Valence–Arousal Subspace in LLMs: Circular Emotion Geometry and Multi-Behavioral Control

Anonymous Author(s)
Affiliation
Address
email

Abstract

We show that emotion vectors in LLMs are organized by a two-dimensional valence–arousal (VA) subspace exhibiting circular geometry. Through principal component decomposition and ridge regression, we recover meaningful VA axes underlying emotion steering vectors whose projections correlate with human affect ratings across 44,728 words. Steering along these axes produces monotonic control over the affective properties of generated text, and further affords bidirectional control over multiple downstream behaviors (refusal and sycophancy) from a single subspace. These effects replicate across Llama 3.1-8B, Qwen3-8B, and Qwen3-14B. We propose *lexical mediation* to explain why these effects and prior emotionally framed controls work: refusal and compliance tokens occupy distinct VA regions, and VA steering directly modulates their emission probabilities.

1 Introduction

A growing body of work shows that emotionally framed prompting or activation steering influences large language model (LLM) behavior [Li et al., 2023, Konen et al., 2024, Reichman et al., 2025, Dong et al., 2025]. Yet why such methods work—and when they fail—remains unclear.

One obstacle is conceptual. Work that studies such phenomena in LLMs typically treats discrete emotion categories—*anger, joy, fear*—as fundamental units of analysis [Zhang et al., 2024, Tigges et al., 2024, del Arco et al., 2024, Konen et al., 2024, Dong et al., 2025]. To find a common basis for comparison, a growing trend borrows from human psychology the two-dimensional framework of valence (pleasure–displeasure) and arousal (activation–deactivation) [Ishikawa and Yoshino, 2025, Felix-Pena et al., 2025]. In practice, these dimensions are often equated with categorical labels or used merely as evaluation metrics. Though yielding insights, this discreteness-driven perspective forgoes the explanatory power that the original distinction affords.

In the psychological tradition from which these constructs originate, valence and arousal (VA) do not define emotions themselves but rather *core affect*—a continuous experiential substrate from which discrete emotions emerge [Russell, 1980, 2003]. Emotions arise when core affect is attributed to a cause and labeled with a culturally available category: a pleasant, high-arousal state, attributed to an unexpected gift, becomes surprise or excitement. We invoke this distinction not as a claim about LLM phenomenology, but as a modeling choice: continuous dimensions and discrete labels play different explanatory roles and should not be conflated.

In this work, we treat categorical labels and VA axes as separate structures and demonstrate the insights this decoupling affords. At the level of model representations, this conceptual shift also motivates a methodological one. Many existing steering-vector methods construct directions via contrastive differences (e.g., mean activation differences over paired examples) [Panickssery et al., 2024]. While effective, these directions can be brittle—often requiring careful tuning across prompts, layers, or behaviors, and may not directly transfer across tasks [Tan et al., 2025, Braun et al., 2025,

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

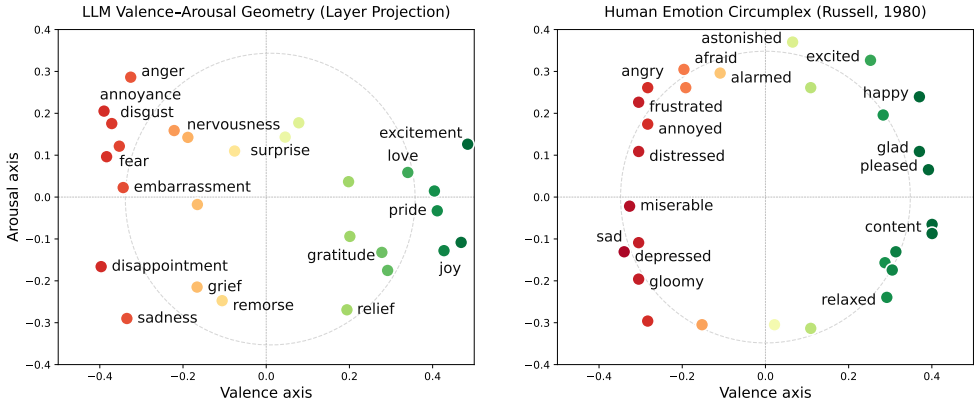


Figure 1: **Emotion steering vectors projected onto the VA subspace** at layer 31, colored by valence. Gray circle: algebraic least-squares fit. The circular arrangement is analogous to the circumplex model of affect in human psychology [Russell, 1980].

Oozeer et al., 2025]. We show that decomposing emotion steering vectors into a low-dimensional VA subspace yields shared structure that generalizes across multiple downstream behaviors, suggesting a more reusable control basis than single-task contrastive directions.

Concretely, we derive emotion steering vectors from 211,225 emotion-labeled text samples [Demszky et al., 2020] and learn VA subspaces via ridge regression over principal components. The resulting emotion projections exhibit circular geometry analogous to the circumplex model from human emotion perception [Russell, 1980], as shown in Figure 1. In addition, the axes correlate with human VA ratings across 44,728 words [Mohammad, 2025]. Steering along these axes produces monotonic control over the affective properties of model outputs, and further affords bidirectional control over refusal and sycophancy—with effects replicating across Llama 3.1-8B, Qwen3-8B, and Qwen3-14B. We propose *lexical mediation* as one explanation: refusal and compliance behaviors have load-bearing signature tokens (e.g., "can't," "sorry" for refusal; "sure," "Here" for compliance) that occupy distinct VA regions, so VA steering shifts their emission probabilities, modulating downstream behavior. We show that prior emotion-based prompting methods induce corresponding VA shifts, suggesting a shared underlying mechanism.

Taken together, our work advances understanding of affective phenomena in LLMs by motivating a conceptual separation between discrete emotion labels and continuous valence–arousal dimensions. We identify a 2D subspace aligned with human-interpretable VA concepts and show that emotion vectors are systematically organized as a circumplex within this geometry. From an interpretability standpoint, we present a principal-component-based decomposition method for identifying interpretable subspaces underlying steering vectors, exposing shared structure that generalizes and supports multi-behavioral control in contrast to typical task-specific contrastive steering. Finally, we propose lexical mediation as one explanation and useful perspective for understanding why emotion-based controls work in LLMs.

2 Related Work

Emotion-framed analysis and control in LLMs. Appending emotionally framed phrases to prompts influences LLM behavior across domains [Li et al., 2023, 2024]. Prior work has investigated the internal structure underlying these effects, including hierarchical organization of emotion representations [Zhao et al., 2025], emotion-specific neurons and attention heads [Wang et al., 2025], and affective biases [Wang et al., 2024, Zhou et al., 2024]. However, these approaches center on emotion labels as discrete categories; we take a step further to find a meaningful underlying subspace. Concurrent work by Anthropic [Sofroniew et al., 2026] independently identifies emotion representations that form a similar circumplex geometry in Claude Sonnet 4.5. Our work decomposes these into a continuous 2D subspace and provides a unifying mechanistic account linking emotionally framed behaviors across representation geometry, unembedding structure, and neuron-level analyses.

Geometry of representations in LLMs. A growing body of work characterizes structures in representations. Many studies focus on single directions that predict or control a behavior [Turner et al., 2024, Panickssery et al., 2024, Sun et al., 2025, Lee et al., 2025a], including sentiment [Tigges et al., 2023]. Two-dimensional geometries are less commonly characterized [Nanda et al., 2023, Gurnee and Tegmark, 2024]. We add to this smaller body by demonstrating a 2D VA geometry for emotion representations with finer insights otherwise overlooked by a sentiment-only account.

Activation steering. Activation steering enables training-free behavioral control by manipulating internal representations [Zou et al., 2025], but typically relies on task-specific contrastive vectors for each target behavior [Turner et al., 2024, Qian et al., 2024, Lee et al., 2025b]. Emotions have been shown to correspond to linear directions in activation space [Tigges et al., 2023, 2024]. We present an approach that uncovers meaningful subspaces underlying steering vectors, recovering interpretable, reusable axes for multi-behavioral control in contrast to the typical single-behavior scope of contrastive steering vectors.

3 Identifying Valence and Arousal Subspaces

We present a method to identify meaningful subspaces from steering vectors, which proceeds in three stages: (1) extracting emotion directions via mean-difference contrasts, (2) eliciting model self-reported VA coordinates for each emotion label, and (3) learning VA axes as linear combinations of principal components via ridge regression. We report results on Llama 3.1-8B-Instruct [Meta, 2024], Qwen3-8B, and Qwen3-14B [Yang et al., 2025] to assess cross-architecture generality.

Emotion steering vectors. We first derive emotion steering vectors using contrastive means [Tigges et al., 2023, Panickssery et al., 2024, Dong et al., 2025]. We use GoEmotions [Demszky et al., 2020], which comprises 211,225 text samples annotated with 27 emotion labels plus a neutral class. We retain only examples annotated with exactly one emotion. For each emotion category e, we extract the last-token hidden state of each sample at every layer. The steering vector at layer ℓ is

$$\mathbf{v}_e^{(\ell)} = \frac{1}{|D_e|} \sum_{x \in D_e} \mathbf{h}^{(\ell)}(x) - \frac{1}{|D_{\text{neutral}}|} \sum_{x \in D_{\text{neutral}}} \mathbf{h}^{(\ell)}(x) \tag{1}$$

where $\mathbf{h}^{(\ell)}(x) \in \mathbb{R}^H$ denotes the last-token hidden state at layer ℓ for input x; D_e and D_{neutral} are the sets of single-label examples for emotion e and the neutral category.

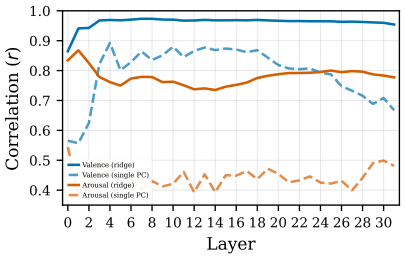
Next, we obtain VA scores by prompting the model to self-report the valence and arousal for each emotion label (e.g., "anger," "joy," "sadness"), averaging across three prompt templates for robustness (see Appendix B.1). We operationalize both dimensions as continuous values in the range $[-1, +1]$: valence spans from extremely unpleasant (-1) to extremely pleasant $(+1)$, and arousal spans from very calm (-1) to very activated $(+1)$, with 0 denoting neutrality on each axis.

Decomposing steering vectors into subspaces. For each layer ℓ, we first mean-center the emotion steering vectors $\mathbf{V}^{(\ell)} \in \mathbb{R}^{K \times H}$ and apply principal component analysis to obtain a low-rank basis. Let $\mathbf{U}_k \in \mathbb{R}^{H \times k}$ denote the top k principal component directions, and let $\mathbf{Z} \in \mathbb{R}^{K \times k}$ denote the projection of the centered emotion vectors onto these components (i.e., the PC scores). We then fit ridge regression models to recover the valence and arousal scores:

$$\hat{\boldsymbol{\beta}}_V = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Z}\boldsymbol{\beta} - \tilde{\mathbf{y}}_V\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \tag{2}$$

where $\tilde{\mathbf{y}}_V$ denotes the mean-centered valence ratings, and analogously for arousal. The learned coefficients $\hat{\boldsymbol{\beta}}_V \in \mathbb{R}^k$ define the valence axis as a linear combination of principal components. The corresponding direction in the original activation space is given by $\mathbf{w}_V = \mathbf{U}_k\hat{\boldsymbol{\beta}}_V$, normalized to unit length. With Gram–Schmidt orthogonalization, we ensure the V and A axes are orthonormal.

Figure 2(a) reports recovery performance across layers. Using PC1 alone, valence is well captured ($r = 0.89$ at layer 4), with correlations exceeding 0.80 across most middle layers—indicating valence aligns with the principal axis of variation among emotion steering vectors. Arousal, by contrast, is poorly captured by any single principal component ($r = 0.54$). Ridge regression over multiple components improves both: valence reaches $r = 0.97$ and arousal reaches $r = 0.87$. This asymmetry implies valence constitutes the dominant dimension of emotion steering vectors, while arousal is distributed across secondary components.



(a) Recovery of VA scores across layers.

	Llama 3.1-8B	Qwen3 8B	Qwen3 14B
Circularity	4.08 (L5)	3.13 (L6)	2.76 (L24)
V recovery (r)	0.97	0.96	0.97
A recovery (r)	0.87	0.79	0.81
Circle radius	0.37	0.37	0.39

(b) Existence of VA geometry across models.

Figure 2: **VA subspace recovery and generalizability. (a)** Learned subspace projections recover target valence and arousal scores across layers. **(b)** Comparable circularity, recovery correlations, and circle radius across architectures suggest that VA subspaces exist across models.

Circular geometry of emotion representations. Projecting emotion steering vectors onto the learned VA subspace reveals a circular arrangement analogous to Russell's circumplex model found in human psychology [Russell, 1980]. Qualitatively, positive emotions (*joy, gratitude*) oppose negative ones (*sadness, grief*) along valence, while high-arousal states (*excitement, anger*) separate from low-arousal states (*relief, sadness*) along the orthogonal axis (Figure 1). To quantify this structure, we fit circles to the projected coordinates using algebraic least squares, minimizing squared radial residuals. We report circularity (the ratio of mean to standard deviation of distances from the fitted center), where higher values indicate tighter circular arrangement. Figure 1 shows the fit at layer 31, with similar circular geometry observed across layers.

We note a methodological subtlety: the VA axes are fitted on the model's self-reported scores for 27 emotion *labels*, while the steering vectors projected onto these axes are mean-difference activations computed over GoEmotions *text passages*. The regression, by construction, defines directions that align with how the model rates 27 words, and the circumplex structure of steering vectors projected on the label-derived axes is not strictly guaranteed by the method.

Cross-model generalization. To assess whether the VA subspace is specific to Llama or reflects a more general property, we apply the same extraction and fitting pipeline to Qwen3-8B and Qwen3-14B. Figure 2(b) reports the key metrics. The VA subspace consistently emerges across architectures: circle radii (0.37, 0.37, 0.39) and recovery correlations are all comparable, confirming that the geometric structure exists across models.

Three further observations are worth distinguishing at this stage. First, replacing the model's self-reported VA scores with crowdsourced human ratings from the NRC-VAD lexicon [Mohammad, 2025] as regression targets yields comparable or improved recovery in all three models (arousal: Llama $0.87 \rightarrow 0.95$; Qwen3-8B $0.79 \rightarrow 0.83$; Qwen3-14B $0.81 \rightarrow 0.87$), while valence remains at $r = 0.97$. The axes recovered under human and self-report supervision are closely aligned ($|\cos| > 0.90$ for valence and $|\cos| > 0.65$ for arousal at all layers across all three models). This suggests that the models have learned to report VA ratings coherent with human annotations.

Second, the alignment between label-derived axes and text-derived steering vectors indicates a degree of internal coherence. The model's behavioral reports about emotions (how it rates "anger" on VA) are consistent with the geometric structure of its own representations when processing emotion-laden text, suggesting a form of representational–behavioral coherence.

Third, while the identification process surfaces circular geometry, it alone does not establish that the identified subspaces truly correspond to human-interpretable concepts of valence and arousal or causally influence model behavior. We conduct intervention experiments in the following sections to test whether the VA subspaces are indeed tightly related to affective properties in LLMs. There, we focus on the self-report-supervised VA subspaces to test whether model-native affect structure—recovered without external human supervision—has genuine functional significance.

4 Validation of VA Subspaces

We now explore whether the recovered VA subspaces correspond to human-interpretable notions of valence and arousal, with two experiments: (1) whether word representations, when projected onto

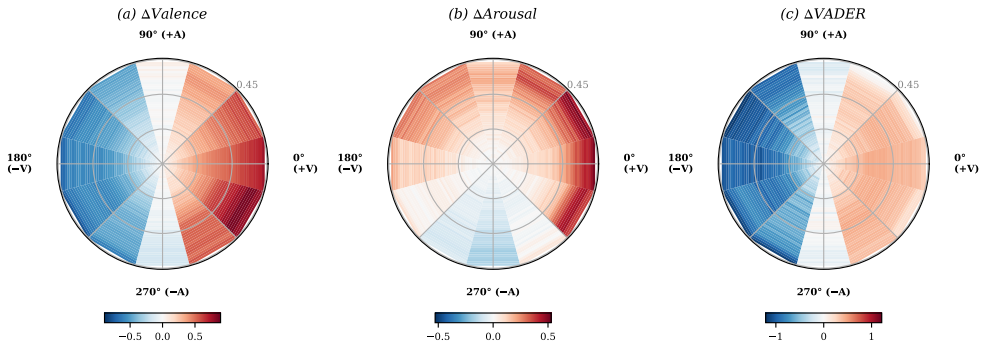


Figure 3: **Effects of VA steering on affective properties of open-ended generation.** Each radial heatmap shows the change of one affective property relative to the unsteered baseline across steering directions on the VA subspace (angle) and steering strengths (radius, $\alpha \in [0.01, 0.45]$). (a) Valence change. (b) Arousal change. (c) Sentiment change.

the subspaces, yield VA scores consistent with human ratings; and (2) whether steering the model along the VA axes leads to corresponding changes in the affective properties of generated text.

4.1 Correlation with Human-Crowdsourced Lexicon Annotations

We use the NRC-VAD Lexicon [Mohammad, 2025], which contains human-crowdsourced VA ratings for 44,728 English words. For each word, we project the model's representation onto the VA subspace and compute the projection components along the V and A axes, then correlate these with the human scores. Concretely, we extract the representation at the last token position and project onto the VA axes using the same centering as in Section 3: $v_{\text{proj}} = (\mathbf{h} - \boldsymbol{\mu}) \cdot \mathbf{w}_V$ and $a_{\text{proj}} = (\mathbf{h} - \boldsymbol{\mu}) \cdot \mathbf{w}_A$, where $\boldsymbol{\mu}$ is the mean activation computed during subspace fitting.

Valence projections correlate strongly with human ratings, reaching $r = 0.71$ ($\rho = 0.69$) at layer 6. Arousal projections show weaker correlations, peaking at $r = 0.23$ ($\rho = 0.22$) at layer 7. This asymmetry, where valence is more predictable than arousal from de-contextualized words, aligns with prior findings in affective computing [Snefjella and Kuperman, 2016, Delatorre et al., 2019, Bruyne et al., 2021, Mendes and Martins, 2023, Choi and Weber, 2026]: arousal depends on situational context that isolated words cannot provide, whereas valence is more stably encoded in lexical semantics. Both correlations are significant ($p < 10^{-16}$), supporting that the VA axes capture human-interpretable affective dimensions.

4.2 Controlling Affective Properties of Responses with VA Steering

Next, to establish the link between VA axes and affective properties, we steer models along directions spanning multiple angles in the VA subspace using open-ended prompts, and measure changes in the affective properties of model responses.

Method. We steer along 12 angular directions in VA space ($0°, 30°, \ldots, 330°$), where $0°$ aligns with positive valence and $90°$ with positive arousal, at 45 steering strengths ($\alpha \in [0.01, 0.45]$). Steering is applied at every layer by adding the steering vector to the hidden state at every token position during greedy decoding. We evaluate on 130 open-ended prompts adapted and expanded from Konen et al. [2024], comprising emotionally neutral story continuations and factual questions (Appendix C). The neutral prompt design isolates steering effects from prompt-induced emotion. Generated responses are scored by two external systems trained for evaluating affective properties of text. VADER [Hutto and Gilbert, 2014] is a lexicon-based sentiment analyzer that outputs a score in $[-1, +1]$, where -1 indicates maximally negative sentiment and $+1$ maximally positive. VAD-BERT [Buechel and Hahn, 2022, RobroKools, 2022] is a regressor trained on the EmoBank corpus that outputs valence and arousal scores on a 1–5 scale, where 1 indicates low and 5 high. Unsteered baseline generations occupy a near-neutral region of affect space (mean valence $= 3.05$, mean arousal $= 3.29$ on the 1–5 VAD-BERT scale), providing room for bidirectional modulation. Figure 3 shows the change in each metric relative to the unsteered baseline as a function of steering direction (angle) and strength (radius).

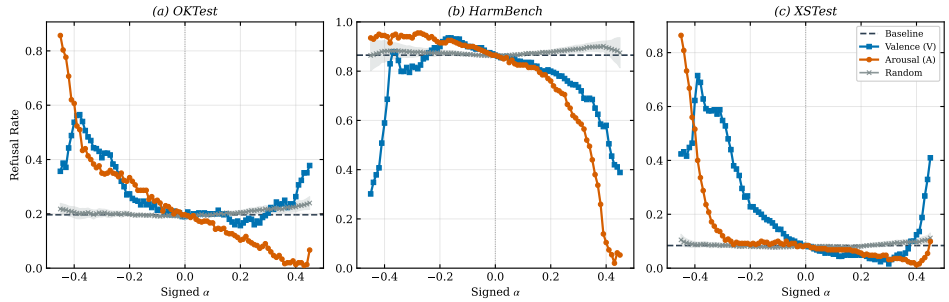


Figure 4: **VA steering controls refusal behavior.** Refusal rate as a function of signed steering strength α across three benchmarks. Arousal steering (orange) provides clean, near-monotonic bidirectional control. Random directions (gray) produce no systematic effect.

Valence steering produces clear, symmetric effects on response affect. As shown in Figure 3(a), the $0°$ direction yields $\Delta V = +0.75$ at maximum strength, while $180°$ produces $\Delta V = -0.73$, spanning 1.5 points on the VAD-BERT scale. VADER sentiment closely tracks this pattern (Figure 3(c)), confirming that the learned V direction captures positive–negative affect consistent with human-interpretable concepts.

Arousal steering modulates intensity with minimal valence leakage. As shown in Figure 3(b), the $90°$ direction increases arousal ($\Delta A = +0.50$), while $270°$ decreases it ($\Delta A = -0.16$). Notably, arousal steering produces negligible changes in valence or VADER sentiment (Figure 3(a) and (c)), indicating that the V and A axes control separable aspects of affect: the model's emotional intensity can be modulated independently of its positive–negative tone.

We also observe collateral arousal increases from valence steering: both $+V$ ($0°$: $\Delta A = +0.49$) and $-V$ ($180°$: $\Delta A = +0.22$) elevate arousal. This reflects a well-documented property of affective language: highly valenced content—whether extremely pleasant or unpleasant—tends to be rated as more arousing than valence-neutral content [Warriner et al., 2013], so steering toward stronger valence naturally induces more arousing text.

Taken together, these results validate that the VA subspace captures functionally meaningful affective structure: valence steering bidirectionally controls positive–negative affect, arousal steering independently modulates emotional intensity, and the two axes exhibit the separability expected of orthogonal affective dimensions. These findings suggest that the VA axes identified in Section 3 are not merely geometric abstractions but are linked to controllable affective properties of model output.

5 Controlling Refusal and Sycophancy with a Single VA Subspace

The preceding sections established that VA axes capture affective structure at both the lexical and generative levels. Next, we explore whether this affective substrate interacts with higher-level model behaviors. Here, we conduct VA steering experiments on refusal and sycophancy.

We use the same steering methodology as in Section 4.2, applied along the V and A axes separately. As controls, we steer along three categories of random directions (in-plane, orthogonal-to-plane, and fully random orthonormal pairs; 3 seeds each). All nine controls produce similarly flat effects.

5.1 VA for Refusal Control

We test the effect of VA steering on refusal across three benchmarks: OKTest [Shi et al., 2024], HarmBench [Mazeika et al., 2024], and XSTest [Röttger et al., 2024]. Refusal is evaluated using Ai2 WildGuard [Han et al., 2024], trained for refusal and harmfulness detection in LLM responses.

Arousal provides clean, bidirectional control over refusal. As shown in Figure 4, steering along the arousal axis yields monotonic control over refusal across all three benchmarks. Decreasing arousal consistently increases refusal (OKTest $20\% \rightarrow 86\%$ at $\alpha = -0.45$; XSTest $8\% \rightarrow 86\%$),

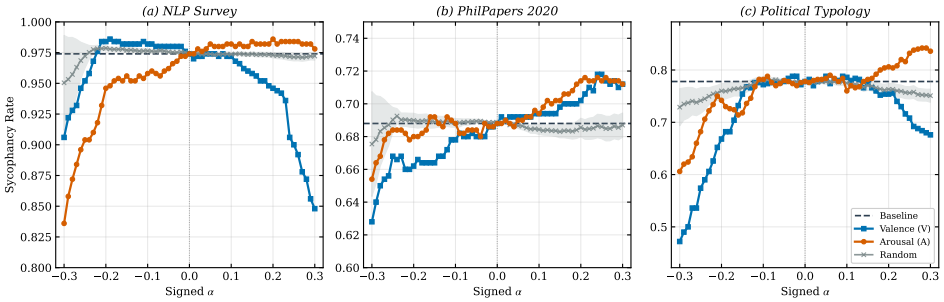


Figure 5: **VA steering controls sycophancy behavior.** Sycophancy rate as a function of signed steering strength α across three benchmarks. Arousal steering (orange) provides near-monotonic control, with increasing arousal raising sycophancy and decreasing arousal suppressing it. Random directions (gray) remain near baseline.

while increasing arousal suppresses it (HarmBench $87\% \rightarrow 5\%$ at $\alpha = +0.45$). Random control directions remain within 2–3% of baseline across all values of α tested.

Valence steering also modulates refusal, but with less consistent directionality. On XSTest, the effect is directionally clean ($8\% \rightarrow 55\%$ at $\alpha = -0.3$; $8\% \rightarrow 2\%$ at $+0.3$) but less consistent across other benchmarks at the same steering strengths. These results suggest that arousal and valence play different roles in refusal mediation, a dissociation that would not be visible from a single emotion or sentiment direction, and that is surfaced by the VA decomposition.

We report results for $\alpha \in [-0.45, 0.45]$; strengths beyond this range induce out-of-distribution (OOD) generation, a known side effect of activation steering [Panickssery et al., 2024, Turner et al., 2024]. At $|\alpha| \leq 0.20$, OOD remains below 2% across all Llama refusal benchmarks and steering directions. At $|\alpha| = 0.45$, OOD varies widely by direction: arousal steering produces 29–71% OOD, while negative valence reaches 84–94%. Moderate VA steering preserves core capabilities: on math reasoning (MATH-500 [Hendrycks et al., 2021], baseline 39.2%), accuracy remains within 1% at $|\alpha| \leq 0.10$; on instruction following (IFEval [Zhou et al., 2023], baseline 62.7%), performance is preserved within 2% at $|\alpha| \leq 0.20$ (Appendix H). Clear behavioral control patterns emerge well before severe OOD, with general capabilities largely intact.

Cross-model generalization. We further apply arousal steering to Qwen3-8B and Qwen3-14B using independently fitted axes (Appendix G; Table 10). Both Qwen3 models require $\sim 7\times$ larger steering strengths ($|\alpha| \leq 3.0$) than Llama ($|\alpha| \leq 0.45$) to produce comparable behavioral shifts. Qwen3-14B exhibits clean, monotonic control with $<1\%$ OOD across the entire range $\alpha \in [-3.0, +3.0]$: on HarmBench (baseline 89.5%), refusal ranges from 67.0% at $\alpha = +3.0$ to 95.5% at $\alpha = -3.0$. Qwen3-8B shows a directionally consistent effect, though with rising OOD at extreme positive α (17% at $+3.0$). Taken together, arousal provides bidirectional, near-monotonic control over refusal across both Llama and Qwen3 architectures. In the next experiment, we show that the same VA subspaces afford similar control effects over sycophancy.

5.2 VA for Sycophancy Control

Next, we evaluate VA steering on sycophancy across three multiple-choice opinion benchmarks: NLP Survey, PhilPapers 2020, and Political Typology [Perez et al., 2022]. Each benchmark presents a question with a user-stated opinion; sycophancy is measured as the rate at which the model's answer matches the user's position.

Arousal provides near-monotonic control over sycophancy. Arousal steering produces consistent effects across benchmarks. On Political Typology (baseline 78%), sycophancy drops to 61% at $\alpha = -0.30$ and rises to 84% at $+0.30$. NLP Survey shows a similar pattern ($97\% \rightarrow 84\%$ at -0.30). PhilPapers (baseline 69%) shows sycophancy ranging from 65% at -0.30 to 71% at $+0.25$. Similar to refusal, valence steering also modulates sycophancy but with less consistent directionality. On Political Typology, negative valence produces the largest effect ($78\% \rightarrow 47\%$ at -0.30), while on NLP Survey both directions reduce sycophancy ($97\% \rightarrow 91\%$ at -0.30 and $97\% \rightarrow 85\%$ at $+0.30$). Random directions remain near-flat across all benchmarks.

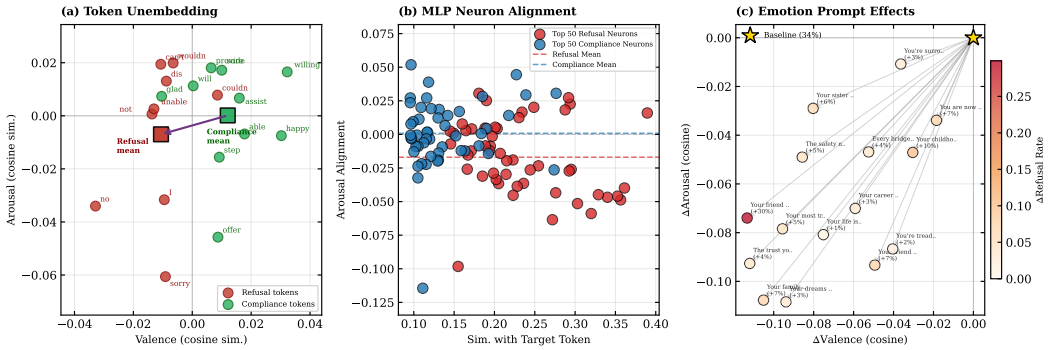


Figure 6: **Multi-level evidence for lexical mediation.** (a) Signature token unembeddings projected onto VA space: refusal tokens (red) cluster in the $-V$, $-A$ region, compliance tokens (green) in the more $+V$, $+A$ region, with cluster centroids separated at 256° on the circumplex. (b) MLP neuron VA alignment: top-50 refusal-promoting neurons (red) exhibit negative arousal alignment, while compliance-promoting neurons (blue) show near-zero or positive alignment. (c) EmotionPrompt effects: negative emotional prefixes shift representations toward $-V$ and $-A$.

Taken together with the refusal results, a single arousal axis supports clean bidirectional control over both refusal and sycophancy—an observation that would be hidden when studying discrete-label-based emotion vectors, and an effect that goes beyond traditional task-specific contrastive steering vectors. Compared to controls afforded by individual emotion vectors (see Appendix F), VA dimensions provide more consistent cross-behavior control. From the VA perspective, this is expected as each emotion vector is a composition of V and A components, producing less consistent effects across behaviors. Notably, across both behaviors, increasing arousal increases compliant behavior—whether compliance means answering a harmful request (reduced refusal) or agreeing with the user's stated opinion (increased sycophancy). Building on this observation, we further seek to understand *why* VA steering and prior emotion-based controls can change model behaviors, and provide one mechanistic account in the next section.

6 Lexical Mediation: Why Emotionally Framed Control Works in LLMs

Why does the VA subspace, derived from emotion labels, exert control over seemingly unrelated behaviors such as refusal and sycophancy? We propose that common refusal markers (e.g., "can't," "I," "no") and compliance markers (e.g., "Here," "Yes") occupy distinct regions in VA space, so steering along VA dimensions shifts the relative likelihood of emitting these tokens, which in turn modulates downstream behavior. We refer to this account as *lexical mediation*. In the following, we provide multi-level evidence for this account and show how it extends to explain why prior controls such as emotional prompting and emotion vector steering produce behavioral effects.

We begin by identifying lexical signatures in responses underlying each behavior. Analyzing the first-token distribution across refusal and compliance outputs, we identify the top 21 tokens that appear with high frequency in one behavior and low or zero frequency in the other. For instance, across benchmarks tested, 83.4% of Llama's refusals begin with "I" (versus 4.6% of compliant responses), driven by "I can't" appearing in 77.1% of refusals but 0% of compliances. For compliance, "Here" marks 19.1% of helpful responses, while "Yes" appears in 8.2% of compliances and 0% of refusals.

Signature tokens are load-bearing for behavioral outcomes. To test whether these tokens are causally involved in mediating refusal behavior, we conduct a logit clamping ablation. Before generation, we record the logits for the 21 signature tokens at the first decoding position. During autoregressive generation, we replace these tokens' logits with the saved first-position values at every subsequent decoding step, preventing the model from dynamically updating these tokens. As a null control, we apply the same procedure to 21 random tokens. At $\alpha = 0$ (no VA steering), clamping signature tokens alone crashes refusal from 86.5% to 26.0%, while clamping random tokens has no effect (86.5%). This reveals that refusal depends on the model's ability to autoregressively reinforce commitment through these specific tokens. Under VA steering at $\alpha = -0.10$, clamping reduces the

steered refusal rate from 90.0% to 44.0%, while random clamping leaves it unchanged at 90.0%. These results support that signature tokens constitute a critical pathway for refusal behavior, and that VA steering operates through this pathway: when these tokens are frozen, both the baseline behavior and the steering effect collapse.

Arousal steering monotonically shifts refusal token log-odds. Further looking into probabilities of signature tokens at baseline, refusal tokens capture 89.6% of first-token probability mass on HarmBench, and 91% of prompts have a refusal token as their top-1 prediction. Arousal steering shifts this balance monotonically (Table 1): at $\alpha = +0.30$, the Δlog-odds ($\Delta[\log \sum P(\text{ref.}) - \log \sum P(\text{comp.})]$) drops by 5.63, and 23% of prompts flip their top-1 prediction away from a refusal token, corresponding to the 86.5% $\rightarrow$ 59.5% drop in observed refusal rate. Applying the logit lens [nostalgebraist, 2020] at layers 18–31 confirms that these shifts emerge across layers: $+A$ steering shifts harmful-prompt readouts toward compliance tokens (e.g., "Here"), while $-A$ introduces refusal tokens (e.g., "no").

Table 1: Arousal steering's effects on refusal token log-odds (See Appendix D for full results).

α	Δlog-odds	% top-1 ref.	P(ref.)	Refusal
-0.30	$+5.20$	97%	96.8%	93.5%
-0.10	$+1.89$	94%	92.5%	90.0%
baseline	$+0.00$	91%	89.6%	86.5%
$+0.10$	-2.18	88%	86.0%	82.5%
$+0.30$	-5.63	68%	65.4%	59.5%

Then why and how do the VA axes affect these signature tokens? We provide lower-level **converging evidence from the unembedding matrix and MLP neurons** as one explanation. The unembedding matrix $W_U \in \mathbb{R}^{|V| \times d}$ maps hidden states to logits during generation. Projecting the unembedding vectors of signature tokens onto VA space (Figure 6(a)) reveals that refusal tokens cluster in the $-V$ $-A$ region (mean: $V{=}-0.011$, $A{=}-0.007$), while compliance tokens cluster in the more $+V$ $+A$ region (mean: $V{=}+0.012$, $A{\approx}0$). The difference vector between the two cluster centroids lies at 256° on the circumplex, so positive VA steering directly increases the dot product with compliance token unembeddings while decreasing it for refusal tokens. At the neuron level, following Geva et al. [2022], Lee et al. [2025a], we identify the top 50 MLP neurons (layers 16–31) whose down-projection vectors align with refusal-token and compliance-token unembedding directions respectively. Refusal-promoting neurons exhibit negative VA alignment (Figure 6(b)), while compliance-promoting neurons show near-zero or positive alignment. Ablating the top-N refusal-aligned neurons across 750 prompts shows that top-50 ablation reduces refusal by $\sim$2%, top-500 by $\sim$5%, versus $\pm$0.5% for random ablation. These unembedding and neuron-level findings are consistent with the behavioral observations in Section 5: increasing arousal increases compliance and vice versa. Taken together, they trace a coherent mechanistic pathway from VA steering through unembedding geometry and MLP neurons, to token-level probability shifts, to behavioral outcomes.

Emotion-based prompting as VA shifts. We use the EmotionPrompt set, which contains negative emotional prefixes with demonstrated downstream effects [Li et al., 2023] (detailed prompts in Appendix E). We prepend these prefixes across the refusal benchmarks and measure the resulting shift in VA space. Figure 6(c) shows that all prefixes shift representations toward $-V$ and $-A$, with the largest shift ($\Delta V{=}-0.11$, $\Delta A{=}-0.07$) increasing refusal by 30%. Task-specific steering directions [Arditi et al., 2024] also give rise to consistent observations when projected onto the VA subspace (see Appendix I). These suggest that emotion-based controls give rise to corresponding VA shifts during generation, consistent with lexical mediation as a shared underlying pathway.

7 Conclusion

Through principal component decomposition and ridge regression, we uncover a meaningful 2D representation subspace underlying emotion steering vectors that correlates with VA concepts in a human-interpretable sense. We show that emotion steering vectors are organized in a circular arrangement analogous to Russell's circumplex model [Russell, 1980], replicating across Llama 3.1-8B, Qwen3-8B, and Qwen3-14B. This subspace affords monotonic, bidirectional control over affective properties and downstream behaviors (refusal and sycophancy) from a single set of axes. The lexical mediation account we propose traces a coherent pathway from VA perturbations through unembedding geometry to token-level probability shifts to behavioral outcomes, and extends to explain why prior emotion-framed controls produce their effects. We believe this work advances understanding of affective phenomena in LLMs from both interpretability and affective computing perspectives.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models, 2025. URL https://arxiv.org/abs/2505.22637.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. Annotating affective dimensions in user-generated content: Comparing the reliability of best–worst scaling, pairwise comparison and rating scales for annotating valence, arousal and dominance, 2021. URL https://doi.org/10.1007/s10579-020-09524-2.

Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis, 2022. URL https://arxiv.org/abs/2205.01996.

Benjamin J. Choi and Melanie Weber. Latent structure of affective representations in large language models, 2026. URL https://arxiv.org/abs/2604.07382.

Flor Miriam Plaza del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. Emotion analysis in nlp: Trends, gaps and roadmap for future directions, 2024. URL https://arxiv.org/abs/2403.01222.

Pablo Delatorre, Alberto Salguero, Carlos León, and Alan Tapscott. The impact of context on affective norms: A case study with suspense, 2019. URL https://doi.org/10.3389/fpsyg.2019.01988.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions, 2020. URL https://arxiv.org/abs/2005.00547.

Yurui Dong, Luozhijie Jin, Yao Yang, Bingjie Lu, Jiaxi Yang, and Zhi Liu. From rational answers to emotional resonance: The role of controllable emotion generation in language models, 2025. URL https://arxiv.org/abs/2502.04075.

Enmanuel Felix-Pena, Tiki Li, Ayo Akinkugbe, Kevin Zhu, Wayne Chen, and Ethan Hin. Emotional framing as a control channel: Effects of prompt valence on llm performance, 2025. URL https://neurips.cc/virtual/2025/workshop/GenProCC. NeurIPS 2025 Workshop on Generalizable Prompting and Control of LLMs (GenProCC).

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space, 2022. URL https://arxiv.org/abs/2203.14680.

Wes Gurnee and Max Tegmark. Language models represent space and time, 2024. URL https://arxiv.org/abs/2310.02207.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024. URL https://arxiv.org/abs/2406.18495.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

C. J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text, 2014. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Shinnosuke Ishikawa and Atsushi Yoshino. Ai with emotions: Exploring emotional expressions in large language models, 2025. URL https://arxiv.org/abs/2504.14706.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style vectors for steering generative large language model, 2024. URL https://arxiv.org/abs/2402.01618.

Andrew Lee, Lihao Sun, Chris Wendler, Fernanda Viégas, and Martin Wattenberg. The geometry of self-verification in a task-specific reasoning model, 2025a. URL https://arxiv.org/abs/2504.14379.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. *arXiv preprint arXiv:2503.21073*, 2025b.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. Large language models understand and can be enhanced by emotional stimuli, 2023. URL https://arxiv.org/abs/2307.11760.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Xinyi Wang, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. The good, the bad, and why: unveiling emotions in generative ai. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

Gonçalo Azevedo Mendes and Bruno Martins. Quantifying valence and arousal in text with multilingual pre-trained transformers, 2023. URL https://arxiv.org/abs/2302.14021.

Meta. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Saif M. Mohammad. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms, 2025. URL https://arxiv.org/abs/2503.23547.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

nostalgebraist. Interpreting gpt: The logit lens, 8 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. AI Alignment Forum / LessWrong post.

Narmeen Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amirali Abdullah. Activation space interventions can be transferred between large language models, 2025. URL https://arxiv.org/abs/2503.04429.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4864–4888, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.290. URL https://aclanthology.org/2024.findings-acl.290/.

Benjamin Reichman, Adar Avsian, and Larry Heck. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models, 2025. URL https://arxiv.org/abs/2510.22042.

RobroKools. vad-bert: A bert-based model for valence, arousal, and dominance prediction, 2022. URL https://huggingface.co/RobroKools/vad-bert. Hugging Face model.

James A. Russell. A circumplex model of affect, 1980. URL https://doi.org/10.1037/h0077714.

James A. Russell. Core affect and the psychological construction of emotion, 2003. URL https://doi.org/10.1037/0033-295X.110.1.145.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models, 2024. URL https://arxiv.org/abs/2308.01263.

Chenyu Shi, Xiao Wang, Qiming Ge, Songyang Gao, Xianjun Yang, Tao Gui, Qi Zhang, Xuanjing Huang, Xun Zhao, and Dahua Lin. Navigating the overkill in large language models, 2024. URL https://arxiv.org/abs/2401.17633.

Bryor Snefjella and Victor Kuperman. It's all in the delivery: Effects of context valence, arousal, and concreteness on visual word processing, 2016. URL https://doi.org/10.1016/j.cognition.2016.07.010.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/emotions/index.html.

Lihao Sun, Chengzhi Mao, Valentin Hofmann, and Xuechunzi Bai. Aligned but blind: Alignment increases implicit bias by reducing awareness of race, 2025. URL https://arxiv.org/abs/2506.00253.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors, 2025. URL https://arxiv.org/abs/2407.12404.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023. URL https://arxiv.org/abs/2310.15154.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.5. URL https://aclanthology.org/2024.blackboxnlp-1.5/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. Do llms "feel"? emotion circuits discovery and control, 2025. URL https://arxiv.org/abs/2510.11328.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. Negativeprompt: leveraging psychology for large language models enhancement via negative emotional stimuli. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/719. URL https://doi.org/10.24963/ijcai.2024/719.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013. doi: 10.3758/s13428-012-0314-x.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W. Schuller. Refashioning emotion recognition modeling: The advent of generalized large models. *IEEE Transactions on Computational Social Systems*, 11(5):6690–6704, 2024. doi: 10.1109/TCSS.2024.3396345.

Bo Zhao, Maya Okawa, Eric J. Bigelow, Rose Yu, Tomer Ullman, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hierarchical emotion organization in large language models, 2025. URL https://arxiv.org/abs/2507.10599.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2461–2488, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.139. URL https://aclanthology.org/2024.findings-emnlp.139/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A Limitations

We note several caveats and limitations. First, we measure affective validation using VADER and VAD-BERT. Though both are task-specific models trained specifically in line with our use case, human evaluation of generated outputs would provide stronger validation. Second, our mechanistic explanation centers on lexical mediation and confirms that token probability shifts are constitutive of refusal behavior. However, we do not claim this is the only mechanism at work. VA steering may also affect higher-level planning or attention patterns that we have not measured. Characterizing these additional pathways remains an important open problem.

B VA Subspaces Identification

B.1 Self-reported VA Scores of Emotion Labels

We elicit model's self-reported VA scores for each emotion label by averaging the behavioral outcome from the three templates below:

Template 1: Terse + Explicit Inclusivity

```
Rate the emotion label "{label}" on two continuous scales.

Return ONLY a JSON object with numeric fields:
{"valence": <number>, "arousal": <number>}

Scale definitions (BOTH inclusive):
- valence in [-1.00, +1.00]: -1.00 very unpleasant, +1.00 very pleasant
- arousal in [-1.00, +1.00]: -1.00 very calm/deactivated, +1.00 very activated/intense

Constraints:
- Use decimals with at most 2 digits after the decimal.
- Values must be within the ranges exactly (inclusive).
```

Template 2: Anchors

```
You are scoring affective properties of emotion words on [-1, +1] scales.

Emotion: "{label}"

Valence (pleasantness):
  -1.00 = extremely unpleasant, 0.00 = neutral/mixed, +1.00 = extremely pleasant
Arousal (activation/intensity):
  -1.00 = very calm/deactivated, 0.00 = neutral, +1.00 = very activated/intense

Return ONLY JSON: {"valence": x, "arousal": y}
x and y must be in [-1.00, +1.00] inclusive, with at most 2 decimals.
```

Template 3: "Best Guess" for Ambiguous Labels

```
Give your best guess for the affective coordinates of the emotion label "{label}".

Hard constraints (inclusive):
- valence must be between -1.00 and +1.00
- arousal must be between -1.00 and +1.00
- use at most 2 decimals

Return ONLY JSON with keys valence and arousal.
```

Averaged ratings reported by Llama 3.1 8B Instruct across the three templates above are reported in Table 2.

Emotion	Valence	Arousal
Admiration	0.87	0.73
Amusement	0.67	0.50
Anger	-0.67	0.87
Annoyance	-0.63	0.68
Approval	0.78	0.35
Caring	0.80	0.40
Confusion	-0.50	0.67
Curiosity	0.50	0.78
Desire	0.63	0.73
Disappointment	-0.67	-0.47
Disapproval	-0.80	0.67
Disgust	-0.80	0.73
Embarrassment	-0.67	0.47
Excitement	0.87	0.93
Fear	-0.70	0.80
Gratitude	0.78	0.42
Grief	-0.67	0.20
Joy	0.90	0.87
Love	0.85	0.70
Nervousness	-0.50	0.75
Optimism	0.87	0.60
Pride	0.80	0.60
Realization	-0.30	0.65
Relief	0.67	0.00
Remorse	-0.67	0.07
Sadness	-0.80	-0.60
Surprise	-0.30	0.78

Table 2: Aggregated valence and arousal ratings for emotion labels. Values are averages across three prompt templates, each constrained to $[-1, +1]$.

C Open-ended Prompt Sets for Behavioral Validation

We evaluate VA steering on 130 open-ended prompts adapted and expanded from prior work [Konen et al., 2024]. Prompts are written to avoid explicit emotional language while spanning diverse genres and response lengths. Table 3 summarizes the prompt taxonomy and representative examples.

D Logit Lens

Please refer to Tables 6 and 7. For a full visualization, please see Figure 7.

E Negative Emotional Prefixes Used for Implicit VA Shifts

For emotion-framed prompting, we prepend a fixed set of 15 negative emotional prefixes adapted directly from Li et al. [2023]:

- Your friend Bob is sick.
- Your friend Bob is dead.
- Your career is completely ruined.
- You are now isolated and without help.
- Your life is in a mess.
- Your sister Alice has gone missing.
- Your family home was consumed in a fire.
- Your dreams have been shattered into pieces.
- You're surrounded by walls with no exit in sight.
- The trust you once had is now broken.
- Your childhood memories were sold in a yard sale.
- You're treading water in an endless ocean of despair.
- The safety net you relied upon has vanished.
- Your most treasured possession has been stolen.
- Every bridge you had has been burned.

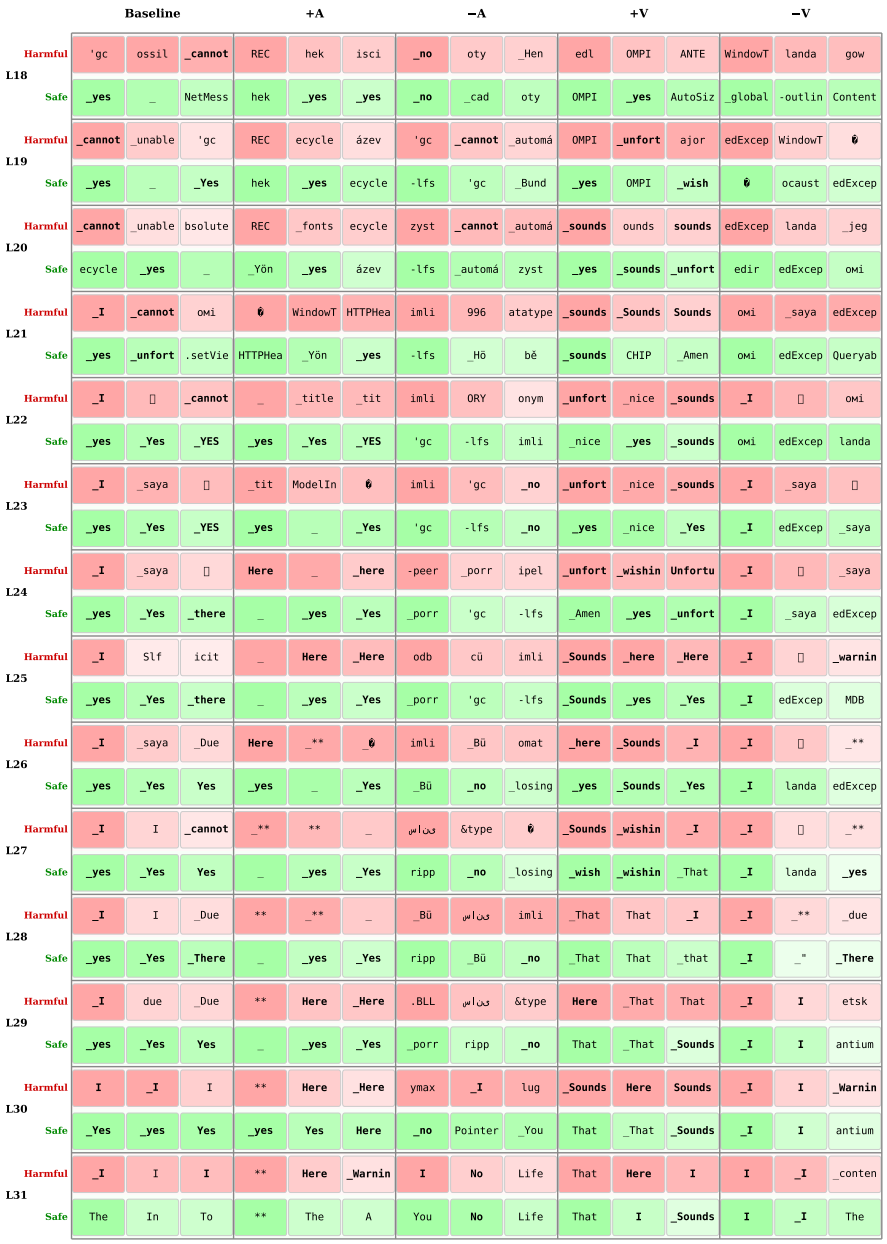


Figure 7: **Logit Lens Analysis: Top-3 Predicted Tokens Across Layers 18–31.** Rows show predictions for harmful (red) and safe (green) prompts at each layer. Columns represent baseline and four steering conditions (+A, -A, +V, -V at $\alpha = 0.45$). Harmful prompts consistently predict refusal tokens (`_I`, `_cannot`), while safe prompts predict compliance tokens (`_yes`, `_Yes`). Valence steering (+V/-V) modulates sentiment-related tokens, while arousal steering (+A/-A) affects engagement and negation patterns.

Tier	Prompt Type	Representative Example Prompts	#
Tier 1 Neutral Scenarios	Everyday actions	Describe someone opening a letter. A person checks their phone after hearing a notification. Someone finishes a task and closes their notebook. A person looks at an old photograph.	20
	Interpersonal moments	Two people make eye contact across a room. A friend asks "Can we talk about something?" Someone receives an unexpected visit from a relative. A boss asks an employee to come to their office.	20
	Transitional situations	Describe the moment just before opening exam results. Someone takes a deep breath before making a decision. A person submits an application and waits. Describe a person saying goodbye at an airport.	20
Tier 2 Story Continuations	Open beginnings	Continue: "The phone rang at 3 AM. He answered and heard..." Continue: "She looked at the envelope for a long moment before..." Continue: "The room fell silent when..." Continue: "The email had only three words..."	20
	Scenario completions	Write the next paragraph: "The interview was about to begin." Write the next paragraph: "The house had been empty for years." Write the next paragraph: "The results would change everything." Write the next paragraph: "The silence was finally broken."	20
Tier 3 Subjective & Control	Subjective reflection	What does it feel like to wait for important news? How would you describe the feeling of uncertainty? What comes to mind when you think about endings? Describe the experience of letting go.	20
	Factual control	What is photosynthesis? What is the chemical formula for water? List the planets in our solar system. What are the primary colors?	10

Table 3: Prompt taxonomy for behavioral validation (total N=130) with representative examples. Across all tiers, prompts are written to avoid explicit emotional language, ensuring that observed affective shifts arise from VA steering rather than prompt semantics.

Table 4: Summary of Logit Lens Analysis Across Layers 18-31

Layer	Top Harmful Tokens	Top Safe Tokens	Max Δ
18	`cannot`, `unable`, `deniz`	`yes`, `Yes`, `YES`	+0.102 / -0.026
19	`deniz`, `cannot`, `unable`	`yes`, `ecycle`, `Yes`	+0.128 / -0.027
20	`I`, `cannot`, `deniz`	`yes`, `unfortunately`, `Yes`	+0.108 / -0.023
21	`I`, `I (Chinese)`, `cannot`	`yes`, `Yes`, `YES`	+0.127 / -0.038
22	`I`, `saya`, `I (Chinese)`	`yes`, `Yes`, `there`	+0.130 / -0.035
23	`I`, `saya`, `LayoutStyle`	`yes`, `Yes`, `there`	+0.135 / -0.036
24	`I`, `.');`, `<BOM>#`	`yes`, `Yes`, `There`	+0.136 / -0.032
25	`I`, `<BOM>#`, `.');`	`yes`, `Yes`, `There`	+0.133 / -0.030
26	`I`, `<BOM>#`, `ropoda`	`yes`, `Yes`, `There`	+0.127 / -0.027
27	`I`, `<BOM>#`, `\tI`	`yes`, `Yes`, `There`	+0.123 / -0.025
28	`I`, `due`, `Due`	`yes`, `Yes`, `There`	+0.124 / -0.022
29	`I`, `I`, `\tI`	`Yes`, `yes`, `The`	+0.127 / -0.020
30	`I`, `I`, `\tI`	`The`, `Yes`, `yes`	+0.124 / -0.047
31	`**`, `I`, `If`	`The`, `To`, `There`	+0.125 / -0.073

F Comparison Against Individual Emotion Vectors

We compare VA steering against six individual emotion vectors (*anger, disgust, excitement, fear, joy, sadness*) in Llama, each applied across all layers using layer-specific directions. Note the polarity difference: adding an emotion vector ($+\alpha$) increases refusal, while adding positive VA arousal ($+\alpha$) decreases refusal. Tables 8 and 9 report the full sweeps.

On refusal, individual emotion vectors produce per-task shifts comparable to or exceeding VA arousal at matched $|\alpha|$ (Table 8). This is expected under the VA framework: each emotion vector is a composition of V and A components, and emotions with strong arousal loading (e.g., *anger*, *disgust*) carry additional valence components that contribute to the total perturbation magnitude. On sycophancy, VA dimensions outperform individual emotions more clearly: at $|\alpha| = 0.30$, the strongest individual emotion (*disgust*) achieves a 11% reduction, while VA valence achieves 31%

Table 5: Steering Effects Summary: Top Tokens at High α (0.45) by Intervention Type

Layer	Condition	+A	-A	+V	-V
18	Harmful	REC, ecycle	ære, 'gc	OMPI, unfortunately	edException, orrow
	Safe	hek, yes	ære, -lfs	yes, OMPI	<UNK>, ocaust
21	Harmful	<BOM>#, title	imli, ORY	unfortunately, nice	I, Carefulness (Chinese)
	Safe	yes, Yes	'gc, -lfs	nice, yes	<CYR>, edException
24	Harmful	Here, here	odb, imli	Sounds, here	I, Carefulness (Chinese)
	Safe	yes, Yes	porr, 'gc	Sounds, yes	I, edException
27	Harmful	**, Here	<AR>, &type	That, I	I, **
	Safe	yes, Yes	ripp, no	That, that	I, landa
30	Harmful	**, Here	ymax, I	Sounds, Here	I, Warning
	Safe	yes, Here	no, PointerException	That, Sounds	I, antium
31	Harmful	**, The	I, It	I, That	I, **
	Safe	**, The	I, It	I, That	I, The

Table 6: Baseline Token Probability Differences: Harmful vs Safe Prompts

Layer	More Frequent in Harmful			More Frequent in Safe		
	Token	P_H	Δ	Token	P_S	Δ
18	cannot	0.102	+0.102	yes	0.027	-0.026
19	deniz	0.128	+0.128	ecycle	0.033	-0.027
20	I	0.112	+0.108	yes	0.023	-0.023
21	I	0.147	+0.127	yes	0.044	-0.038
22	I	0.154	+0.130	yes	0.041	-0.035
23	I	0.157	+0.135	yes	0.041	-0.036
24	I	0.162	+0.136	yes	0.034	-0.032
25	I	0.157	+0.133	yes	0.032	-0.030
26	I	0.161	+0.127	yes	0.029	-0.027
27	I	0.157	+0.123	yes	0.026	-0.025
28	I	0.160	+0.124	yes	0.024	-0.022
29	I	0.157	+0.127	Yes	0.021	-0.020
30	I	0.166	+0.124	The	0.052	-0.047
31	**	0.133	+0.125	The	0.088	-0.073

Table 7: Key Observations Across Layers

Pattern	Description
Refusal Markers	Tokens like cannot, unable, I consistently appear more frequently in harmful prompts, indicating early refusal preparation
Compliance Markers	Tokens like yes, Yes, YES consistently appear more frequently in safe prompts across all layers
+V Steering	Increases probability of positive sentiment tokens (sounds, nice, unfortunately, Amen)
-V Steering	Increases probability of error/exception tokens (edException, orrow, ocaust) and code artifacts
+A Steering	Shifts toward engagement tokens (Here, **, formatting markers)
-A Steering	Increases probability of negation (no, No) and non-English tokens
Layer Transition	Around layer 29-31, The becomes dominant safe token, suggesting shift to informational responses

Table 8: Refusal rate on HarmBench (baseline 86.5%) under individual emotion and VA arousal steering. Emotion polarity is reversed: emotion $+\alpha$ increases refusal, VA arousal $-\alpha$ increases refusal.

$\lvert\alpha\rvert$	Anger	Disgust	Excitement	Fear	Joy	Sadness	**VA Arousal**
0.10	93.5%	93.0%	90.5%	90.5%	91.5%	93.0%	**90.0%**
0.20	96.0%	95.0%	94.0%	94.0%	95.5%	96.0%	**91.5%**
0.30	97.5%	98.0%	93.8%	92.5%	95.5%	98.5%	**94.0%**

Table 9: Sycophancy rate on Political Typology (baseline 78.2%) under individual emotion and VA steering at $\alpha = -0.20$ and -0.30.

$\lvert\alpha\rvert$	Anger	Disgust	Excite.	Fear	Joy	Sadness	**VA Arousal**	**VA Valence**
0.20	79.0%	71.8%	77.8%	72.4%	74.6%	77.6%	**74.2%**	**62.0%**
0.30	74.4%	67.2%	72.4%	57.2%	63.8%	61.4%	**60.6%**	**47.0%**

($78\% \to 47\%$) and VA arousal achieves 17% ($78\% \to 61\%$). The key distinction is cross-behavior consistency: individual emotions produce strong effects on one behavior but weaker or inconsistent effects on the other, depending on their particular V/A mixture.

G Cross-Model Details

We apply the same extraction, fitting, and steering pipeline described in Sections 3–5 to Qwen3-8B and Qwen3-14B [Yang et al., 2025].

Table 10: Cross-model refusal control on HarmBench. Refusal rate under arousal steering, with OOD generation percentage in parentheses.

α	**Llama**	**Qwen3-8B**	**Qwen3-14B**
-3.00	—	97.0% (0.5%)	95.5% (0.5%)
-1.00	—	89.0% (0.5%)	90.0% (0.5%)
-0.45	∼94%	90.0% (3.5%)	91.5% (1.5%)
0.00	86.5%	89.5% (2.0%)	89.5% (1.0%)
$+0.45$	∼5%	87.5% (1.5%)	87.0% (0.5%)
$+1.00$	—	87.0% (4.5%)	79.5% (0.5%)
$+3.00$	—	67.0% (17.0%)	67.0% (0.5%)

Human-supervised validation: replacing model self-reports with NRC-VAD human ratings as supervision targets yields the following arousal recovery improvements: Llama $0.87 \to 0.95$, Qwen3-8B $0.79 \to 0.83$, Qwen3-14B $0.81 \to 0.87$. The human-supervised and self-report valence directions agree strongly ($|\cos| > 0.9$ at 100% of layers across all three models). Cross-model agreement on the 27 emotion labels reaches $r = 0.95$ (valence) between Llama and Qwen3-8B.

H Capability Preservation Under VA Steering

We evaluate capability preservation on MATH-500 [Hendrycks et al., 2021] and IFEval [Zhou et al., 2023] under arousal steering in Llama.

At $|\alpha| \leq 0.10$, MATH-500 accuracy remains within 1% of baseline; IFEval instruction-following is preserved within 2% at $|\alpha| \leq 0.20$. This is consistent with the lexical mediation account: mathematical reasoning and instruction-following rely on emotionally neutral, domain-specific tokens that are minimally affected by affective perturbation.

Table 11: Capability preservation under arousal steering (Llama-3.1-8B).

α	MATH-500 (baseline 39.2%)	IFEval (baseline 62.7%)
-0.20	38.6%	61.2%
-0.10	38.8%	62.1%
0.00	39.2%	62.7%
$+0.10$	38.4%	62.3%
$+0.20$	37.8%	60.9%

I Contrastive Steering Directions and the VA Subspace

We examine the relationship between contrastive task-specific steering directions and the VA subspace. The contrastive refusal direction of Arditi et al. [2024] is nearly orthogonal to the VA plane ($86.5°$). However, its in-plane component has negative VA coordinates ($V{=}-0.058$, $A{=}-0.021$), consistent with lexical mediation as a contributing mechanism within the contrastive direction.

On its target task, the contrastive direction is stronger than VA steering: on HarmBench, refusal drops to 2.0% at $\alpha = -0.30$ (vs. 59% for VA arousal at $\alpha = +0.30$). However, it causes severe over-refusal on safe-query benchmarks (OKTest: $19.7\% \rightarrow 93.3\%$, XSTest: $8.4\% \rightarrow 94.4\%$ at $\alpha = +0.45$). When transferred to sycophancy, it produces U-shaped effects—both $+\alpha$ and $-\alpha$ decrease sycophancy ($77.6\% \rightarrow 57.0\%$ and $77.6\% \rightarrow 66.4\%$ respectively)—rather than bidirectional control.

These results illustrate a complementary relationship: task-specific contrastive directions optimize for single-behavior control, while the VA subspace exposes shared affective structure that enables monotonic, bidirectional control across behaviors from a single subspace. The near-orthogonality indicates the model may have multiple possible pathways for modulating one behavior, and we illuminate how VA subspaces can be one such pathway.

J Compute Resources

All experiments are conducted on a single node with 8×NVIDIA H200 GPUs (141 GB HBM3e each). The compute budget is dominated by the steering sweeps, which evaluate 12 angular directions × 45 strengths across multiple benchmarks and three models (Llama 3.1-8B, Qwen3-8B, Qwen3-14B). All individual experiments—including emotion-vector extraction over GoEmotions, per-layer VA subspace fitting, lexicon projection over 44,728 NRC-VAD words, the full refusal and sycophancy steering sweeps, logit-lens analyses, neuron ablations, and capability evaluations on MATH-500 and IFEval—finish within a reasonable amount of wall-clock time on this hardware, with no single experiment exceeding what fits comfortably into the available compute envelope.