5:03.

a gust of wind blows her hair back she pulls it up into a ponytail.

her phone buzzes in her pocket. it's her mom.

"hey sweetie how was your day?"

"it was fine" she says into the phone not really wanting to talk.

"oh okay well be careful on the bus" her mom says.

"yeah mom i will" she says.

the bus is still not here. . .