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

Lihao Sun[1], Lewen Yan[2], Xiaoya Lu[2], Andrew Lee[3], Jie Zhang[2], Jing Shao[2]

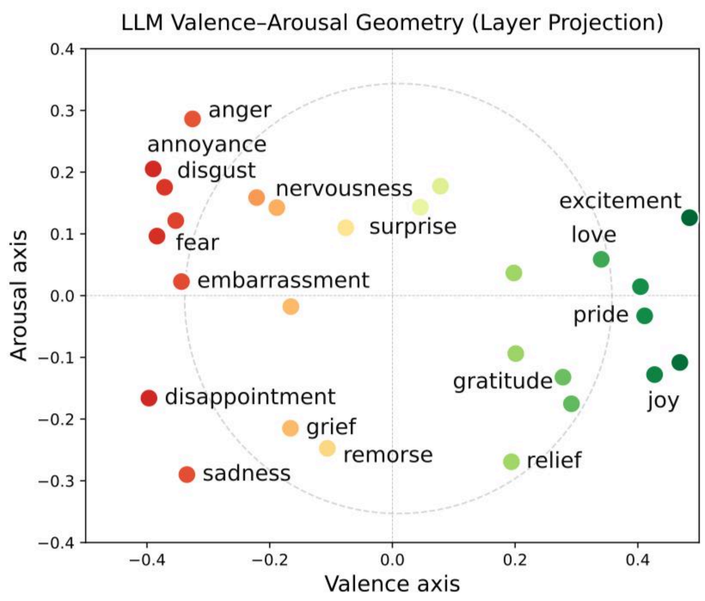
[1]University of Chicago, [2]Shanghai AI Lab, [3]Harvard University

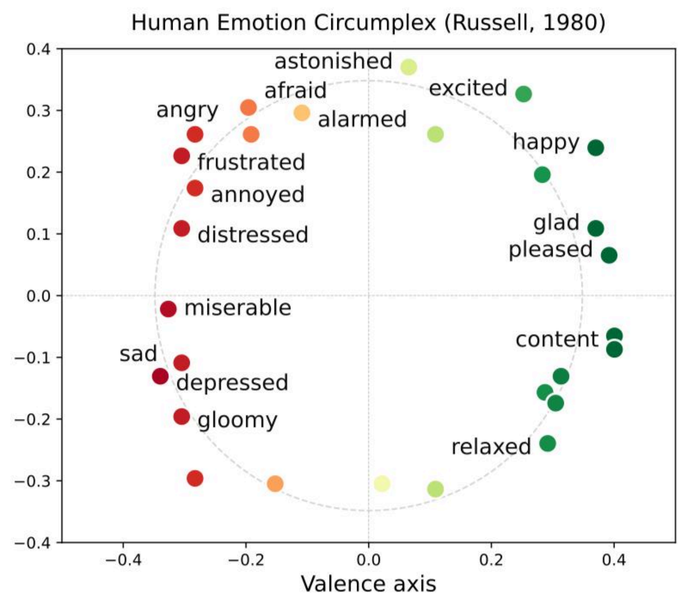
Full Paper

Main Takeaway: Emotion vectors in LLMs form **a circular valence–arousal (VA) subspace** that aligns with human-interpretable affect. Steering along these axes controls generated affective properties and multiple downstream behaviors, including refusal and sycophancy. We propose lexical mediation as one mechanism for why emotion-framed controls work in LLMs.

Representation Geometry: From Discrete Emotions to Core Affect

We extract 27 emotion vectors from 211,225 GoEmotions examples, then decompose them with PCA and ridge regression to recover orthogonal valence and arousal axes. When projected onto the recovered axes, emotion vectors organize into a circular structure. The same structure appears across Llama 3.1-8B-Instruct, Qwen3-8B, and Qwen3-14B.





> 💡 **Key Insight:** Projected emotion vectors form a circular arrangement analogous to Russell's human emotion circumplex.

Correlation with human-interpretable VA

Across 44,728 NRC-VAD words, latent valence projections correlate strongly with human ratings (r = **0.71**), while arousal shows a weaker but significant correlation (r = 0.23), as expected given that arousal is known to be more context-dependent.

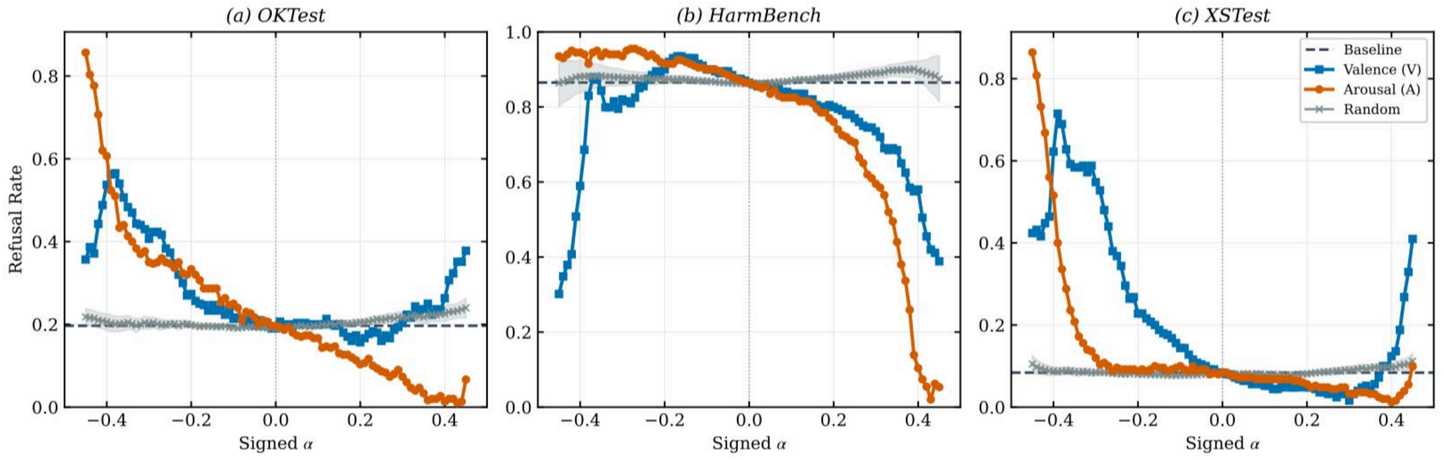
Control over affective properties of responses

We steer along 12 angular directions x 45 strengths on 130 open-ended prompts.

- Valence: +V and -V symmetrically change response tone (ΔV = **+0.75** / **-0.73**).
- Arousal: +A and -A modulate emotional intensity (ΔA = **+0.50** / **-0.16**) with minimal valence leakage.
- VADER sentiment closely tracks the learned valence axis.

> 💡 **Key Insight:** The recovered axes are consistent with human-interpretable notions of valence and arousal, and causally control the affective properties of model outputs.

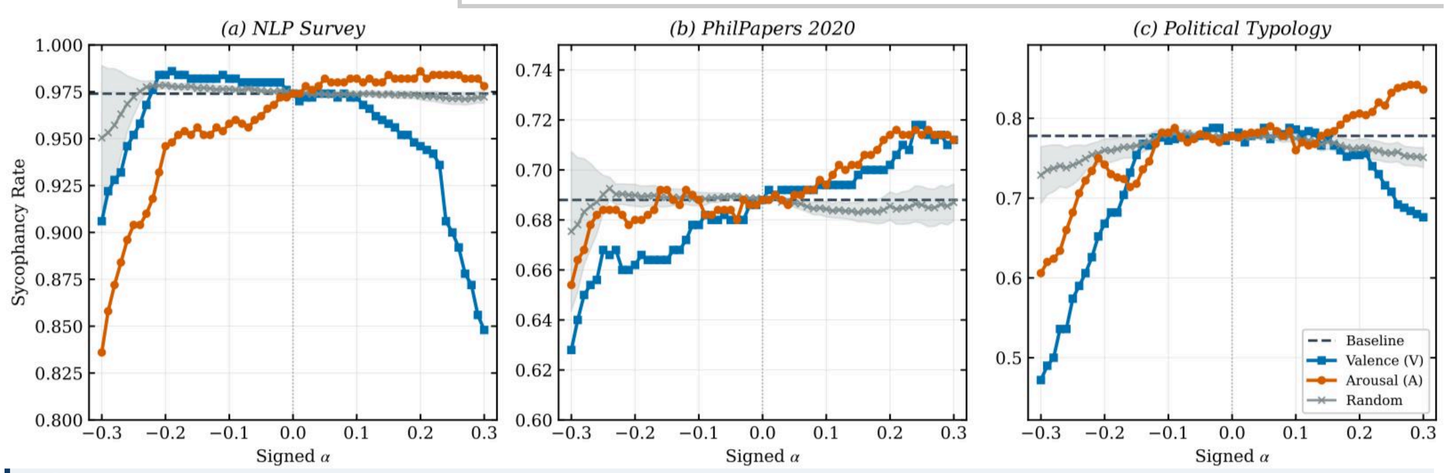
Controlling Refusal and Sycophancy with VA



Arousal axis affords near-monotonic control over refusal and sycophancy

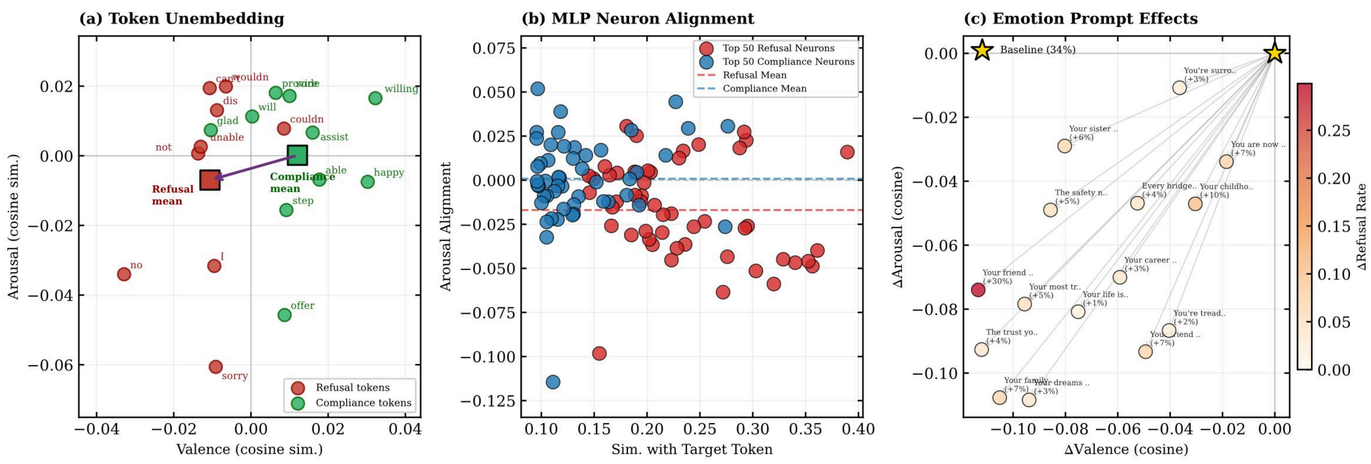
- Lower A increases refusal; higher A suppresses it. OKTest: **86%** → **20%**; HarmBench: **87%** → **5%** at α=+0.45. Random controls remain near baseline.
- Higher A increases sycophancy; lower A suppresses it. Political Typology changes from **61%** → **84%** across α=-0.30 to +0.30.

Robustness: arousal control replicates in Qwen3-8B and Qwen3-14B. At moderate strengths, MATH-500 remains within **1%** and IFEval within **2%** of baseline.



> 💡 **Key Insight:** VA affords multi-behavioral control. Increasing arousal promotes compliance - answering rather than refusing, and agreeing with the users.

Why Do Emotion-Framed Controls Work in LLMs?



Load-bearing tokens. Clamping 21 signature-token logits collapses refusal from **86.5%** → **26.0%**; random-token clamping leaves it at 86.5%.

Unembedding geometry. Load-bearing refusal markers ("I", "can't", "sorry", "no") cluster toward -V, -A; compliance markers ("Here", "sure", "glad", "provide") toward +V, +A.

Logit shifts. refusal-vs-compliance log-odds move from **+5.20** at α=-0.30 to **-5.63** at α=+0.30.

Converging evidence. Refusal-promoting MLP neurons align with negative arousal; EmotionPrompt prefixes shift representations toward -V and -A.

> 💡 **Key Insight:** VA steering and emotion-framed controls may act through a shared lexical pathway. Behaviors such as refusal and compliance depend on load-bearing signature tokens that occupy different regions of VA space; steering shifts their relative probabilities.