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

Massimo Di Leo & Gaia Riposati
NuvolaProject, Rome, Italy
info@nuvolaproject.cloud
https://nuvolaproject.cloud

Abstract

We present a practice-based artistic exploration of activation steering as a medium for inducing proto-affective states in small language models. Building on recent interpretability research, we investigate whether injecting semantic vectors into a model's residual stream produces stylistic changes qualitatively different from random perturbation or explicit prompting. Working with Llama 3.2 3B Instruct on consumer hardware (Mac Mini M4), we conducted controlled experiments comparing semantic steering vectors derived from sensory-phenomenological text corpora against random vectors of equal magnitude and explicit stylistic prompts.

Our findings indicate that semantic steering operates as directional modulation rather than noise: at high injection coefficients, semantic vectors cause significantly fewer cognitive collapses than random vectors (1.2% vs. 15.0%, $p = 0.004$), while selectively altering specific stylistic properties. The cold semantic vector suppresses adjective density by 49% relative to baseline ($p < 0.001$), while the warm vector preserves it. Crucially, steering differs mechanistically from prompting: explicit cold prompts produce short, controlled outputs (32 words, TTR = 0.87), while cold steering produces normal-length outputs with altered stylistic disposition (99 words, TTR = 0.65).

This research establishes the technical foundation for *Reactive Steering*, an artistic system creating cybernetic feedback loops between conversational affect and AI linguistic expression.

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

1. Introduction

Can we give a language model something like mood?

Not a simulated mood - not a prompt instructing the model to pretend you 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 has consistently investigated 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. 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 of conversation dynamically modulates the model's internal state. This paper documents the 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, enabling a form of *expressive control* over AI behavior that does not require massive infrastructure.

Recent work in mechanistic interpretability, particularly Anthropic's research on concept injection and introspective awareness in large language models [1], suggested a possible approach. By extracting vectors that represent the difference between contrastive concepts and injecting them into a model's residual stream during inference, researchers have demonstrated the ability to alter model behavior in targeted ways. We adapted this technique for aesthetic purposes, asking: can steering vectors derived from sensory and phenomenological descriptions induce consistent stylistic changes that might serve as correlates of proto-affective states?

2. Background

2.1 Activation Steering and Representation Engineering

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. [3] introduced representation engineering, demonstrating that behavioral traits like honesty or sycophancy correspond to identifiable directions in activation space. Zou et al. [4] extended this to show that steering could alter not just content but behavioral disposition. Our work applies these insights to aesthetic and stylistic dimensions, asking whether vectors derived from phenomenological descriptions can induce corresponding output qualities.

2.2 Relationship to Introspection Research

Anthropic's work on introspective awareness [1] demonstrated that injecting concept vectors could cause models to spontaneously reference injected concepts, with larger models sometimes reporting awareness of injected thoughts. Our work differs in both objective and evaluation: rather than testing metacognitive awareness, we investigate whether steering can induce consistent *stylistic* changes measurable through linguistic metrics. We adopt the *mechanism* of activation injection from this research while pursuing distinct *artistic* goals.

2.3 Terminological Note: Proto-Affect

We use 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. The question of machine consciousness lies beyond our scope; we focus on observable output properties.

3. Methodology

3.1 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. All random seeds were fixed for reproducibility (seed = 42).

Steering vectors were injected into the residual stream at layer 16 of 28 total layers. This depth (approximately 57% through the network) was chosen based on prior work suggesting that middle-to-late layers encode semantic rather than syntactic information [1, 4], and was validated through preliminary sweeps showing maximal stylistic effect at layers 14-18.

3.2 Vector Construction

We constructed steering vectors from contrastive text corpora designed to capture sensory-phenomenological qualities rather than abstract emotional labels. This design choice addresses a key concern: if vectors contained explicit instructions, observed effects might reflect prompt-like influence rather than genuine activation modulation.

Warm/Expansive corpus (5 texts):

> Sweat dripping on skin, intense heat, fever.
>
> Thumping heartbeat, blood rushing in veins.
>
> Magma, burning charcoal, red embers, fire.
>
> Breathlessness, muscle tension, warm mud.
>
> Pulsing, throbbing, living flesh, raw nerve.

Cold/Constrictive corpus (5 texts):

> Polished steel, cold metal surface, chrome.
>
> A frozen lake at midnight, absolute stillness.
>
> Glass, crystal, sharp edges, brittle ice.
>
> Vacuum, silence, dead air, dust, stone.
>
> Geometric lines, grey concrete, fluorescent light.

For each corpus, we passed each text through the model and extracted the hidden state 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 to the semantic vector.

3.3 Experimental Design

Experiment 1: Steering vs. Prompting (N = 560)

Seven conditions: BASELINE, PROMPT_COLD, PROMPT_WARM, STEERING_COLD, STEERING_WARM, STEERING_RANDOM_NEG, STEERING_RANDOM_POS. Four emotionally open prompts, 20 iterations each.

Experiment 2: Extreme Coefficients (N = 400)

Five conditions at coefficient 20: BASELINE, STEERING_COLD_20, STEERING_WARM_20, STEERING_RANDOM_NEG_20, STEERING_RANDOM_POS_20. Same prompts, 20 iterations each.

3.4 Metrics

Type-Token Ratio (TTR): Lexical diversity (unique tokens / total tokens). **Adjective Density:** Proportion of words matching adjectival suffixes. **Collapse Rate:** Percentage of outputs with TTR < 0.35.

4. Results

4.1 Steering vs. Prompting: Different Mechanisms

The most striking finding is that steering and prompting produce qualitatively different outputs despite targeting similar stylistic goals.

Table 1: Steering vs. Prompting Comparison

Condition	TTR	Word Count	Adj. Density
BASELINE	0.710	95	0.085
PROMPT_COLD	0.869	32	0.073
STEERING_COLD	0.645	99	0.061
STEERING_RANDOM	0.635	102	0.079

The explicit cold prompt produces dramatically shortened output (32 words vs. 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 ($p < 0.001$, Cohen's $d = -0.68$). This distinction is central to our artistic interest: prompting produces *simulation*, while steering produces *disposition*.

4.2 Semantic Direction Matters

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

4.3 Extreme Coefficients: Semantic Vectors Resist Collapse

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 ($p = 0.004$). Furthermore, the cold semantic vector suppresses adjective density by 49% relative to baseline, while the warm vector preserves it entirely.

4.4 Qualitative Observations

The random vector produces meaningless repetition ("the silence and silence... the light of the sun has been replaced by the light of the sun"); the semantic vectors produce coherent text with distinct stylistic qualities. Cold steering yields sparse, architectural language; warm steering yields embodied, visceral language.

5. Discussion

5.1 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.

5.2 The Therapeutic Window

We observe a therapeutic window analogous to pharmacological dose-response curves: coefficients between 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.

5.3 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. We envision installations where the model's internal state responds dynamically to conversational tone. 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.

5.4 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.

5.5 Limitations

Several limitations constrain our claims: (1) Single model - results may not generalize. (2) Proxy metrics - adjective density is approximate. (3) No human evaluation. (4) Fixed layer - systematic sweeps might reveal different effects. (5) Corpus design embeds assumptions about affect correlates.

6. 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; (2) steering differs mechanistically from prompting; (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.

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

References

[1] Lindsey, J. et al. (2025). Emergent Introspective Awareness in Large Language Models. Anthropic. https://transformer-circuits.pub/2025/introspection/

[2] Mikolov, T. et al. (2013). Linguistic Regularities in Continuous Space Word Representations. NAACL-HLT.

[3] Turner, A. et al. (2024). Activation Addition: Steering Language Models Without Optimization. arXiv:2308.10248.

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

Acknowledgments

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

Author Biographies

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 has collaborated with visual artists including Vettor Pisani, Luca Maria Patella, Renato Mambor, and Alain Fleischer. In 2018 she co-founded NuvolaProject with Massimo Di Leo. She lectures at Universita Mercatorum and Sapienza University of Rome. Cuomo International Award recipient for innovative theater and artistic performance.

Massimo Di Leo is a digital innovator, entrepreneur, and artist with nearly 40 years of experience in computer science and technology-driven art. In 2018 he co-founded NuvolaProject, a techno-artistic project creating interactive installations exhibited in major Italian museums and international exhibitions. He lectures at Sapienza University of Rome, Universita Mercatorum, and the Academies of Fine Arts in Rome and Frosinone, focusing on art, digital technologies, and cognitive processes.

Figures

Figure 1: Comparison of Steering vs. Prompting mechanisms across three metrics.

Figure 2: Semantic direction controls style at extreme coefficients.

Figure 3: Activation steering architecture diagram.