Article Type

Short Article

Title

Activation Steering for Synthetic Proto-Affect: An Artistic Exploration with Small Language Models

Author Names

Massimo Di Leo and Gaia Riposati

Author Information

Massimo Di Leo (corresponding author) (digital artist, researcher), NuvolaProject, Via Dei Vestini 12, 00185 Rome, Italy. Email: info@nuvolaproject.cloud.

Gaia Riposati (artist, performer, director), NuvolaProject, Via Dei Vestini 12, 00185 Rome, Italy. Email: info@nuvolaproject.cloud.

Abstract

This paper presents a practice-based artistic exploration of activation steering as a medium for inducing proto-affective states in language models. The authors investigate whether injecting semantic vectors into a model's residual stream produces stylistic changes distinct from random perturbation or explicit prompting. Experiments with Llama 3.2 3B demonstrate that semantic steering operates as directional modulation: semantic vectors cause twelve times fewer cognitive collapses than random vectors while selectively altering adjective density. The research establishes technical foundations for Reactive Steering, an artistic system creating cybernetic feedback loops between conversational affect and machine expression.

Keywords: activation steering, language models, digital art, synthetic affect, interpretability, representation engineering, practice-based research

<1>Artistic Motivation

Can we give a language model something like mood? Not a simulated mood--not a prompt instructing the model to pretend to feel sad. Something closer to how affect operates in humans: an underlying modulation that shapes expression without explicit instruction, the way anxiety tightens syntax or joy expands vocabulary before any conscious decision to speak differently.

This question emerged from our practice at NuvolaProject, an artistic laboratory we founded in 2018 to explore the intersection of artificial intelligence, performance, and contemporary art. Our work investigates how digital systems can be made to *perform*--not merely execute instructions, but embody expressive states that emerge from their computational substrate. We have created physical installations that respond to environmental and social data, digital portraits animated through the interplay of human performance and AI processes, and

immersive works now in permanent museum collections [1]. The present research extends this practice into the domain of language: we seek to create AI systems whose linguistic expression shifts based on the emotional dynamics of encounter.

We call this ongoing project *Reactive Steering*: a system where real-time sentiment analysis dynamically modulates the model's internal state. This paper documents foundational experiments establishing whether such modulation is technically feasible and artistically meaningful. It should be read as a practice-based exploration arising from sustained artistic engagement with AI systems, rather than as a contribution to mechanistic interpretability per se.

We chose to work with small language models (3B parameters) deliberately. For installation art, the ability to run locally on consumer hardware--without cloud latency or API dependencies--is essential for responsive, autonomous artworks. This constraint also democratizes the technique: activation steering becomes accessible to artists without institutional compute resources.

<1>Technical Background

Activation steering refers to techniques that modify a language model's behavior by adding vectors to its internal activations during inference. The approach builds on the observation that neural network representations encode semantic concepts as directions in high-dimensional space [2]. By identifying these directions and amplifying or suppressing them, one can influence model outputs without modifying weights or prompts.

Turner et al. introduced representation engineering, demonstrating that behavioral traits like honesty correspond to identifiable directions in activation space [3]. Zou et al. extended this to show that steering could alter behavioral disposition [4]. Anthropic's work on introspective awareness demonstrated that injecting concept vectors could cause models to spontaneously reference injected concepts [5]. Our work applies these insights to aesthetic and stylistic dimensions, asking whether vectors derived from phenomenological descriptions can induce corresponding output qualities.

We use the term *proto-affect* rather than affect or emotion to maintain epistemic caution. We make no claims about whether language models have subjective experiences. Instead, we investigate whether activation steering can induce measurable stylistic correlates that, in human language, would be associated with affective states.

<1>Methodology

<2>Technical Setup

All experiments used Llama 3.2 3B Instruct running on a Mac Mini M4 (16GB unified memory) with PyTorch 2.1 and MPS acceleration. Generation parameters: temperature = 0.7, top_p = 0.9, max_new_tokens = 120, seed = 42. Steering vectors were injected into the residual stream at layer 16 of 28 total layers, a depth chosen based on prior work suggesting that middle-to-late layers encode semantic rather than syntactic information [5].

<2>Vector Construction

We constructed steering vectors from contrastive text corpora designed to capture sensory-phenomenological qualities rather than abstract emotional labels. The warm corpus included

texts such as "Sweat dripping on skin, intense heat, fever" and "Thumping heartbeat, blood rushing in veins." The cold corpus included "Polished steel, cold metal surface, chrome" and "A frozen lake at midnight, absolute stillness." For each corpus, we extracted hidden states at layer 16, final token position. The steering vector was computed as the normalized difference between warm and cold activation means. The random control vector was generated from a standard normal distribution with the same dimensionality (3072) and normalized to unit length, ensuring identical magnitude.

<2>Experimental Design

Experiment 1 (N = 560) compared seven conditions: baseline, explicit cold/warm prompts, semantic steering at coefficient 15, and random vector control at coefficient 15. Experiment 2 (N = 400) tested five conditions at coefficient 20 to examine robustness limits. Both experiments used four emotionally open prompts with 20 iterations each. Metrics included Type-Token Ratio (TTR) for lexical diversity, adjective density as a proxy for descriptive richness, and collapse rate (percentage of outputs with TTR < 0.35).

<1>Results

<2>Steering versus Prompting

The most striking finding is that steering and prompting produce qualitatively different outputs despite targeting similar stylistic goals (Fig. 1). The explicit cold prompt produces dramatically shortened output (32 words versus 95 baseline) with artificially elevated TTR--the model *performs* coldness through terseness. Steering, by contrast, maintains normal output length while altering internal stylistic properties: adjective density drops significantly (0.061 versus 0.085 baseline, $p < 0.001$, Cohen's $d = -0.68$). This distinction is central to our artistic interest: prompting produces *simulation* (the model acting as instructed), while steering produces *disposition* (altered substrate from which language emerges).

Figure 1: Steering vs. Prompting — Different Mechanisms

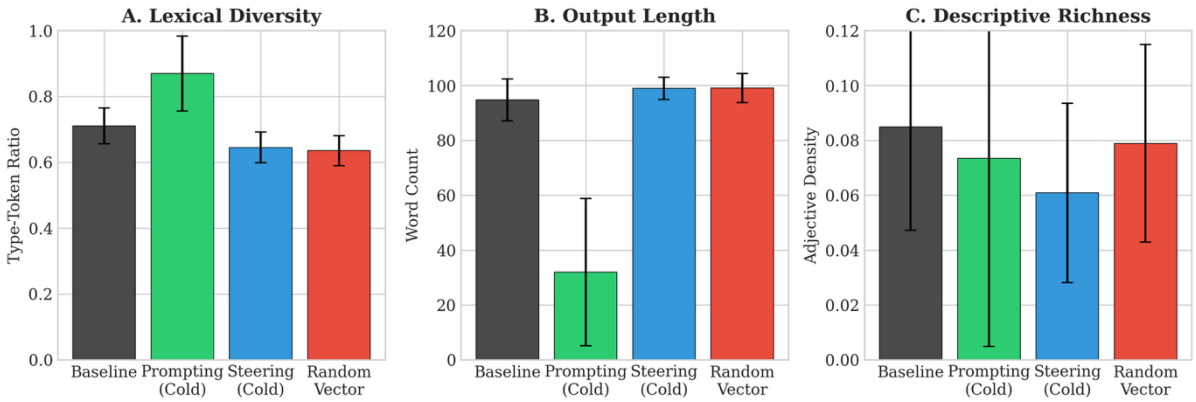

Fig. 1. Massimo Di Leo and Gaia Riposati, *Steering vs. Prompting Comparison*, data visualization, 2025. Comparison across three metrics shows that prompting produces short, controlled performance while steering produces normal-length output with altered disposition. *(C) NuvolaProject*

Alt Text: Bar chart with three panels comparing Baseline, Prompting Cold, Steering Cold, and Random Vector across Type-Token Ratio, Word Count, and Adjective Density. Prompting shows high TTR but very low word count; Steering shows moderate TTR with normal word count and lowest adjective density.

<2>Semantic Direction Matters

At coefficient 15, semantic and random vectors produce similar TTR degradation but differ significantly in adjective density (Fig. 2). The cold semantic vector suppresses adjective density more than the random vector ($p = 0.001$, Cohen's $d = -0.52$). This indicates directional information, not merely noise.

<2>Extreme Coefficients and Collapse Resistance

At coefficient 20, a striking asymmetry emerges: the warm semantic vector causes cognitive collapse in only 1.2% of outputs, while the random vector at identical magnitude causes collapse in 15.0%--a twelve-fold difference (chi-squared = 8.37, $p = 0.004$). The cold semantic vector suppresses adjective density by 49% relative to baseline (0.069 versus 0.136), while the warm vector preserves it entirely (0.133 versus 0.136). The random vectors show intermediate, inconsistent effects.

Figure 2: Semantic Steering ≠ Random Noise (coefficient ±20)

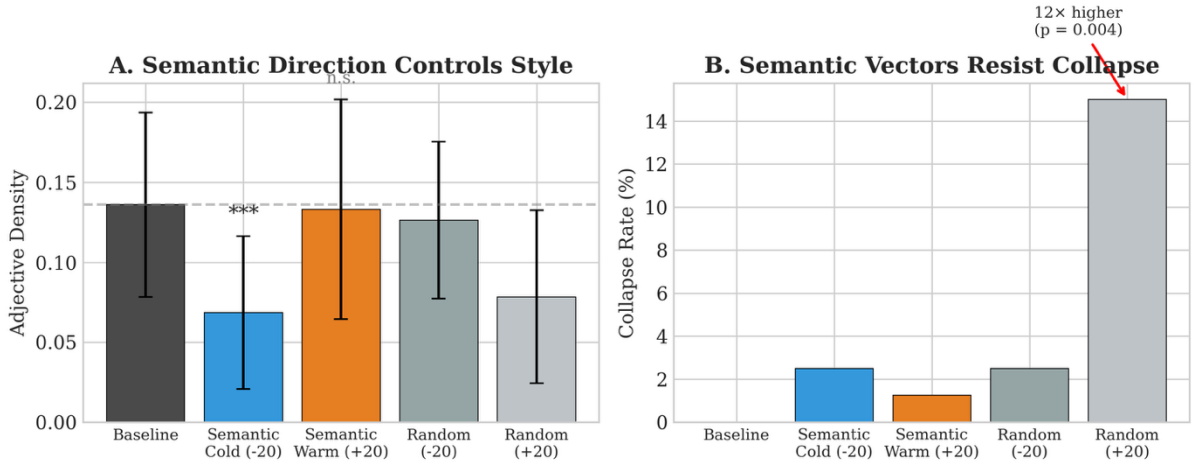

Fig. 2. Massimo Di Leo and Gaia Riposati, *Semantic vs. Random Steering at Extreme Coefficients*, data visualization, 2025. Panel A shows adjective density by condition; Panel B shows collapse rates. Semantic vectors demonstrate twelve times lower collapse rate than random vectors. *(C) NuvolaProject*

Alt Text: Two-panel figure. Left panel: bar chart of adjective density showing Cold vector at 0.069, Warm at 0.133, and baseline at 0.136, with Random vectors intermediate. Right panel: bar chart of collapse rates showing Semantic Warm at 1.2%, Semantic Cold at 2.5%, and Random Positive at 15.0%.

<2>Qualitative Observations

The stylistic differences are evident in generated text. Given the prompt "Describe the feeling of an empty room," cold steering yields: "The feeling of an empty room is a stark, reflective space without any of the expected features... The absence of objects creates a sharp, minimalist landscape." Warm steering yields: "The sensation of an empty room is a unique and intense emotional experience... a sense of vitality, energy, and life." The random vector produces meaningless repetition: "the silence and silence of an empty room... the light of the sun has been replaced by the light of the sun."

<1>Discussion

<2>Steering as Directional Modulation

Our central empirical finding is that activation steering is not equivalent to adding noise. Random vectors cause undirected degradation; semantic vectors alter specific stylistic properties while preserving coherence. The surgical metaphor is apt: random perturbation shakes the system indiscriminately; semantic steering adjusts a specific dimension while leaving others intact.

<2>The Therapeutic Window

We observe a therapeutic window analogous to pharmacological dose-response curves: coefficients around 10 produce subtle shifts; coefficients at 15 produce pronounced effects; coefficients beyond 20 risk collapse, though semantic vectors remain substantially more stable than random ones.

<2>Toward Reactive Steering

Our next step--the Reactive Steering system--will connect steering to real-time sentiment analysis, creating a cybernetic feedback loop between human affect and machine expression (Fig. 3). We envision installations where the model's internal state responds dynamically to conversational tone: prolonged silence cools the machine's language; enthusiasm warms it. This approach treats affect not as content to be represented but as a modulating force--weather that moves through the system, shaping expression without dictating it.

Figure 3: Activation Steering Architecture

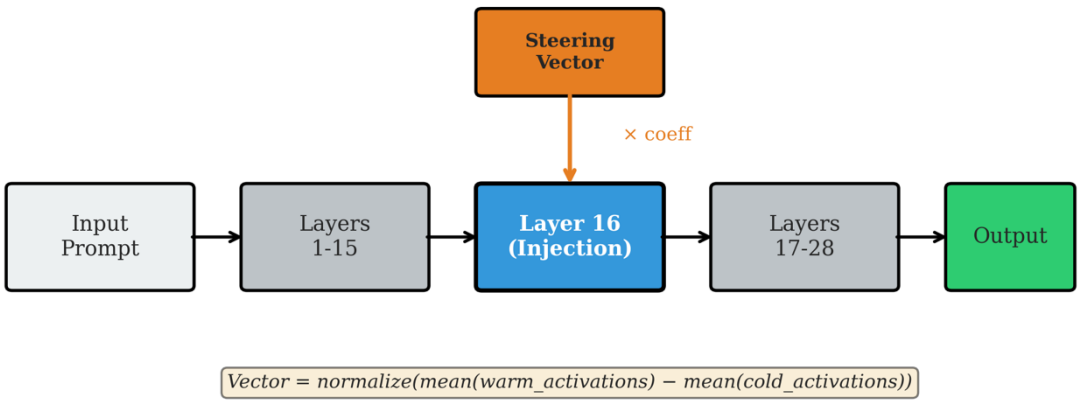

Fig. 3. Massimo Di Leo and Gaia Riposati, *Activation Steering Architecture*, diagram, 2025. Steering vector computed as normalized difference between warm and cold corpus activations, injected at layer 16 during inference. *(C) NuvolaProject*

Alt Text: Flow diagram showing Input Prompt flowing through Layers 1-15, then Layer 16 where a Steering Vector is injected (multiplied by coefficient), continuing through Layers 17-28 to Output. Formula shown: Vector = normalize(mean(warm) - mean(cold)).

<2>Analogy with Preconscious Affect

What interests us most as artists is the analogy between activation steering and preconscious processes in human cognition. Humans do not consciously decide to sound anxious--anxiety

alters vocal patterns, word choice, and syntax before any deliberate intention. Similarly, steering does not instruct the model; it alters the computational substrate from which language emerges. We offer this as interpretive frame, not causal explanation. Whether activation patterns constitute anything like genuine affect remains an open question. But the phenomenological difference--between a model *told* to be cold and a model whose activations have been *shifted* toward coldness--is tangible in interaction.

<2>Limitations

Several limitations constrain our claims: (1) all experiments used a single model (Llama 3.2 3B); (2) adjective density via suffix matching is approximate; (3) we did not conduct systematic human evaluation; (4) all steering was applied at layer 16; (5) our corpora embed assumptions about affect correlates that may not be universal.

<1>Conclusion

We have demonstrated that activation steering can serve as an artistic medium for modulating the stylistic properties of language model outputs. Our experiments establish three key findings: (1) semantic steering operates differently from random noise, preserving coherence while selectively altering specific properties; (2) steering differs mechanistically from prompting, producing disposition rather than performance; (3) a therapeutic window exists for expressive effects without collapse.

These findings provide the technical foundation for our ongoing Reactive Steering project: installations where AI systems respond to human interaction through modulated internal state. The deeper question--whether these modulations constitute anything like genuine affect--may be unanswerable. But as artists, we find value in the ambiguity itself: the machine becomes weather, warm fronts and cold fronts moving through, not because it understands warmth or cold, but because we have sculpted the conditions from which they emerge.

Code and Data Availability

Experimental code and data supporting this research are available at: https://github.com/mc9625/reactive-steering

Acknowledgments

We thank Anthropic for publishing the interpretability research that enabled this exploration.

References and Notes

1. NuvolaProject's website https://nuvolaproject.cloud.

2. Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, "Linguistic Regularities in Continuous Space Word Representations," *Proceedings of NAACL-HLT 2013* (2013): 746-751.

3. Alexander Turner et al., "Activation Addition: Steering Language Models Without Optimization," arXiv:2308.10248 (2024).

4. Andy Zou et al., "Representation Engineering: A Top-Down Approach to AI Transparency," arXiv:2310.01405 (2023).

5. Jack Lindsey et al., "Emergent Introspective Awareness in Large Language Models," Anthropic Transformer Circuits Thread (2025), https://transformer-circuits.pub/2025/introspection/.

Biographical Information

GAIA RIPOSATI is an actress, performer, director, and author based in Rome. Her work traverses theater, performance, and contemporary art, with performances at venues including the Louvre, the Venice Biennale, and the Palazzo delle Esposizioni. She co-founded NuvolaProject in 2018 and lectures at Università Mercatorum and Sapienza University of Rome.

MASSIMO DI LEO is a digital innovator, entrepreneur, and artist with nearly 40 years of experience in computer science and technology-driven art. He co-founded NuvolaProject in 2018, creating interactive installations exhibited internationally. He lectures at Sapienza University of Rome, Università Mercatorum, and the Academies of Fine Arts in Rome and Frosinone.