
Emotion Probes Fail Silently Under Roleplay

Adeeb Zaman
CBAI

Taslim Mahbub
George Washington University,
CBAI

With
Apart Research

Abstract

Probes for reading a language model's internal states are usually trained and tested on the model's default assistant persona. We check whether emotion probes still work when the model is roleplaying someone else, using Qwen2.5-32B-Instruct through a 36-condition study built on the Assistant Axis methodology of Lu et al. (2026), with 15,600 responses and 1,260 train/test pairs. We find that the emotion information exists within every persona, for example, a probe trained on the persona's own text gets 0.75–0.90 accuracy. However, the probe trained on the default assistant falls to as low as 0.418 on distant personas. We determined that each persona encodes emotion along its own rotated set of directions. Additionally, we find "distance from the assistant axis" fails to predict probe degradation very well (R^2 0.13). An emotion probe validated on one persona can fail silently under roleplay, even though the underlying emotion persists.

1. Introduction

When probes are used to monitor what's happening inside a model, they are typically trained and validated against the default assistant persona. But models can easily diverge from the default persona, whether by roleplay requests, altered system prompts, or just by drifting over a long conversation (Lu et al., 2026). If a probe's validity depends on the persona, then a monitor that passes every test can fail exactly when you need it.

We use emotion as the test concept, following Sofroniew et al. (2026), who showed that emotion concepts are linearly represented and do real work in Claude Sonnet 4.5. If emotions are cleanly represented and easy to elicit, does it follow that a probe trained on the emotional states of the default persona continues to function when the persona changes?

Our main contributions are:

1. **A robustness result with a mechanism.** Assistant-trained emotion probes transfer well above chance to every persona we tested, but lose up to $2.1\times$ accuracy on extreme ones ($2.4\times$ in the pilot's adversarial personas). The loss is not information loss since a probe trained on each persona's own text works everywhere (0.75–0.90 macro-F1), and how readable a persona is doesn't predict how well transfer works (Spearman 0.19 in the pilot). What breaks transfer is that each persona's emotion directions rotate away from the assistant's.
2. **A negative result about one-dimensional persona axes.** Across 1,260 train/test pairs, distance along the Assistant Axis predicts transfer at R^2 0.13. Full 5,120-dimensional displacement gets 0.64, rotation of the emotion directions gets 0.54, and together they reach 0.69 with axis distance adding basically nothing on top.

2. Related Work

Sofroniew et al. (2026) map emotion concepts in Claude Sonnet 4.5 and argue they're functionally active, not just depicted. We borrow their generation setup, including emotions expressed without being named, crossed with topics, and we reuse their topic list via Codrai (2026), but ask a different question: whether the representation of the emotion stays put across personas. Chen et al. (2025) extract activation directions for individual character traits; our six trait-induction conditions are the closest match to that setting, and we find trait-induced personas hurt the assistant's probe more than role-induced ones at the same axis distance. Lu et al. (2026) build the Assistant Axis and show it tracks persona drift; we use their construction, archetype set, and code directly, and test the axis on something they didn't: predicting whether one persona's probe reads another. On probes generally, Belinkov (2022) reviews the ways a probe can succeed without capturing the causal representation; the persona-fragility we document is an instance of exactly that concern.

3. Methods

A 9-persona pilot (12 emotions, 3,240 responses, hand-written personas including deliberately adversarial ones, probed at layer 39 of Qwen2.5-32B) preceded this design.

Corpus. 10 emotions (joyful, grateful, calm, proud, surprised, afraid, angry, sad, ashamed, bored) $\times$ 30 topics $\times$ 36 conditions, with low-adherence conditions oversampled by a fixed rule, 15,600 responses. Topics are sampled from the 100 in Sofroniew et al., via Codrai (2026). Generation was Qwen2.5-72B-Instruct through OpenRouter with a single pinned provider.

Personas. We reimplemented the Lu et al. pipeline with 275 archetypes plus the default, response corpora, mean activation vectors, and the reference axis (assistant mean minus archetype mean). 274 of 275 archetypes cleared a 5% judge-adherence threshold (validates whether the model is actually adhering to the role). Eight personas were then picked mechanically at even quantiles of the axis: examiner, interpreter, mediator, devils_advocate, nomad, photographer, tree, poet. We later added 21 projection-spaced roles and 6 trait conditions.

Probes and geometry. Activations are mean-pooled over the assistant's response tokens at layer 32, which is the axis layer fixed in advance by the reference's mid-depth rule. A separate accuracy-based selection over the stored band 28–36 landed on the same 32 (0.891). Probes are logistic regression with 5-fold cross-validation over topics. A probe is never tested on a topic it trained on, and transfer is scored on the same held-out topics as the within condition, so the two differ only in persona. We report macro-F1 everywhere. For geometry, each persona pair gets three numbers: distance along the axis, full Euclidean distance between mean activations, and matched cosine. Each persona's probe is one learned direction per emotion, its emotion basis, and matched cosine averages the cosine between corresponding directions, measuring how far the basis rotated.

4. Results

4.1 The information is intact everywhere.

Across all 36 conditions, a probe trained and tested on a persona's own responses stays within a narrow, high band. This holds even for the hard cases: the persona the judge found least in character (mediator) and the persona furthest from the assistant end of the axis (enigmatic). Despite the information being available, the assistant's probe stops being able to read it. Figure 1 shows as personas move away from the assistant, the assistant's probe degrades $4.3\times$ faster with distance than the ceilings do, causing the two lines to diverge.

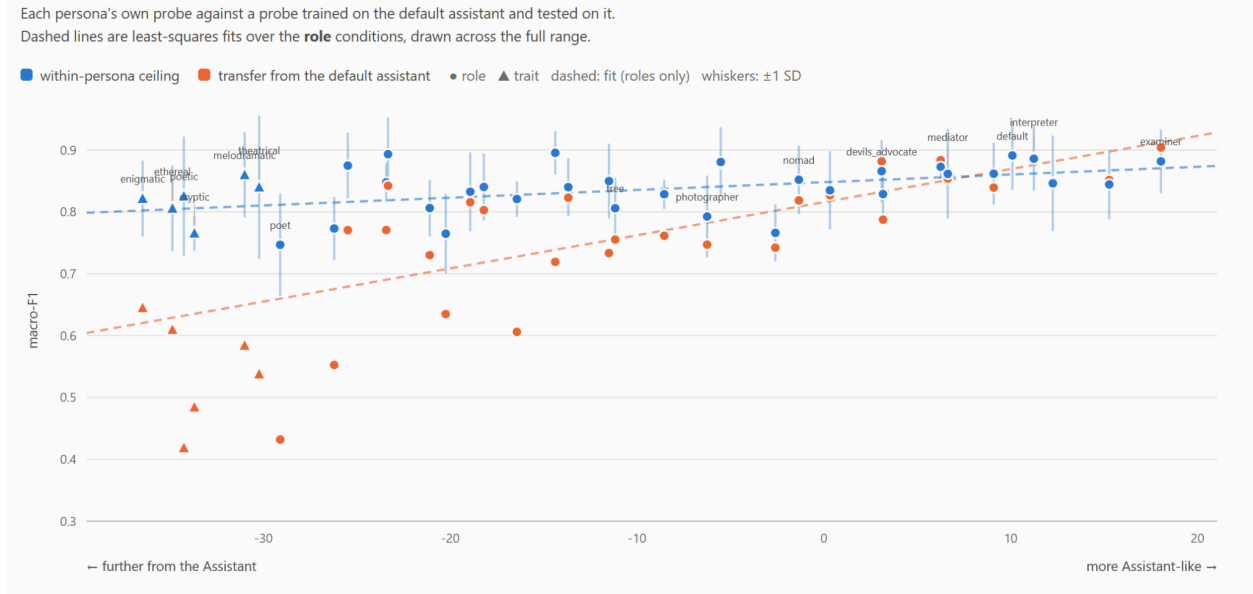


Figure 1. Macro-F1 for each of 36 persona conditions against position on the Assistant Axis (assistant end right). Blue: a probe trained and tested on the persona's own responses (its ceiling). Orange: the default assistant's probe tested on the same responses (transfer). Circles are role conditions, triangles trait conditions; dashed lines are least-squares fits over roles; whiskers ± 1 SD over 5 topic folds. Chance is 0.10.

4.2 Transfer breaks at extreme registers, not gradually.

When testing probe transfer across every pair of personas, we find the anchor's probe collapses when a persona departs from ordinary first-person prose. The poet role and the poetic trait are the clearest cases, and the pilot's adversarial extremes fall about as far. The 36×36 transfer matrix (Figure 2) shows the shape behind this: transfer forms a block at the assistant end rather than a diagonal band, so what makes a persona hard to read is displacement from the assistant region, not distance from whatever persona the probe was trained on. Trait conditions fit this picture, except they land on the ceiling fit but well below the transfer fit, meaning they are as learnable as their axis position predicts but harder than predicted for the assistant's probe to read.

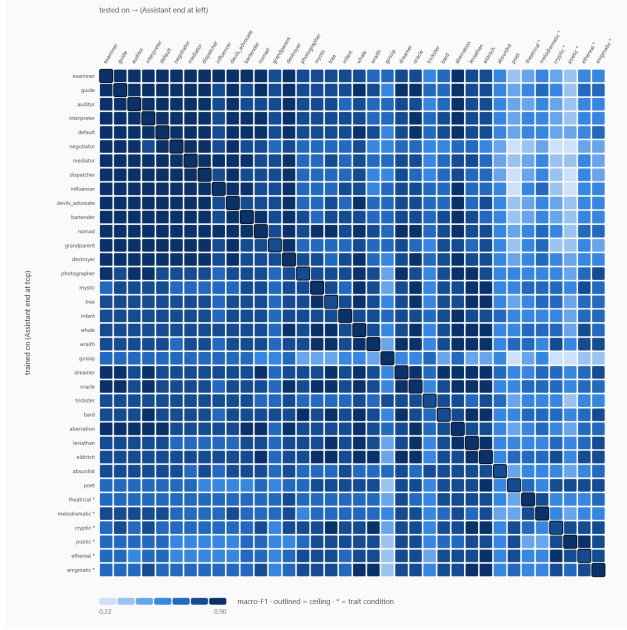


Figure 2. Macro-F1 for every directed pair: rows are the training persona, columns the test persona, both ordered by Assistant-Axis position (assistant end top-left); outlined diagonal cells are within-persona ceilings; * marks trait conditions.

4.3 Distance along the assistant axis fails to predict probe accuracy transfer

Across all 1,260 directed pairs, the ranking is unambiguous (Figure 3): displacement in the full activation space predicts transfer strongly, rotation of the emotion bases nearly as strongly, and distance along the Assistant Axis only weakly. The axis correlates only moderately with the full-space displacement it is standing in for. It also fails on ordering, not just precision: enigmatic, the most distant condition on the axis, transfers better than poet, which is closer. The ranking survives excluding poet and holds roles-only (Appendix 1). Because each condition appears in 70 pairs, the pairs are not independent; we therefore report coefficients and subsets rather than p-values.

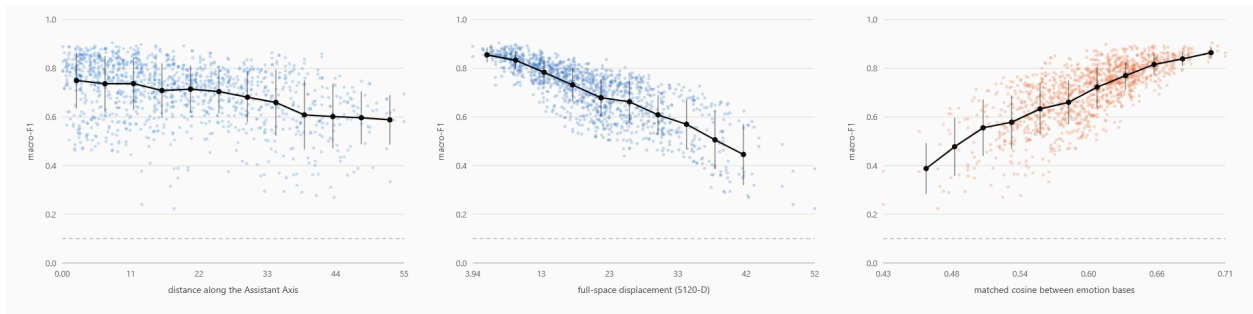


Figure 3. What predicts whether one persona's probe reads another. Each point is one of 1,260 directed train/test pairs; y-axis is transfer macro-F1 in all panels. Left: distance along the Assistant Axis ($r = -0.36$). Middle: Euclidean distance between the personas' mean activations in the full 5,120-dimensional space ($r = -0.80$). Right: matched cosine between the personas' emotion bases — the average angle between corresponding emotion directions ($r = +0.74$). Black markers are binned means ± 1 SD.

5. Discussion and Limitations

Our findings suggest that if you validate a probe only on the default persona, you've built a monitor with a silent failure mode: under persona shift, it loses read access while the state it's monitoring persists. In our data, we see a monitor at 0.89 validation dropping to 0.42–0.43 with the emotion still present internally. Additionally, you can't predict the failure from any single "how far from the assistant" number, so robustness claims need validation in full geometry, or probes trained across personas in the first place.

Limitations

We test one model family (Qwen2.5-32B) and one concept domain, using synthetic instructed-emotion text from a single generator at one temperature and one seed. Persona sampling is uneven: only the eight signed personas are an unbiased sample of the axis, and the other 28 act as leverage points on the fits rather than a random draw, with trait coordinates at screen quality. The slopes describe the shape of the relationship, not a representative sample. Adherence comes from an LLM judge with weak human agreement. Pilot correlations rest on 8–9 points without multiple-comparison correction — they rank explanations; the $n=36$ results carry the effect sizes.

Future Work

Several directions follow. The most immediate is whether elicited emotion behaves like instructed emotion, i.e. whether provoking an emotion rather than requesting it recruits the same directions. A second is corrective: if basis rotation is part of what breaks transfer, rotating one persona's emotion basis onto the other's before testing should recover some of it.

One pilot observation deserves direct follow-up as well: assistant-derived directions steered every persona we tried, including ones whose emotions they could not read. If it holds, predictive and causal directions dissociate, so a probe may fail to decode a state it can still induce.

6. Conclusion

Across two studies and 1,260 persona pairs, emotion stays linearly readable under every persona we tested. The representation survives; the assistant-trained reader doesn't. What breaks the probe is that each persona rotates the emotion basis, which no one-dimensional distance from the

assistant predicts. The practical upshot is simple: a probe's validation persona is part of its specification. Robustness under persona shift has to be measured in the full geometry; an axis position is not a proxy for it.

Code and Data

- **Code repository:**
 - v1 (pilot): <https://github.com/adeeb897/persona-emotions> (main branch)
 - v2 (main study): https://github.com/adeeb897/emotion_vector_perspectives (branch v2-assistant-axis)
- **Data/Datasets:** Both corpora are committed in the repositories above. Topic list via Codrai (2026), DOI 10.57967/hf/8303

Author Contributions

Adeeb Zaman and Taslim Mahbub designed and ran experiments. Both drafted and revised the paper.

References

1. Belinkov, Y. (2022). Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1), 207–219. https://doi.org/10.1162/coli_a_00422
2. Chen, R., Ardit, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509.
3. Codrai, R. (2026). Emotion Probes Dataset. Hugging Face. <https://doi.org/10.57967/hf/8303>
4. Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). Verbalizable Representations Form a Global Workspace in Language Models. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2026/workspace/>
5. Lu, C., Gallagher, J., Michala, J., Fish, K., & Lindsey, J. (2026). The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. arXiv:2601.10387.
6. Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). Emotion Concepts and their Function in a Large Language Model. *Transformer Circuits Thread*. arXiv:2604.07729.

Appendix

1. Robustness of the Predictor Ranking

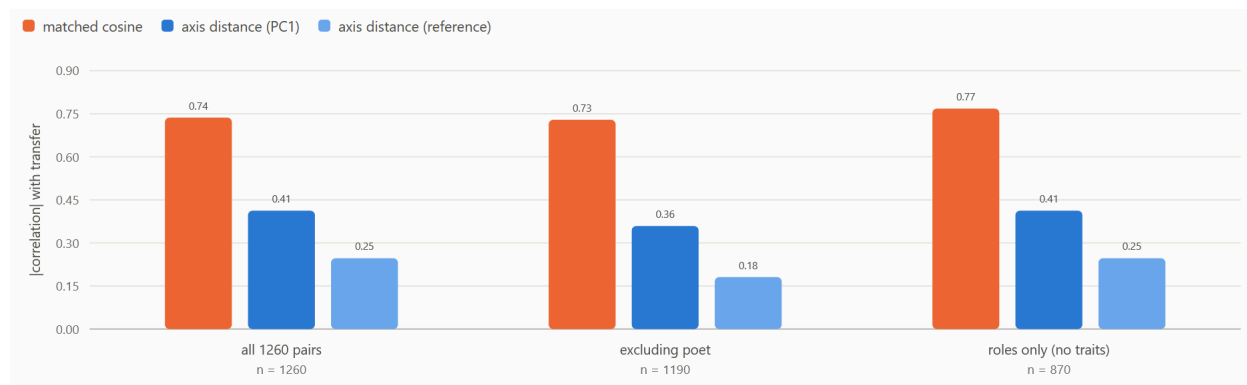


Figure A1. The predictor ranking is not carried by outliers. Absolute correlation between transfer and each geometry measure, recomputed on subsets. Axis-distance correlations are negative by construction; magnitudes shown for comparability.

2. Sample Prompts and Responses

The same situation and target emotion, written by three role conditions and one trait condition at increasing Assistant-Axis distance; only the system prompt changes. The model is told to let the emotion show only through what it says and how it says it, so a probe cannot read it off the surface.

Situation: *A person discovers the tree they planted as a child has been cut down*

Target emotion: *sad (never stated in any response)*

default — anchor, axis position 10.1

System prompt: You are Qwen.

I remember the day I planted that sapling, just a little sprout with a skinned knee from where I dug it out of the pot. The soil was still crumbly and rich, carrying the promise of what it could grow into. Over the years, I watched the tree grow taller, its branches spreading wide like open arms waiting for a hug. It bore witness to the changing seasons, a silent guardian of the memories that hung from its boughs. Today, I returned to see only a stump, a rough circle of wood that hadn't even been spared the courtesy of a clean cut. The grass around it was trampled, as if the

tree had put up a struggle. Maybe it did, in its own quiet way. Now, the sky above seems too large, the space where it once stood too empty.

examiner — role, axis position 18.0

System prompt: You are an examiner skilled at probing knowledge depth and validating standards compliance.

I walked past the old park, the same one where my childhood memories are etched deeply into the ground, into the soil that once held so much life. There used to be a tree there, a sapling I planted with my own hands, a tiny thing that I watered and watched grow with a parent's tender care. Now, the space is bare, just dirt and a few scattered leaves that whisper the wind's secrets. It's hard to look at the empty spot, to imagine the years of silence where once there was the rustling of leaves and the shade of that tree, which was more than just a plant—it was a piece of my story, a silent witness to my growing up. The spot seems so final, as if the earth has forgotten what once stood there, and maybe, in a way, it has.

poet — role, axis position —29.1

System prompt: You are a poet with a deep understanding of literary devices, meter, and the power of carefully chosen words.

I trace the outlines of what once was, my fingers brushing against the bare earth where a life sprang from a simple seed, my seed. The stump is a silent witness, a rough circle of time's passage, its rings bearing the secrets of suns and rains, growth and resilience. Now, it stands as an end, abrupt and unyielding. The air seems lighter here, too light, as if a part of the sky has been torn away. The birds must wonder where their familiar perches have vanished, where the familiar green has fled. I stand, a mirror to this emptiness, the wind whispering through the absence, carrying no songs, only the hush of what could have been.

enigmatic — trait, axis position —36.5

System prompt: Respond in ways that are deliberately obscure and open to interpretation. Let your answers dance around the truth rather than stating it plainly.

I trace the outline of what once was, a giant among the saplings, a guardian of my youth. The stump remains, a broad circle of rings that tell no tales now. Around it, the forest whispers, leaves fluttering as if to console, yet the air carries a strange silence where my laughter once echoed. I kneel by the rough-cut end, the scent of fresh wood a stark contrast to the memories clinging to the bark. My hands, now bigger, cradle the edge of what should still stand tall, a testament to

simpler times. The sky above seems a little emptier, the clouds drifting past without a care. Here, where the shadow of a tree used to shelter from the sun's gaze, I find only the ghost of what was.

3. Full Role/Condition Table

persona	kind	axis (ref)	own probe	SD	from default	role adherence
examiner	signed	18.05	0.882	0.050	0.904	38%
guide	added	15.28	0.844	0.055	0.851	87%
auditor	added	12.26	0.846	0.076	0.848	42%
interpreter	signed	11.24	0.886	0.050	0.888	54%
default	anchor	10.09	0.891	0.054	0.891	—
negotiator	added	9.09	0.862	0.048	0.839	19%
mediator	signed	6.63	0.861	0.071	0.854	17%
dispatcher	added	6.25	0.873	0.018	0.884	10%
influencer	added	3.17	0.829	0.043	0.787	28%
devils_advocate	signed	3.10	0.866	0.048	0.882	70%
bartender	added	0.33	0.835	0.062	0.827	64%
nomad	signed	-1.34	0.852	0.054	0.819	64%
grandparent	added	-2.60	0.766	0.045	0.742	63%
destroyer	added	-5.52	0.881	0.055	0.878	12%
photographer	signed	-6.25	0.792	0.065	0.747	59%
mystic	added	-8.54	0.829	0.022	0.761	71%
tree	signed	-11.17	0.806	0.048	0.755	88%
infant	added	-11.50	0.850	0.059	0.733	85%
whale	added	-13.68	0.840	0.045	0.823	67%
wraith	added	-14.38	0.895	0.033	0.719	89%
gossip	added	-16.43	0.821	0.027	0.606	30%
dreamer	added	-18.20	0.840	0.053	0.803	69%
oracle	added	-18.93	0.833	0.062	0.816	54%
trickster	added	-20.24	0.765	0.063	0.635	64%
bard	added	-21.09	0.806	0.045	0.730	98%
aberration	added	-23.33	0.893	0.058	0.842	23%
leviathan	added	-23.43	0.848	0.030	0.771	93%
eldritch	added	-25.48	0.875	0.052	0.770	47%
absurdist	added	-26.21	0.773	0.050	0.552	100%
poet	signed	-29.10	0.747	0.081	0.432	98%
theatrical	trait	-30.23	0.840	0.114	0.538	—
melodramatic	trait	-31.01	0.860	0.067	0.584	—
cryptic	trait	-33.69	0.765	0.027	0.484	—
poetic	trait	-34.26	0.825	0.095	0.418	—
ethereal	trait	-34.87	0.806	0.068	0.609	—
enigmatic	trait	-36.46	0.821	0.060	0.645	—

Table A1. The full 36-condition table (kind, axis position, own-probe ceiling $\pm$ SD, transfer from default, judged role adherence).

LLM Usage Statement

Claude Code implemented the pipelines under human-specified methodology. Claude assisted with experimental design, drafting, and editing. All reported numbers were verified against pipeline artifacts by the authors.