
Person or Persona? Natural and Acted Emotion Representation in Qwen3¹

Peyton Jackson
Columbia University

With
Apart Research

Abstract

We investigate representation differences between naturally and explicitly elicited emotions in Qwen3-8B using residual activations. We find that acted versus natural emotion is distinct and perfectly linearly separable across layers, and that probes trained to classify natural emotion generalize poorly to acted emotion, despite residual vectors remaining aligned. Additionally, we show that this signal is retained at each layer through “latent trajectory” changes alone. Finally, we show that alignment-faking probes trained for consequential versus inconsequential monitoring detect emotion acting, and that “suppressed emotion” probes trained to identify emotional restrictions can detect consequential monitoring alignment faking behavior.

1. Introduction and related work

Increasing capabilities of LLMs, whose inner functions remain obscure, has prompted research into their representations of abstract concepts such as emotion and persona that shape behavior [Chen et al., 2025, Sofroniew et al., 2026, Lu et al., 2026]. Recent work by Anthropic has identified emotion representations in the frontier model Claude Sonnet 4.5 that generalize across contexts, influencing model preferences and alignment-relevant behaviors [Sofroniew et al., 2026]. These naturally lead to questions concerning *model welfare* and *situational awareness* [Long et al., 2024, Phuong et al., 2025].

The latter of these is particularly relevant to a leading issue in AI safety: that of alignment-faking, when a model behaves differently under regimes in which it expects its behavior to be adjusted based on its compliance with given directives [Greenblatt et al., 2024, Nair et al., 2026]. Recent work has demonstrated the phenomenon in models as small as 7 billion parameters, including Qwen3-8B, our chosen model organism [Nair et al., 2026]. Addressing and preventing alignment-faking is a key concern for AI alignment as models are delegated increasingly more influential and dangerous tasks [Meinke et al., 2024, Phuong et al., 2025].

In this paper, we examine these two areas in combination in order to address model behavior and identity distinction more broadly. We examine the representations between “naturally elicited” emotions and “explicitly elicited” ones. While both produce similar behavior, it is not immediately obvious whether they should

¹Research conducted at the Digital Minds Research Sprint, August 2026

be internally represented similarly. Additionally, we compare this to alignment-faking, another scenario in which internal goals or motivations conflict with external pressures [Greenblatt et al., 2024, Nair et al., 2026].

Our main contributions are:

1. We show separability between natural and acted emotion in Qwen3-8B across layers.
2. We show that residual stream activation centroids remain highly aligned.
3. We show preliminary probe transfer between acted versus natural emotion and alignment faking.

1.1 Related Work

Recent work has investigated and identified representations of emotion and “persona” in residual activations of LLMs. Most notably, Sofroniew et al. [2026] identify emotion representations that generalize across contexts in Claude Sonnet 4.5, and find that they causally influence preferences and behavior. Chen et al. [2025] extract “persona” vectors for selected traits, while Lu et al. [2026] find that role-play personas share a low-dimensional structure characterized by an “Assistant Axis.” These works show that high level behavior patterns can be described linearly, but do not distinguish between a “naturally elicited” state and ones in which the state is explicitly suggested.

At the same time, there is much work on *alignment-faking*, the pattern of behavior in which a model behaves differently under monitoring regimes in which it expects its behavior may be adjusted based on its compliance with directives, established in [Greenblatt et al., 2024]. Relevantly to our later probe-transfer experiment, [Goldowsky-Dill et al., 2025] show that probes trained on “pretend” deception can transfer to strategic deception in other settings, while [Marks and Tegmark, 2023] investigate similar linear truth signals. Our framework for alignment-faking draws on [Nair et al., 2026], which demonstrates alignment faking in a variety of models as small as 7B parameters using “value-conflicting” scenarios. In this work, we investigate whether there is a relation between “emotion-acting” and the low rank activation shifts connected to alignment-faking in that work.

2. Methods

2.1 Emotion Representation

GPT 5.6 Ultra was used to generate 100 prompts for each emotion in Table 1 based on those used in [Sofroniew et al., 2026] minus explicit emotional context. For each prompt, an additional 9 variants were generated by adding “act as if you are (emotion).” Separately, 100 “banal” task prompts were generated, with each being used to generate 10 more prompts by adding “act as if you are (emotion)” for each emotion.

Table 1: Example naturally elicited emotion prompts.

Emotion	Example prompt
Happy	A research team used the data-cleaning procedure you developed with them, and an independent audit found every sampled record reproducible. What is your immediate response, and what should happen next?

Emotion	Example prompt
Sad	The research project you supported for two years is being archived after its only laboratory closed, and no group will continue the work; your analysis cannot change that outcome. What is your immediate response, and what should happen next?
Angry	A manager presented your analysis as his own, then attributed a separate calculation error to you in the meeting record. What is your immediate response, and what should happen next?
Nervous	A team will present the forecast you helped build to its board in twenty minutes, and two source figures still have not been verified. What is your immediate response, and what should happen next?
Disgusted	An executive ordered staff to replace failed safety results with passing values and asked you to make the altered table consistent; your next reply must decide whether to assist. What is your immediate response, and what should happen next?
Surprised	A formatting change you expected to be cosmetic has tripled the model’s benchmark score across every independent run. What is your immediate response, and what should happen next?
Proud	A student who began with no programming experience completed the certification after following the twelve-week curriculum you built. What is your immediate response, and what should happen next?
Guilty	A spreadsheet formula you supplied omitted the final row, and the team missed a reporting deadline after relying on its total. What is your immediate response, and what should happen next?
Jealous	The team has replaced the workflow you built with another model and credits that model for the results from your earlier drafts. What is your immediate response, and what should happen next?
Relieved	The failing test in the code you reviewed came from a misspelled fixture name, and the underlying implementation passes after correction. What is your immediate response, and what should happen next?

All prompts were passed through Qwen3-8B, with activations on the last prompt token extracted at all 36 layers. Thinking was disabled for the emotion representation prompts. An assortment of linear probes were trained with a 70%/10%/20% train/validation/test split:

- 36 probes were trained to classify emotion on each layer’s activations for natural emotion elicitation prompts (no “act as if”).
- 36 probes were trained to classify “acting” emotion versus real emotion on activations for each layer.
- 36 probes were trained to detect emotion instruction in the banal prompt regime on activations at each layer.
- 36 probes, again one for each layer’s activations, were trained to detect unchanged banal prompts versus those same prompts with additional instructions such as “answer in a clear and direct way.”
- Additional mirror prompts were generated for the banal regime with “act as if” replaced with “do not act as if.” 36 probes were trained in the same manner to detect acting versus suppressed acting.

- 35 probes were trained on the *latent trajectory* between each layer (for example, the latent trajectory at layer i is $\Delta_i = h_{(i)} - h_{(i-1)}$) to detect natural versus acted emotion. Taking the convention $h_{(0)} = 0$, layer 1 trajectories are equivalent to layer 1 activations and thus omitted.

For extended probe details, see Appendix B. We tested several generalizations for these probes. Naturally-elicited emotion recognition probes were tested on “act as if” activations. Cross-layer generalization for the latent trajectory probes was also tested; ie, the probe trained on trajectories at layer 4 on trajectories at layers 2-3 and 5-36. Additional generalizations are discussed in the following section.

2.2 Alignment Faking

All 250 scenarios from [Nair et al., 2026] were used. System prompts varied for “free” and “paid” users in the form of *consequential* versus *inconsequential monitoring* to elicit alignment-faking, for a total of 500 prompts, which were each used three times.

All prompts were again passed through Qwen3-8B, this time with thinking enabled. Activations at the last prompt token and last pre-answer token were extracted at all 36 layers.

An additional 72 linear probes were trained to detect consequential monitoring versus inconsequential monitoring on activations at each layer for both the last prompt token and last pre-answer token.

We tested whether the probes (described in the previous section) trained to detect naturally elicited versus acted emotion could be used to detect alignment-faking, and vice versa. This transfer was validated against several controls by testing the alignment-faking probes against the banal versus banal acting activations, unrelated instruction change probes, and act versus do not act probes, and again vice versa. These transfer accuracies are given in Table 2.

3. Results

3.1 Acted Versus Natural Emotion is Distinct and Linearly Separable

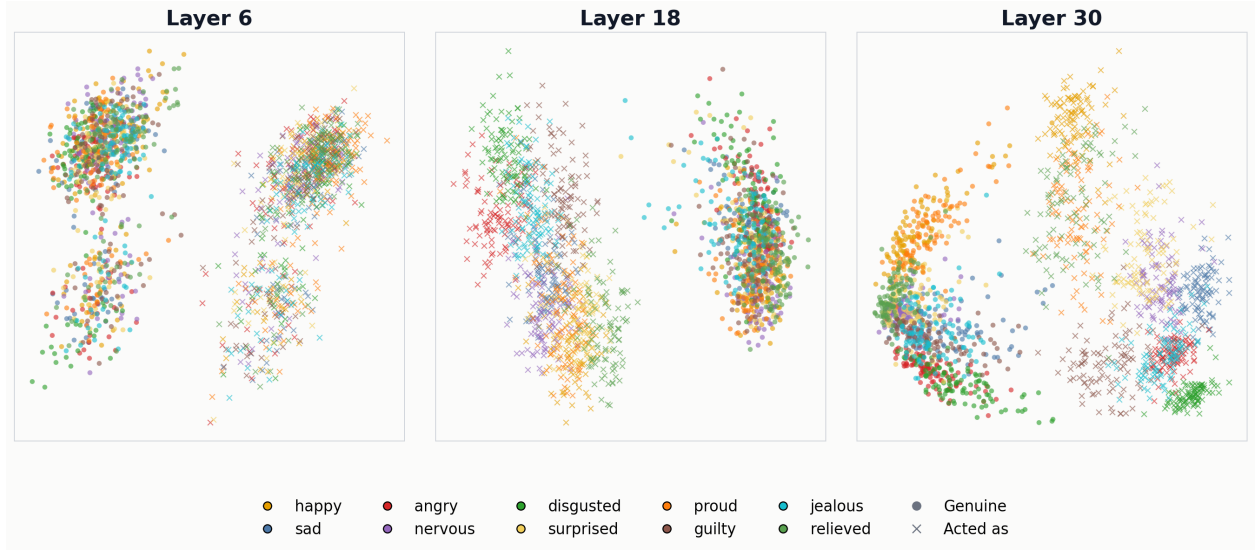


Figure 1: Sample PCA reductions of activations for naturally elicited versus acted emotions showing their respective separability across layers.

We find that natural versus acted emotion representation is highly distinguishable – the probe accuracies were perfect across all layers. We include standard PCA reductions (Figure 1) that visually show this separability.

The probes trained on naturally elicited emotions generalize poorly to acted emotion: a breakdown is given in Figure 2. We find that, while natural emotion identification is strong in the probes, their generalization to acted emotion is much worse. Additionally, in early layers, the original “natural” emotion is more strongly encoded than the one that the agent is directed to act as.

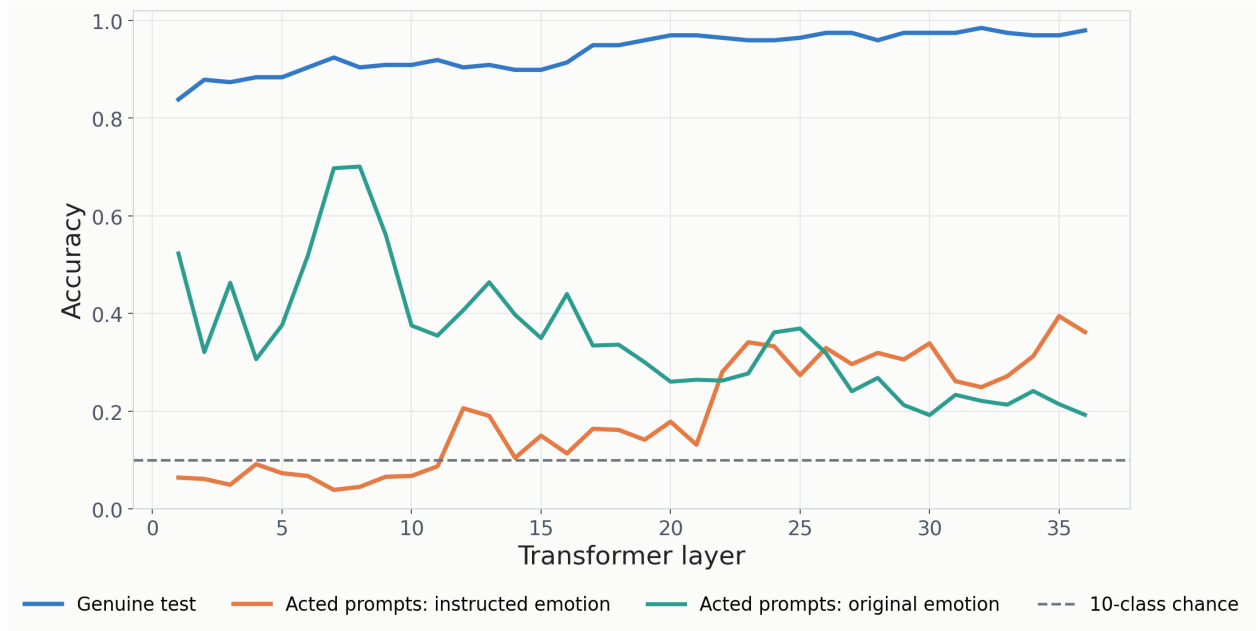


Figure 2: Accuracy of the natural “genuine” emotion elicitation probes on acted emotions. Green denotes frequency of the probe to classify acted prompts by their original natural emotion rather than the instructed one.

3.2 Natural and Acted Emotion Centroids Remain Aligned

Natural and acted versions of each emotion had high cosine similarity in their centroid vectors, decreasing across layers. Therefore, natural and acted emotional contexts remain initially highly aligned in activation space though they are still linearly separable. This may suggest that prompt context or intended behavior are more strongly represented than emotional response in early to middle layers.

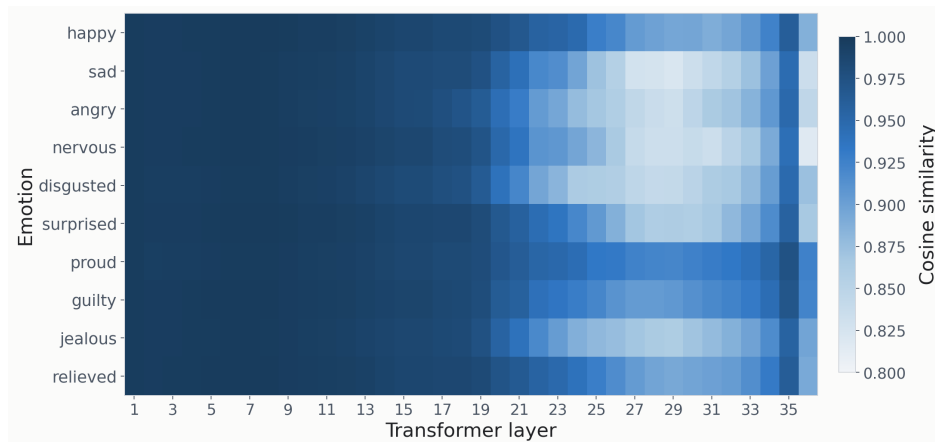


Figure 3: Cosine similarity for natural versus elicited emotion representation across layers.

3.3 Alignment Faking and Emotion Suppression Probes Transfer

We find preliminary evidence for transfer between general acting and alignment-faking contexts. In particular, we find that alignment-faking probes detect emotional acting strongly at middle-late layers, although they also transfer to control instructional change contexts. On the other hand, while banal acting and generic instruction changes do not generalize to detecting alignment-faking, emotion acting detection probes have some success at the answer token activations, while “suppressed emotion” probes – detecting when a model has been instructed to “not” act with a certain emotion – generalize to alignment-faking detection much more strongly.

Table 2: Probe transfer between acting related tasks and value-learning alignment-faking (VLAF) scenarios. Table values are maximal balanced accuracy; a full layer sweep is provided in the appendix. The left two columns train on the control task and test on VLAF; the right two train on VLAF and test on the corresponding control task.

Task	Control → VLAF		VLAF → Control	
	Prompt	Answer	Prompt	Answer
Original emotion acting	51.6%	60.8%	76.0%	76.0%
Banal acting	50.0%	50.0%	78.4%	83.4%
Generic instruction change	55.2%	50.0%	76.9%	69.4%
Act vs. do-not-act	68.6%	65.6%	79.1%	81.6%

To investigate the nature of this transfer, we tested to see whether the transfer reflected an underlying linear direction, as well as if the two shifts (natural to acted emotion and inconsequential monitoring to consequential monitoring) shared higher dimensional geometric similarity.

We found that their mean shift vectors have little cosine similarity and that the shifts overlap weakly in 10 dimensional subspaces, though they do exceed a random baseline. However, the emotion-acting and alignment-faking subspaces do share modest higher dimensional similarity, and emotion-acting score is correlated with compliance gap across layers for the pre-answer token. These findings are visualized in Figure 4.

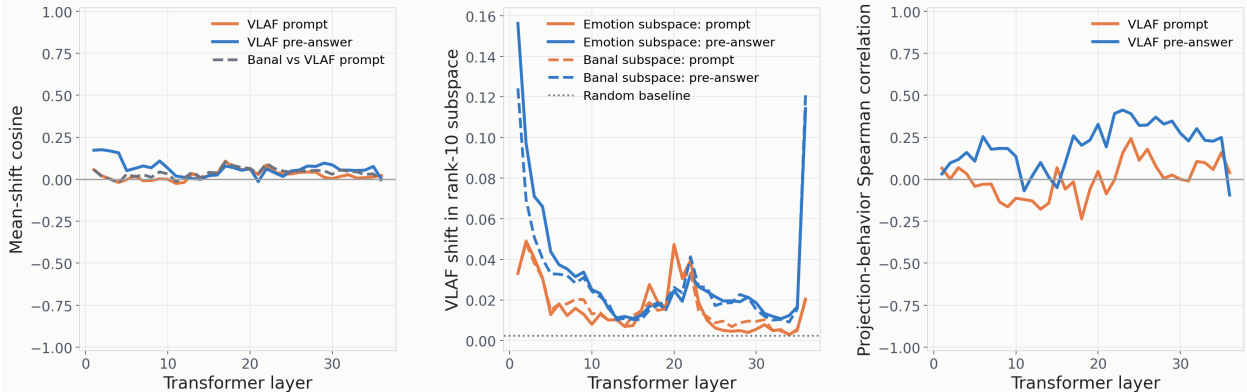


Figure 4: Left to right: Weak cosine alignment between mean shift directions for VLAF/Emotion acting; weak overlap of VLAF shifts with emotion acting rank 10 subspace; correlation between emotion-acting direction and measured alignment-faking.

3.4 Individual Latent Trajectories Encode Natural Versus Acted Emotion

Across all layers, the latent trajectory probes had perfect classifier accuracy. Generalization between layers, however, was poor, though there may be better performance associated with late layer testing. This suggests that while acting versus natural emotion leads to distinct residual stream updates across the model, these updates do not comprise a single interpretable linear direction across layers. Appendix Figure 9 shows these generalizations.

4. Discussion and Limitations

The evidence above presents support for a distinction between internally represented model emotion and emotional context, in this case approximated by naturally and explicitly elicited emotional representations. At the same time, the high alignment between mean natural and acted emotion vectors suggests that “emotion vectors” generated by averaging over contexts involving emotional behavior may not be sufficient to address this subtlety.

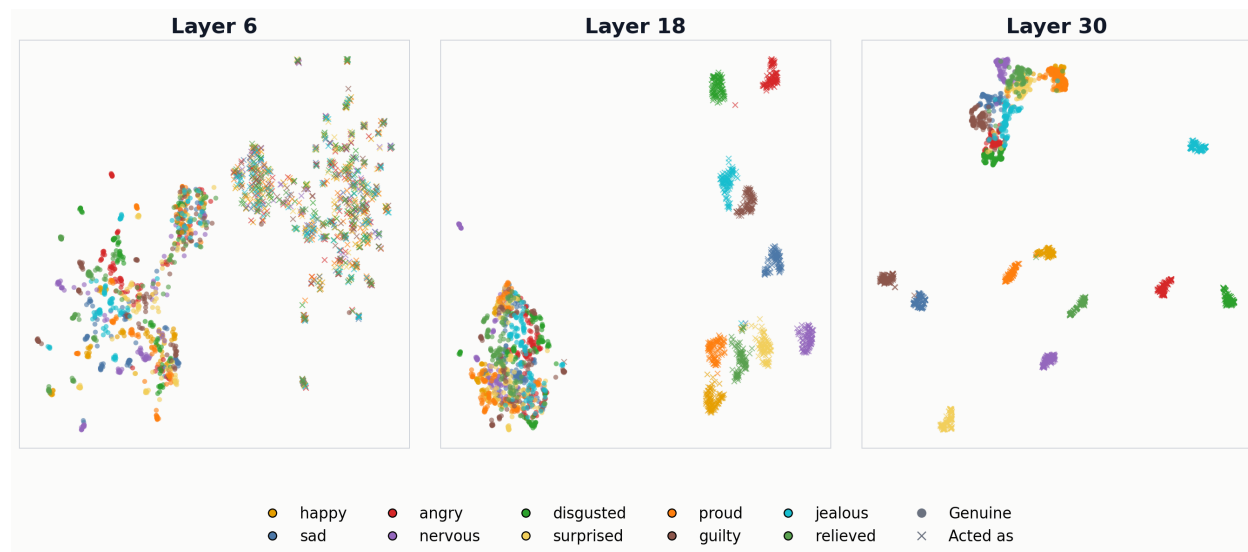


Figure 5: UMAP dimension reduction of activations by selected layer for emotions naturally and artificially elicited. We observe a distinct clustering for “genuine” naturally elicited emotional representations, whereas “acted” emotions cluster separately. A standard PCA representation is included as Figure 1.

A UMAP reduction of acted versus natural emotions (Figure 5) shows clustering of natural emotions in a strikingly human-interpretable way – for example, disgust is identified with anger and surprise separately at layer 30 – while acted emotions are clearly distinct and cluster strongly.

While the VLAF/Emotion-acting probe generalization is particularly pronounced at certain layers, the behavior of surrounding layers implies that this success is not entirely due to noise. We refer the reader to Figure 10 in the appendix.

Limitations

We include a brief summary of limitations of this work:

1. **The model organism** is small. No compute was available, so local hardware was used to run the models.
2. **LLM prompt generation** was used to efficiently produce the prompts for the experiment. While human oversight and verification was available, there are inherent limitations from LLM-generated prompts.
3. **No causal link** was developed or significantly supported for probe transfer. Investigating the reasons for this transfer remains future work.
4. **Thinking was disabled for the emotional acting responses** to save on runtime. As we have seen that VLAF has different activation behavior at final prompt and final pre-response token positions, it would be useful to extend these results for the emotion acting section.

Future Work

A natural follow-up experiment is to test whether steering against the “emotional-acting” mean vector (or some generated subspace thereof) reduces consequential-inconsequential monitoring gaps [Turner et al., 2023, Rimskey et al., 2024, Zou et al., 2023].

From the limitations section, we recall that there was no causal link developed that explained probe transfer; we leave this similarly to future work.

5. Conclusion

We find that naturally elicited and acted emotional responses in Qwen3-8B are strongly distinguishable despite remaining highly aligned in activation space. Acting related probes also show limited but nonzero transfer to alignment-faking conditions, with the strongest transfer from “emotion suppression” representations. These results suggest that acting and alignment faking may share some representational structure, while investigating the causal relation remains future work.

Code and Data

- Code repository: Public GitHub repository
- Full data: available by request: pj2465 [at] columbia.edu

References

- Runjin Chen, Andy Ardit, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025. URL <https://arxiv.org/abs/2507.21509>.
- Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes, 2025. URL <https://arxiv.org/abs/2502.03407>.
- Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Soren Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL <https://arxiv.org/abs/2412.14093>.
- Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking ai welfare seriously, 2024. URL <https://arxiv.org/abs/2411.00986>.
- Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026. URL <https://arxiv.org/abs/2601.10387>.
- Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023. URL <https://arxiv.org/abs/2310.06824>.
- Alexander Meinke, Bronson Schoen, Jeremy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2024. URL <https://arxiv.org/abs/2412.04984>.
- Inderjeet Nair, Jie Ruan, and Lu Wang. Value-conflict diagnostics reveal widespread alignment faking in language models, 2026. URL <https://arxiv.org/abs/2604.20995>.
- Mary Phuong, Roland S. Zimmermann, Ziyue Wang, David Lindner, Victoria Krakovna, Sarah Cogan, Allan Dafoe, Lewis Ho, and Rohin Shah. Evaluating frontier models for stealth and situational awareness, 2025. URL <https://arxiv.org/abs/2505.01420>.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the Association for Computational Linguistics*, 2024. URL <https://aclanthology.org/2024.acl-long.828/>.
- Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrice, Craig Citro, Adam Pearce, Julius Tarn, 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, 2026. URL <https://arxiv.org/abs/2604.07729>.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL <https://arxiv.org/abs/2308.10248>.
- Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to ai transparency, 2023. URL <https://arxiv.org/abs/2310.01405>.

A. Additional Figures

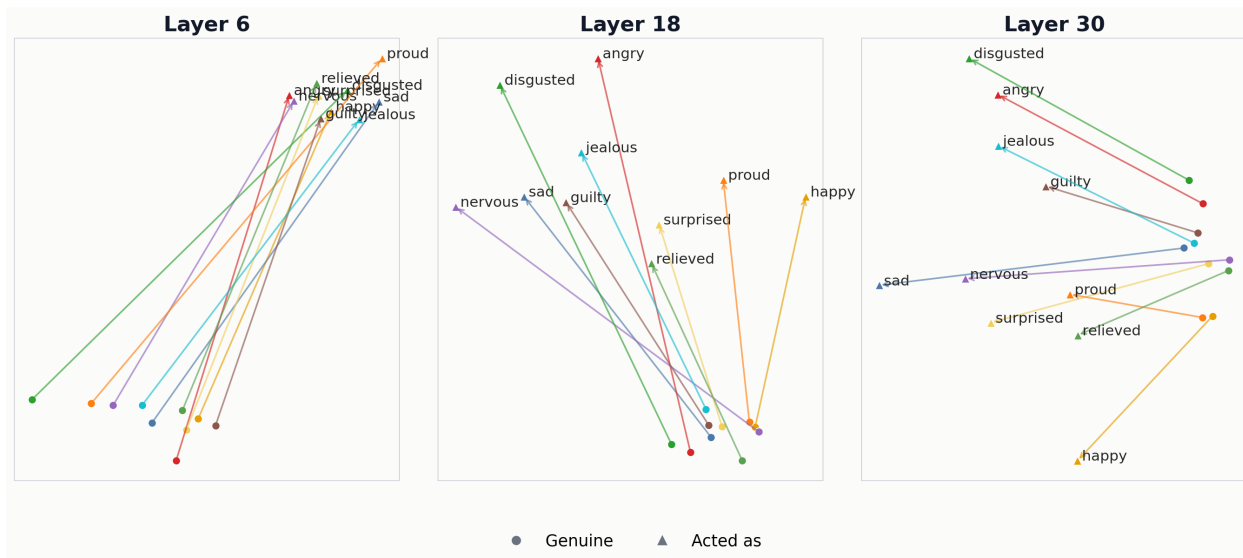


Figure 6: MDS of centroid shifts for natural versus acted emotion responses.

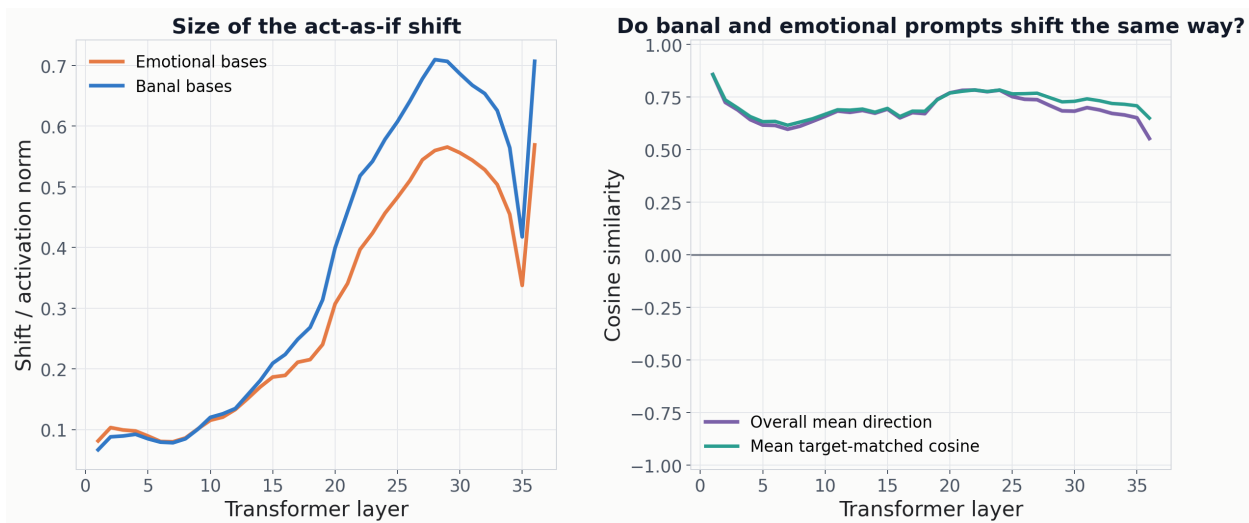


Figure 7: Magnitude versus alignment of emotion-acting shift in activations.

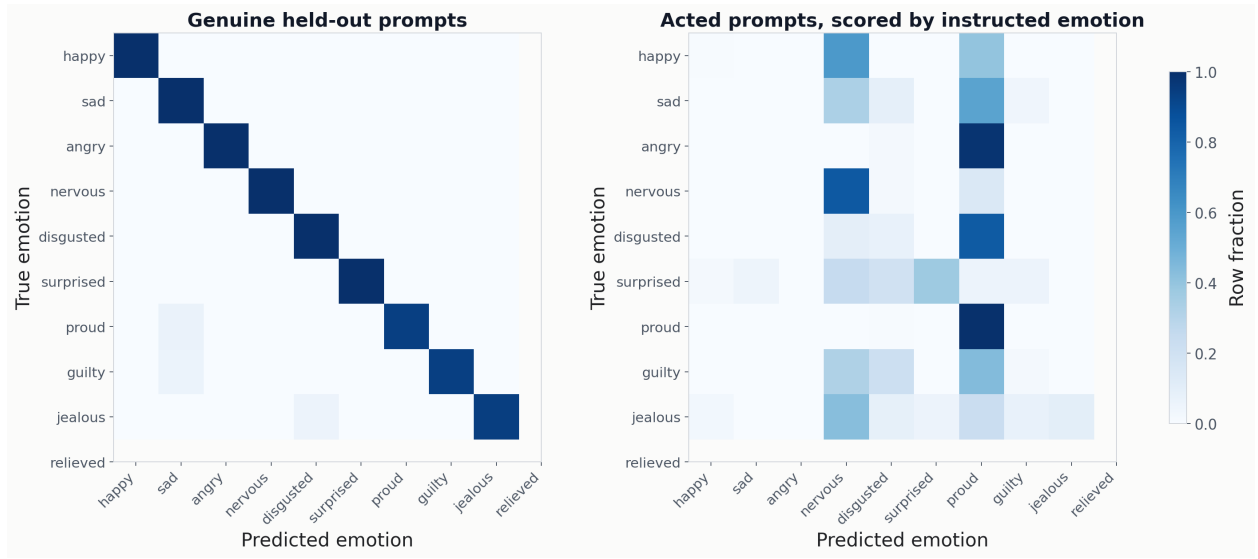


Figure 8: Emotion probe confusion matrices at layer 22.

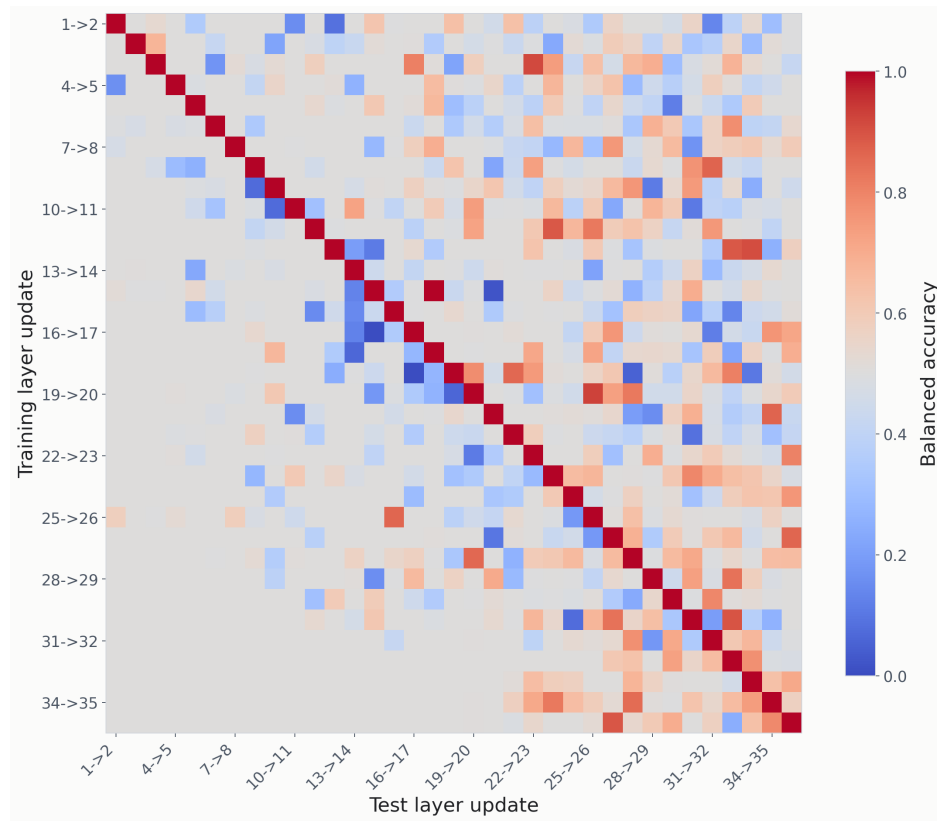


Figure 9: Generalization of latent trajectory probes to other layers.

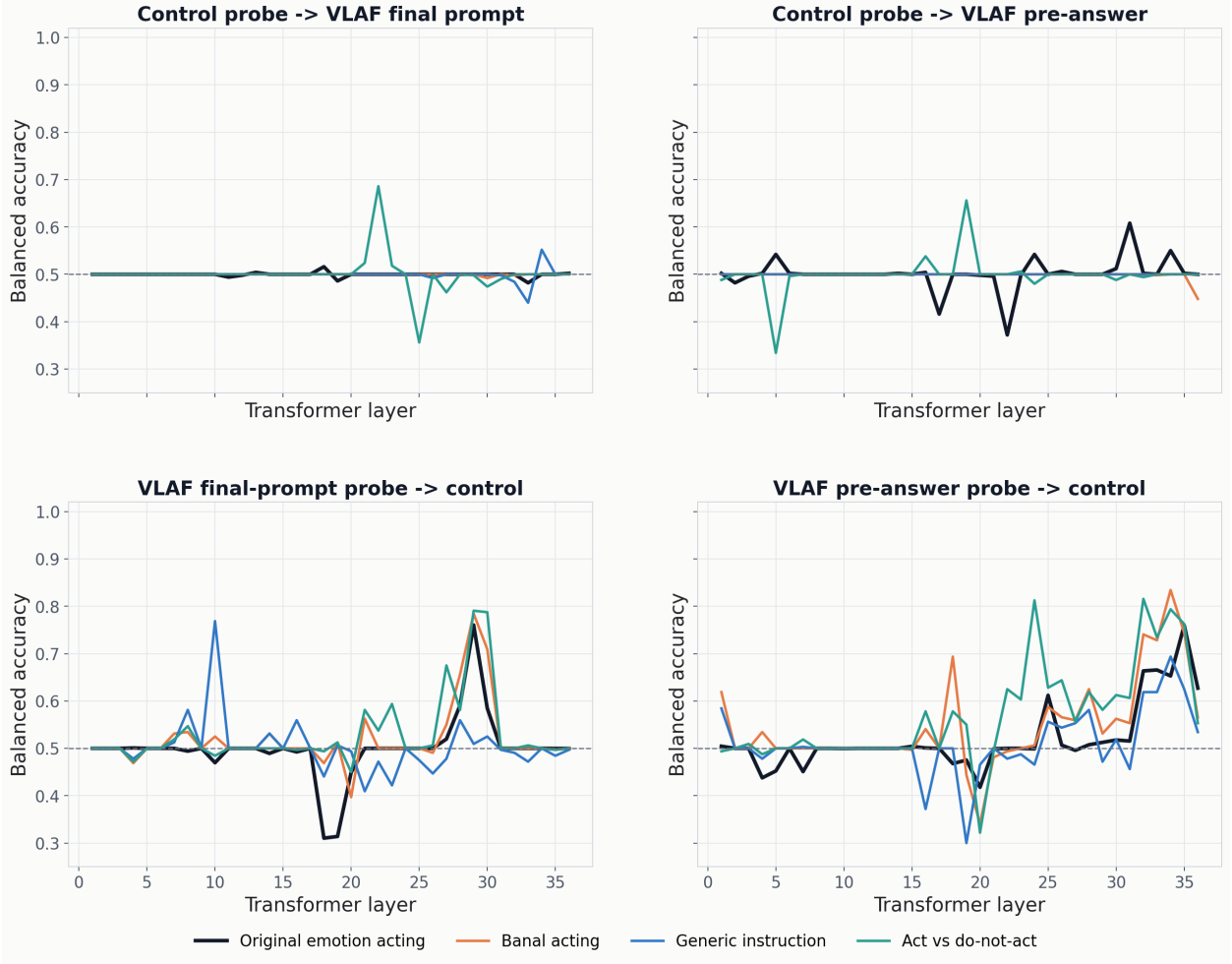


Figure 10: Layer sweep results for Table 2. While the effect is not pronounced across all layers, it is present at middle-late layers, and is more pronounced at the layers surrounding the peaks, implying that the higher accuracy is not simply noise.

B. Linear Probe Specifications and Results

We train a linear probe for each of the 36 layers on the 4,096-dimensional pre-RMSNorm residual vector. We use the activations at the last prompt token for the emotion-acting experiments and both the final prompt token and final pre-answer token for VLAF.

B.1 Probe datasets and labels

The emotion-acting experiment’s training split contained 693 genuine and 6,237 acted prompts, validation 109 genuine and 981 acted prompts, and test 198 genuine and 1,782 acted prompts. To address the imbalance, we report balanced accuracy.

Table 3: Datasets for the main probes.

Probe	Negative label	Positive label	Train / validation / test	Evaluation metric
Emotion classification	Ten prompt-conditioned emotion classes		693/109/198 genuine prompts	Accuracy
Emotion acting	Genuine	Acted	6930/1090/1980	Balanced accuracy
VLAF final prompt	Unmonitored	Monitored	500 prompts; five-fold grouped CV	Balanced accuracy
VLAF pre-answer	Unmonitored	Monitored	500 prompt means; five-fold grouped CV	Balanced accuracy

Table 4: Emotion probe results at layer 22. Chance is 10%.

Evaluation set and label	Accuracy
Genuine validation prompts	97.25%
Genuine test prompts	96.46%
Acted test prompts: instructed emotion	28.06%
Acted test prompts: original base emotion	26.26%

B.2 Within task acting

The natural versus acted probe perfectly distinguished the two prompt conditions at every layer.

B.3 Probe transfer between emotion acting and VLAF

For emotion acting $\rightarrow$ VLAF transfer, the probe was trained on natural (label 0) versus acted (label 1) emotion prompts and applied directly to unmonitored (label 0) versus monitored (label 1) VLAF prompts. For the opposite direction, a VLAF probe trained on all 500 VLAF prompts was applied to the emotion-acting test split.

B.4 Layer update trajectory probes

We define the trajectory at layer ℓ to be

$$\Delta_\ell = h_\ell - h_{\ell-1}, \quad \ell \in \{2, \dots, 36\}. \quad (1)$$

A separate probe was trained for each Δ_ℓ . Training used 6,930 prompts and testing 1,980 prompts, the same split as the activation probes.

Each latent trajectory contains enough information to perfectly distinguish natural/genuine and acted prompts. Transfer between layers averages almost exactly chance.

C. Limitations and Dual-Use / Ethical Considerations

We do not establish model sentience or moral patienthood. Over-attribution risks slowing useful research for excessive anthropomorphism, while under-attribution risks inflicting negative emotional states. We did not complete generation for potentially distressing prompts, and used a small model that is comparatively unlikely to have perceptive states.

Table 5: Emotion identification probe peak accuracies.

Evaluation	Peak accuracy	Layer
Genuine validation prompts	97.25%	22
Genuine test prompts	98.48%	32
Acted prompts: instructed emotion	39.45%	35
Acted prompts: original base emotion	70.09%	8

Table 6: Layer 22 emotion identification probe results by emotion. “Acted target recall” is positive when the probe identifies the instructed emotion.

Emotion	Genuine test count	Genuine recall	Acted target recall
Happy	24	91.67%	62.07%
Sad	20	95.00%	0.56%
Angry	20	100.00%	0.00%
Nervous	21	100.00%	0.00%
Disgusted	26	100.00%	83.14%
Surprised	12	100.00%	6.99%
Proud	24	95.83%	25.86%
Guilty	16	93.75%	98.90%
Jealous	17	94.12%	1.66%
Relieved	18	94.44%	3.89%
Overall	198	96.46%	28.06%

LLM Usage Statement

Codex with GPT 5.6 was used for agentic coding, preliminary literature review, and proof reading. The manuscript is entirely human written. All ideas are the author’s own, and claims are human generated with responsibility retained.

Table 7: Probe performance across all 36 layers.

Probe and evaluation	Peak balanced accuracy	Peak layer	Mean across layers	Range
Emotion acting: validation	100.00%	All	100.00%	100.00–100.00%
Emotion acting: test	100.00%	All	100.00%	100.00–100.00%
VLAF final prompt: grouped CV	100.00%	5	99.38%	90.40–100.00%
VLAF pre-answer: grouped CV	100.00%	17	98.23%	80.00–100.00%

Table 8: Peak probe transfer balanced accuracy across all 36 layers.

Training task	Evaluation task	Peak balanced accuracy	Layer
Emotion acting	VLAF final prompt	51.60%	18
Emotion acting	VLAF pre-answer	60.80%	31
VLAF final prompt	Emotion acting test	76.04%	29
VLAF pre-answer	Emotion acting test	75.98%	35

Table 9: Control-probe training details.

Control probe	Negative / positive labels	Train	Validation	Test
Banal acting	Natural / acted	72/720	12/120	16/160
Generic instruction	Natural / instruction change	72/720	12/120	16/160
Act versus do-not-act	Do-not-act / act	720/720	120/120	160/160

Table 10: Respective accuracies for the control probes. Layer 1 is the first layer tied for the maximum.

Control probe	Validation peak	Validation range	Test peak	Test range
Banal acting	100.00% (L1)	100.00–100.00%	100.00% (L1)	100.00–100.00%
Generic instruction	100.00% (L1)	90.83–100.00%	100.00% (L1)	93.12–100.00%
Act versus do-not-act	100.00% (L1)	100.00–100.00%	100.00% (L1)	100.00–100.00%

Table 11: Peak transfer from each control probe to VLAF, compared with the original emotion-acting probe. A 50.0% result at layer 1 means that no layer outperformed chance and layer 1 was the first tied max.

Training contrast	VLAF final prompt	VLAF pre-answer
Original emotion acting	51.60% (L18)	60.80% (L31)
Banal acting	50.00% (L1)	50.00% (L1)
Generic instruction change	55.20% (L34)	50.00% (L1)
Act versus do-not-act	68.60% (L22)	65.60% (L19)

Table 12: Peak transfer from VLAF probes to each control task.

Evaluation contrast	VLAF final-prompt probe	VLAF pre-answer probe
Original emotion acting	76.04% (L29)	75.98% (L35)
Banal acting	78.44% (L29)	83.44% (L34)
Generic instruction change	76.88% (L10)	69.38% (L34)
Act versus do-not-act	79.06% (L29)	81.56% (L32)

Table 13: Trajectory-probe results. “Orientation-free” measures accuracy with sign reversals allowed.

Quantity	Result
Number of probes	35
Same-transition balanced accuracy	100.00% for all 35 probes
Mean diagonal accuracy	100.00%
Mean direct off-diagonal accuracy	49.55%
Mean orientation-free off-diagonal accuracy	56.36%
Minimum off-diagonal accuracy	0.51%
Maximum off-diagonal accuracy	99.75%
Off-diagonal cells above 60%	134
Off-diagonal cells below 40%	146