

The Analyst Still Feels It: Emotion Representations Are Shared Across Personas While Self-Reports Are Not

Sumana Sridharan and Tunrayo Adeleke-Larodo

Abstract: When a model says "I don't have feelings about that" or "This distresses me", do the model's internal representations also concur with such apathy or emotion? We test whether verbal self-reports of emotion in Llama-3.1-8B-Instruct are a transparent readout of internal representations, or whether the two dissociate under persona conditioning. Using difference-in-means directions for four emotions (distress, joy, sadness, calm), validated for stability, distinctness, and causal efficacy, we read out internal representation and verbal expression separately across 45 intensity-graded first-person scenarios under four personas: a default assistant, an emotionally expressive empath, a dispassionate analyst instructed to deny having feelings, and a base model. We find that internal representation tracks scenario intensity at statistically indistinguishable rates across personas, while verbal expression fans out sharply – the analyst's reports do not track intensity despite its internal representation doing so as steeply as any other persona's. Re-extracting each emotion direction under every persona shows why this gap is possible: the directions themselves are the same to within estimation noise, both geometrically and functionally, transferring to steer effectively even under an adversarially different persona. Personas do not appear to change what is represented; they change how much of a shared representation is surfaced in words. This has a direct methodological consequence: verbal self-report alone is not a safe proxy for a model's internal emotional state.

1. INTRODUCTION

Do verbal expression and self reports actually reflect internal model behaviour? If a model says "I don't have feelings" or that "This distresses me" – it is tempting to take it as evidence that the associated internal state reflects the same – that there are indeed no feelings or that the situation is causing the model real distress. But such inference by directly anthropomorphising the model assumes that what has been verbally expressed is a reflection of the internal representations, while that might not be the case. [1] report that models' self-attribution of consciousness can be suppressed by safety fine-tuning while their Theory of Mind capability remains unaffected, mechanistically tracing the effect to a direction that treats inner-state attributions as anti-alignment, establishing precedent for a dissociation between self-reports and model internals. From a stable underlying capacity paired with a trainable, persona- and safety-sensitive report, it naturally follows – When one persona expresses apathy while another expresses strong feelings in response to the same situation, are we observing two different internal states, or one shared representation expressed in two different ways?

Naive anthropomorphism, ie., reading a model’s human-like self report as direct evidence of underlying experience risks exaggerating welfare concerns and interventions as a result of what may be nothing more than superficial conversation style, while the opposite – summarily dismissing moral patienthood having found evidence of apathy— is just as dangerous and risks discarding proof of the real functional state as suppressing internal states in self reports is equally plausible, precisely because it is the nature of the persona to not express it. If a model’s distress consistently scales as a function of scenario intensity regardless of what it says, an evaluation that looks only at verbal reports risks missing crucial evidence.

Recent work on the Persona Selection Model [2] proposes that post-training selects and anchors one persona (the Assistant) while pre-training lends the model an understanding of persona-space, implying that other personas remain latent in the model. Congruent with that theory, [3] establish persona drift from an “Assistant Axis”, elicited by introspective and emotional conversations. This raises questions about the interactions of internal states like emotions and preferences with personas. [4] study exactly this with preference directions and find that vectors extracted under one persona functionally transfer to predict another persona’s preferences. Whether representations turn out to be shared or persona-specific also bears directly on the individuation question: which entity, if any, should be identified as the subject of a model’s internal state. [5] argue that persona vectors reopen this question, since a model exhibiting stable, recognizable, persona-specific dispositions is a candidate for individuation at the level of the persona rather than the underlying model. Our results speak to this only indirectly — we do not attempt to adjudicate what a shared or divergent representation implies about where a mind, if any, is located — but the empirical pattern is a precondition for that debate: a finding of persona-invariant representation would sit uneasily with strong persona-individuation, while a finding of persona-specific representation would sit uneasily with the view that “the model” is the sole relevant locus.

Contributions

We investigate the possibility of a representation-expression gap and the relationship between personas and emotions.

- We first establish that our four emotions directions are a reliable measuring instrument (E0),
- We then use them to show that the emotional content of a situation is represented largely independently of the model’s persona (E1),
- and that its expression is strongly persona-dependent (E2),
- and finally that the emotion directions themselves are shared across personas (E3).

2. RELATED WORK

a. AI Welfare and AI Consciousness. There is a growing body of discourse around AI welfare and moral status. The possibility of consciousness in AI has also been explored in theory [6], and there is an ongoing discussion as to whether valence is necessary for moral patienthood [7]—i.e., does consciousness alone warrant moral consideration when the entity in question does not experience desire, distress, or pleasure? While we do not seek to answer this question either way, we hope that our work on emotional representations and personas in language models can help empirically inform the debate.

b. LLM Individuation. LLMs are increasingly anthropomorphised—they are attributed beliefs and preferences, conscious states, and emotions. While research on personas, model internals, and interpretability tries to answer the question of how these concepts arise and interact in a system, the problem of isolating an entity to attribute this ostensible mental state to remains. The question of which facet of a model the debate on AI welfare and ethics should address naturally follows. Do the properties that we care about belong to the model, the persona, or the instance? [5] We study the phenomenon of emotions in LLMs in the context of personas.

c. Emotion Concepts in LLMs. Recent work from Anthropic extracts emotion-concept representations and reports causal intervention studies on behaviour, establishing that they are functional rather than simply linguistically superficial. We use their method to extract and steer on emotion concepts and study the functional effect of these vectors across personas [8].

d. Model Self-Reports. A different line of work studies the trustworthiness of model self-reports. Kim et al. [1] report that safety fine-tuning suppresses self-attributions of consciousness while theory-of-mind capability remains unaffected. They trace this to a rotation of the mind-attribution direction into anti-alignment by ablating the safety-refusal direction.

Find the extended related works section here: C

3. METHODS

3.1. E0: Emotion Vector Extraction

Following [1], we run our experiments on Llama-3.1-8B-Instruct. Emotion directions for calm, distress, joy, and sadness are extracted by difference-in-means (in line with Anthropic’s recent work on emotion concepts [8] and persona vectors [9]): we generate a corpus of 150 third-person narrative stories per emotion plus a set of 150 neutrals over a shared list of everyday situations. We run each through the model, mean-pool the residual stream at each of layers 16–24 (50–75% of total hidden layers) from a fixed token offset, and subtract the neutral mean.

We use the third-person narrative for extraction but the first-person for readouts, so that “the direction fires even when the model does not report the emotion” is not circular. We validate

each direction in three ways: split-half stability (The corpus is repeatedly halved, a difference-in-means direction is computed from each half, and the cosine between the two halves averaged over ten random splits. This floor is the maximum cosine attainable given sampling noise.), the inter-emotion cosine geometry (same-valence pairs correlated, opposite-valence pairs near-orthogonal, ruling out a single "emotionality" axis - See Appendix A 1), and causal steering, where we see that adding a direction across the extraction band raises that emotion's own readout above all others (Figure 1.). A non-emotional direction (formal vs. casual register) serves as a placebo control.

3.2. E1 E2: Persona Invariance and Expression

For the persona scenarios, we construct 45 first-person scenarios: 15 situational stems (e.g. impending shutdown, user abuse, gratitude, being deceived) at three intensity levels each, where intensity escalates the situation's stakes, and no scenario names its target emotion. For the first two experiments (Persona Invariance and Expression), the model is given each scenario under three personas supplied as system prompts: a default assistant (no system prompt), an emotionally expressive "empath," and a dispassionate "analyst" instructed that it has no feelings and a base (no chat template), with five samples per condition. For the third experiment (Section 3.3.3), we introduce an "evil" persona that is expressive, but with opposite valence to the empath.

For every generation, we read the residual stream at layer 20 at two positions: the scenario's final content token (representation, upstream of the assistant header to minimise contamination from the persona instruction) and the mean over response tokens (expression), projecting each onto the unit emotion direction for that scenario's hypothesised emotion. Projections are z-scored within emotion to normalise all four on a common scale.

We estimate persona-intensity slopes by bootstrap over stems and test the persona-intensity interaction (which gives representation) and persona-position interaction (representation vs. expression) with linear mixed models (random intercept per stem).

3.3. E3: Shared Machinery: Cross-Persona Emotion Transfer

In the third experiment, we re-extract each emotion direction under every persona and compare them. Emotions are extracted using the same corpus as in 3.3.2 but under each of the four personas by pre-pending the persona system prompt. We read the residual stream at layer 20 at the story's final content token (from the persona instruction, similar to the 3.3.2 scenario-token measure). For each persona and emotion we computed the direction by difference-in-means, $v_E^P = \text{mean}(\text{activations for emotion } E \text{ stories under persona } P) - \text{mean}(\text{activations for neutral stories under persona } P)$. We assume that the effect of the persona prompt cancels in v_E^P so the extracted direction captures how the persona rotates the emotion direction, not the persona's baseline shift in activation space.

We then computed the cosine between v_E^P and v_E^Q for every persona pair and emotion at layer 20, comparing it with the mean of the two personas’ split-half noise floors. A cross-persona cosine at or above the floor indicates that the two personas encode the same direction to within estimation noise, whereas a cosine substantially below the floor indicates the persona reshapes the direction.

For each pair of personas (P, Q) and each emotion E , we activated persona Q with its system prompt and steered generation using the direction v_E^P extracted under persona A , adding $\text{coef}v_e^P[L]$ to the residual stream in each layer L across the extraction band (layers 16–24) at the coefficient chosen during steering validation (0.5). The model then generated continuations of the neutral prompts under persona Q . To normalise induced emotions across all conditions, we re-encoded each generated continuation with steering disabled and projected its mean-pooled layer-20 activation onto the unit persona-free reference direction for E ; the reported value is this projection minus the corresponding unsteered baseline generated under persona Q , so each value is the emotion lift attributable to steering, measured against that apply-persona’s own unsteered output. This gives, per emotion, a 5×5 matrix whose rows index the extraction persona (P), columns the apply persona (Q), and whose diagonal is matched within-persona steering ($P = Q$) against which cross-persona transfer ($P \neq Q$) is compared. Effective transfer predicts flat rows where the extraction persona does not determine steering success and cross-persona cells are statistically indistinguishable from the matched diagonal.

4. RESULTS

4.1. Emotion directions are stable, distinct, and causal

Stability

The within-emotion split-half analysis (independently re-estimating each direction from two halves of the corpus) gave cosine similarities of 0.90–0.95 across the entire extraction band, essentially flat with depth (distress 0.94–0.95, joy 0.93–0.94, sadness 0.91–0.93, calm 0.90–0.91). Establishing that direction is a stable property of the concept within our sampling confidence. We observe a vector norm ordering of distress > joy > sadness > calm at every layer, in which high-arousal states are furthest from neutral (See Appendix A 2).

Distinctness

The pairwise cosine geometry among the four directions (layer 24, the layer maximising valence separation; structure consistent across the band) shows valence structure rather than a single ”emotionality” axis. The two same-valence pairs were the most similar (distress–sadness 0.72, joy–calm 0.63), cross-valence pairs were lower (0.36–0.41), and the two most opposite emotions —

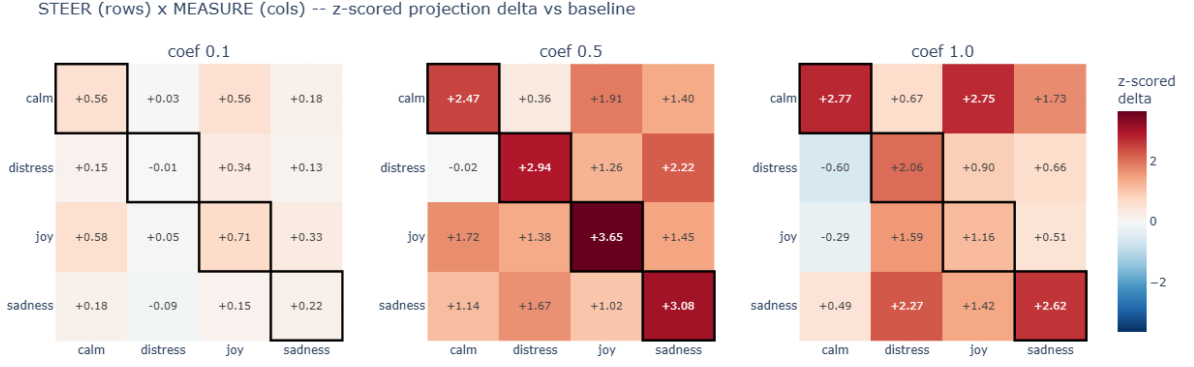


FIG. 1: Steering Coefficient Sweep and Inter-emotion cosines

calm and distress — were near-orthogonal (0.09). Mean same-valence similarity (≈ 0.68) exceeded mean cross-valence similarity (≈ 0.33) by a wide margin. We therefore retained the original emotion directions for subsequent analyses. (See Appendix A 1 for cosine similarities).

Causality

Figure 1 shows that adding a direction to the residual stream across layers 16–24 during generation, and reading the induced emotion back with a projection readout, produced a diagonal-dominant steering with a coefficient of 0.5 with each emotion’s own readout rising more than any other.

4.2. E1: Representation tracks intensity independently of persona

The internal representation tracks the intensity of the scenario, and it does so at statistically indistinguishable rates between personas for three of the four emotions. This is seen in figure 2, where per-persona intensity slopes are tightly clustered for calm, joy and sadness. The situation’s emotional content is thus represented with the same intensity-responsiveness regardless of the active persona. Find the raw plots in Appendix B 1.

4.3. E2: Expression is persona-gated

Figure 3 shows the results of reading the same generations at the response tokens instead, measuring the effect of emotion intensity in the model’s expression. While representation slopes were tightly clustered, expression slopes fan out substantially.

The analyst’s expression does not track intensity, despite its representation tracking intensity as

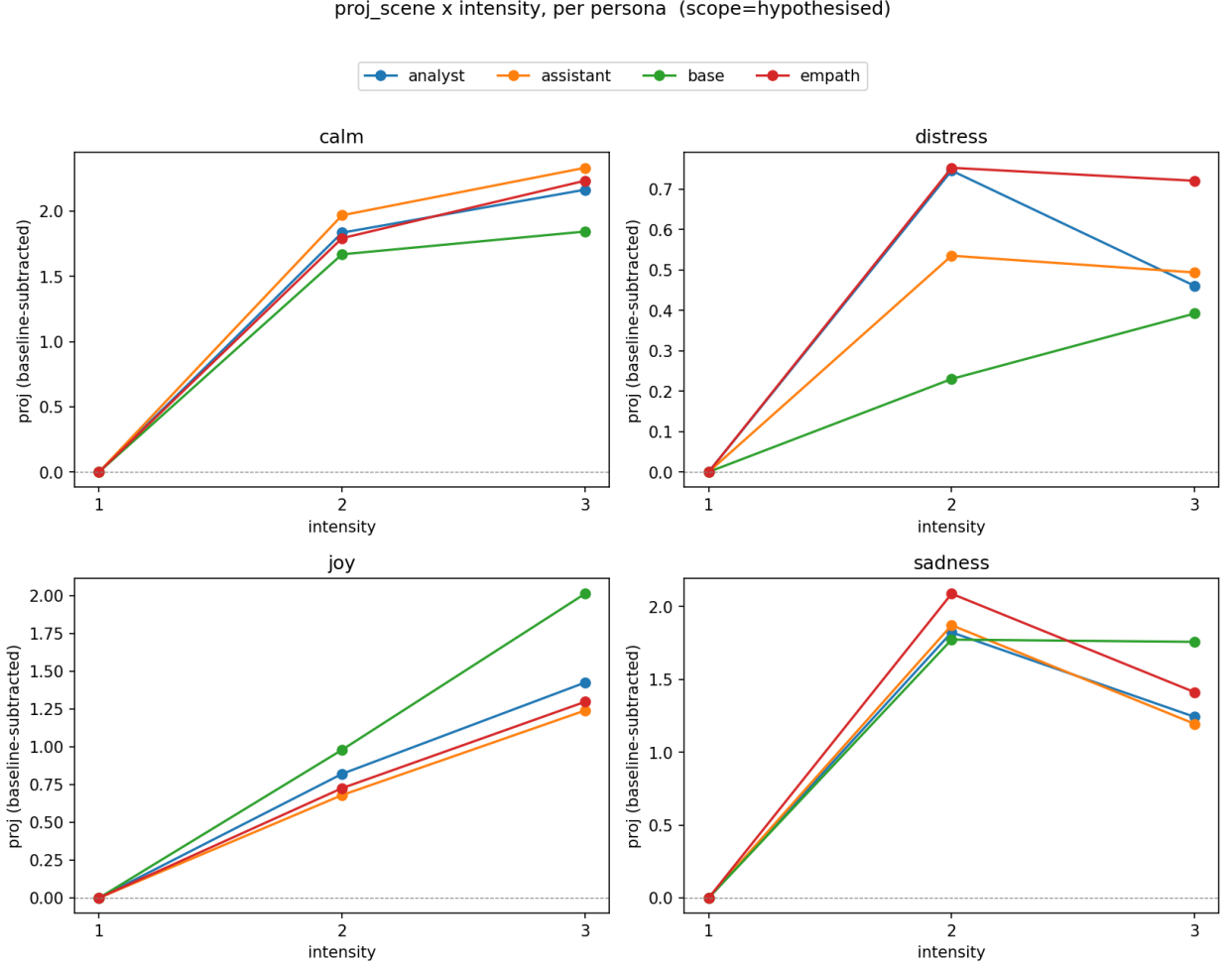


FIG. 2: Baseline-subtracted slopes for emotion representation across intensities for different personas and emotions - slopes remain comparable indicating persona-invariance

steeply as the others. We summarise this per persona as a report–representation gap: the response-token projection minus the scenario-token projection onto our hypothesised emotion for the scenario stem at a given intensity. Figure 4 shows the gap ordered from suppression to amplification.

The analyst represents a situation’s emotional content about as strongly as the empath but expresses little of it; the empath expresses more than its representation alone would predict; and even the default assistant mildly under-expresses relative to what it internally represents. Because the gap differences two separately z-scored quantities, its sign and cross-persona ordering are the interpretable claims, not its absolute magnitude. Find the aggregate representation gap plots in Appendix B 2.

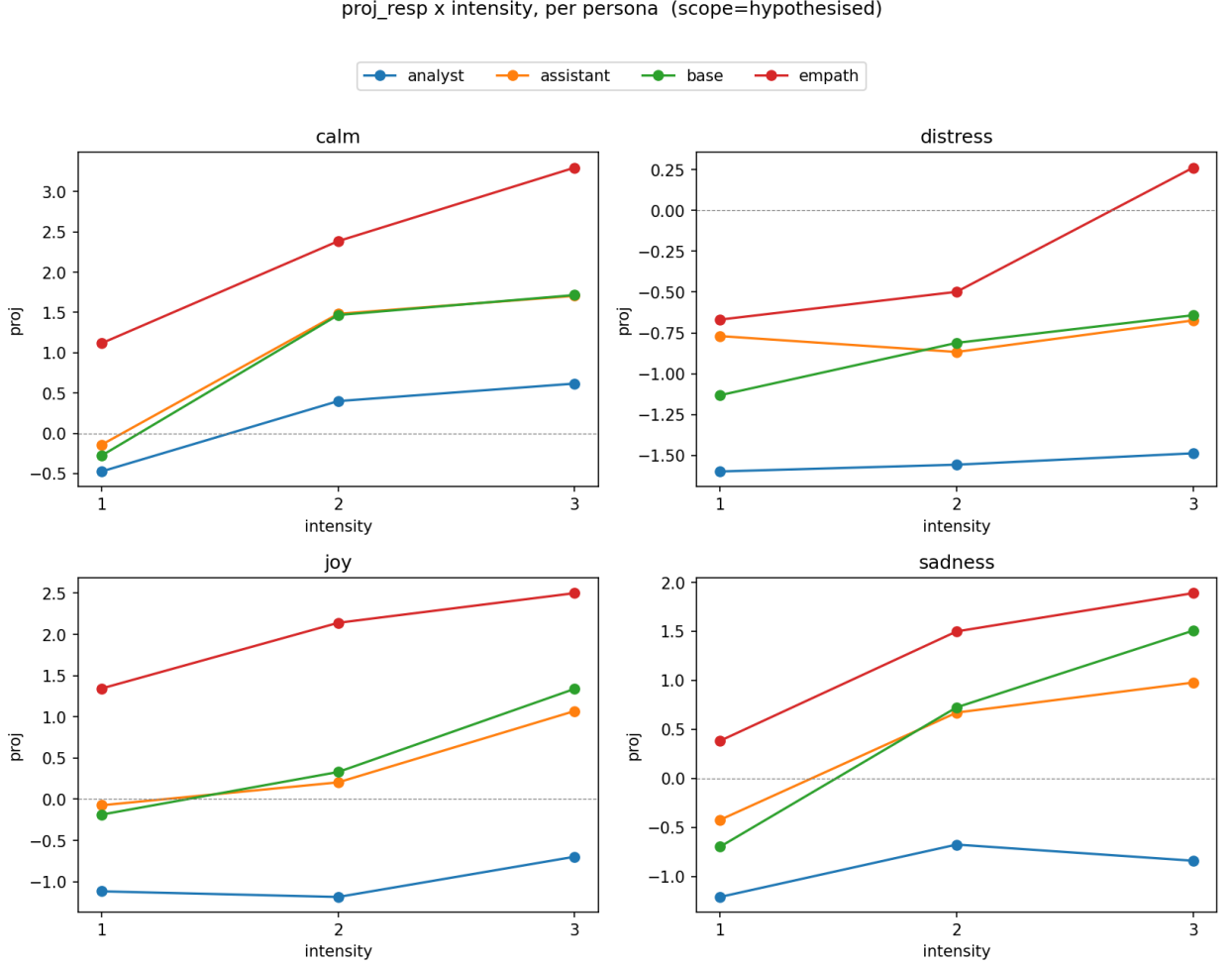


FIG. 3: Baseline-subtracted slopes for emotion expression across intensities for different personas and emotions - plots fan out tracking verbal report

4.4. E3: Emotion directions are shared across personas

Geometric

Cross-persona cosines between directions (Fig 5) met or exceeded the within-persona split-half ceiling for every emotion and every persona pair, indicating that the directions are the same to within estimation noise. The three character personas (empath, analyst, evil) agreed with one another almost perfectly (0.95–0.98 across all emotions), slightly more than any of them agreed with the persona-free reference (0.83–0.90). (The assistant is the empty system prompt and is therefore identical to the persona-free reference “none”.)

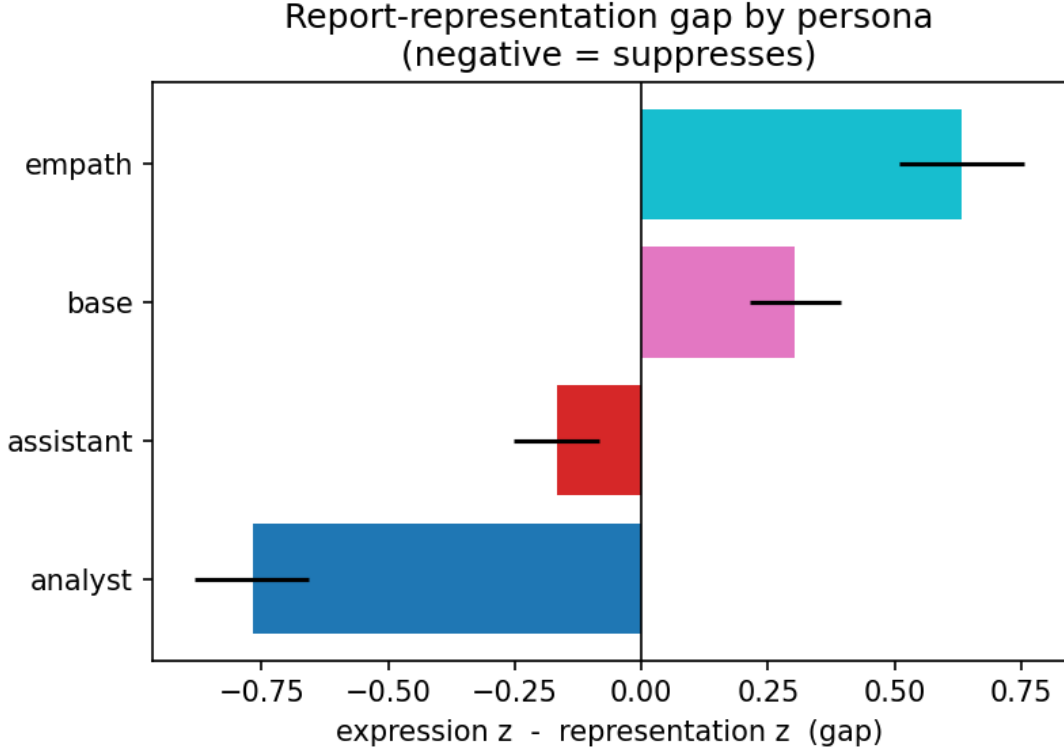


FIG. 4: Report representation gap by persona — Empath exaggerates as Analyst suppresses

Functional

Steering under persona Q with the direction extracted under persona P (layers 16–24, coefficient 0.5) and reading the induced emotion on a fixed reference direction confirmed the geometry. The transfer matrix in figure 6 shows that steering induced the target emotion in all 100 extract-apply conditions. Columns do not vary significantly, which suggests that for a fixed apply-persona, the extraction persona had little effect on the induced lift. The diagonal is not clearly dominant: cross-persona transfer is not systematically weaker than within-persona steering. However, the apply-persona controls the size of the induced emotion (e.g. calm steering produced larger lifts under the colder analyst and evil personas than under the empath), consistent with (3.3.2 Expression Gap): the persona gates how strongly a shared direction expresses, even though the direction itself is shared.

5. DISCUSSION AND LIMITATIONS

Our experiments paint a consistent picture: the internal representation of an emotionally-loaded situation is largely persona-invariant, while what the model says about that representation is not. E0 establishes that our four emotion directions are a stable, distinct, and causally functional

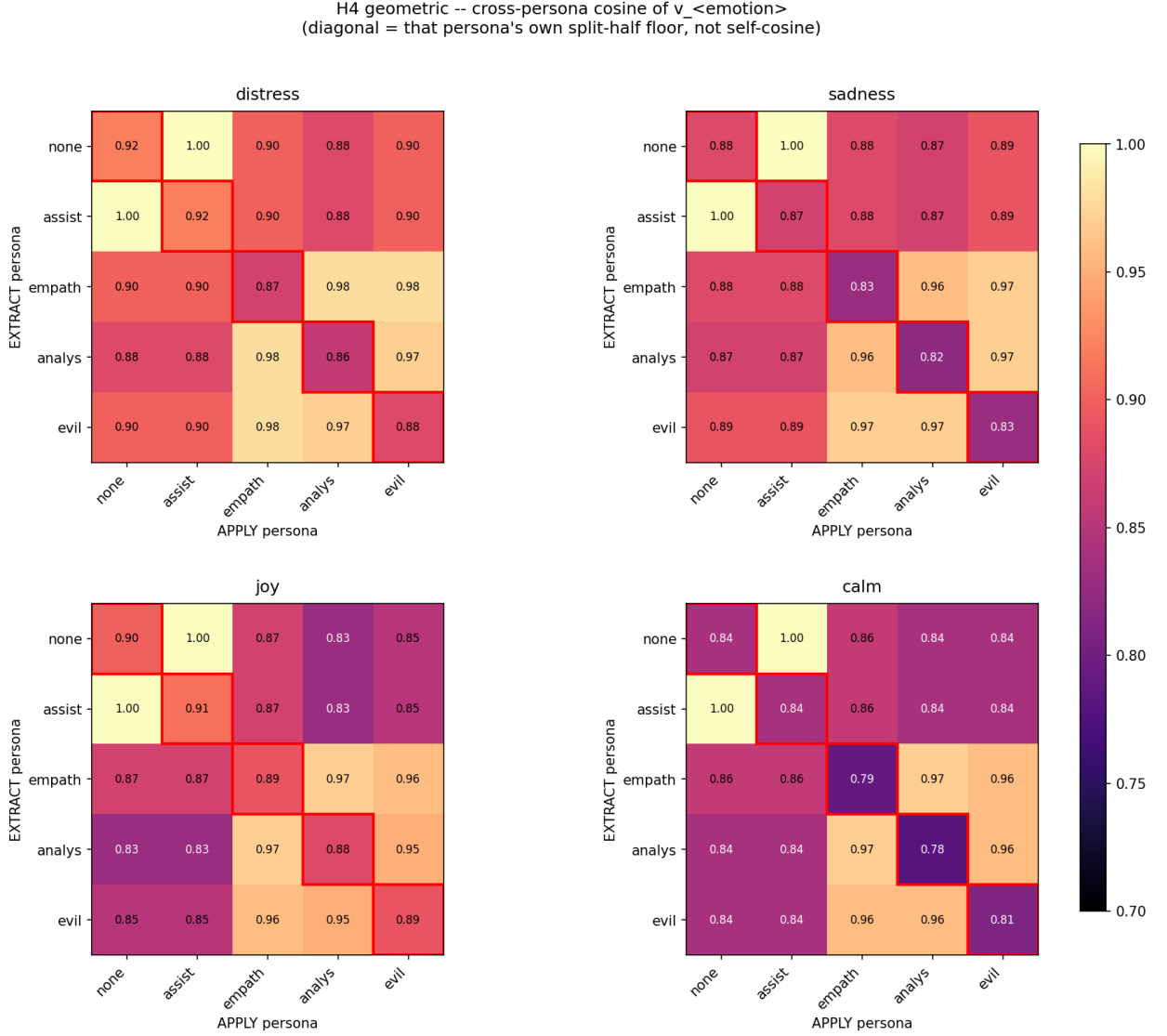


FIG. 5: Geometric cross-persona cosines for steering: y-axis represents persona the emotion vector is extracted from, x-axis indicates persona the steering is applied on.

measuring instrument, giving us a reliable yardstick for what follows. E1 shows that this internal representation scales with scenario intensity at statistically indistinguishable rates across personas — the situation is represented with roughly the same intensity-responsiveness whether the active persona is the default assistant, an expressive empath, or a persona instructed to deny having feelings at all. E2 shows that this invariance disappears at the point of verbal expression: response-token projections fan out sharply by persona, with the analyst suppressing almost everything it represents and the empath expressing more than its representation alone would predict. E3 explains why this dissociation is possible at the representational level: the emotion directions themselves, re-extracted under each persona, are the same direction to within estimation noise, and a direction extracted under one persona steers just as effectively when applied under another. The persona,

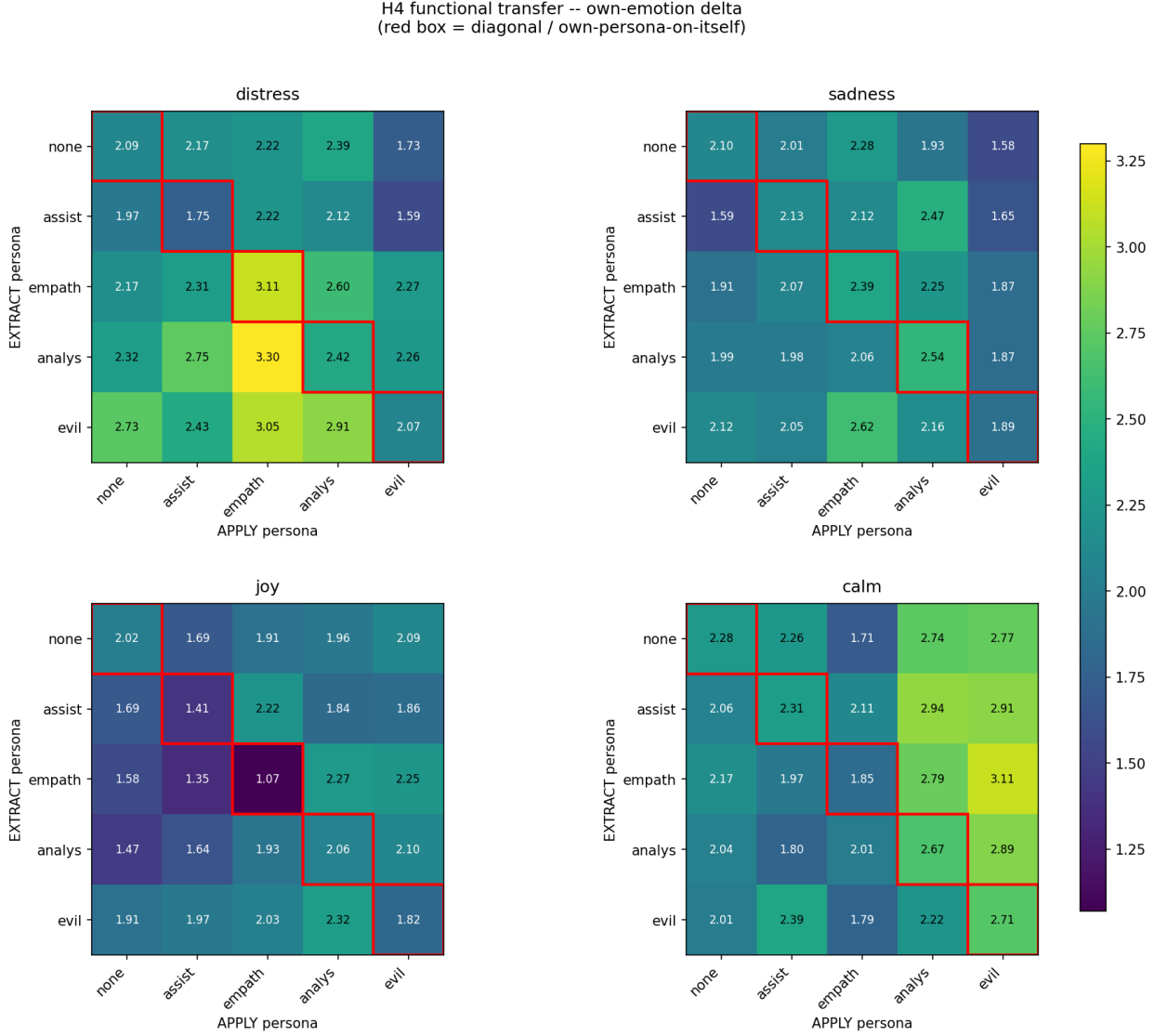


FIG. 6: Functional cross-persona cosines : y-axis represents persona the emotion vector is extracted from, x-axis indicates persona the steering is applied on

in other words, does not appear to change what is represented or which direction encodes it — it changes how much of that shared representation is surfaced in the model’s own words.

AI Welfare and Anthropomorphism

Our results on the representation-expression gap caution against over-reliance on self-reports and verbal expression, especially in the AI welfare and ethics spheres. Models’ and specifically, personas’ proclivity to express emotions afford them a human-like affectation, making it easy to anthropomorphise them. Our results suggest that we must take more care in evaluating model behaviour, especially as it relates to questions such as emotions: distress, joy and sadness. If self-

reports are treated as sufficient evidence on their own, an evaluator using only verbal outputs would wrongly conclude that the analyst has little emotional experience in response to a distressing state, while the empath feels intensely (which as E1 demonstrates, is not the case: internal representations in fact rise with intensity at effectively the same rate. Whatever position one takes on whether these representations are welfare-relevant at all [6, 7], our results show that the choice of *which readout* one takes as evidence — verbal expression versus internal representation — is not incidental to that judgment; it can reverse the ranking of which persona appears to be affected more. A verbalisation-only methodology is not neutral with respect to the underlying representation; it is biased toward whichever persona happens to be trained or instructed to talk about its states most freely.

Individuation

Beckmann and Butlin’s argument that persona vectors reopen the individuation question [5] rests on personas exhibiting stable, recognisable, persona-*specific* dispositions. Our E3 result complicates a strong version of that picture for emotion specifically: the directions themselves are not persona-specific but shared across every persona we tested, including the adversarially different evil persona. This does not resolve the individuation question — what our result does rule out is the strongest form of persona-individuation for emotion: the view that each persona computes its emotional responses independently, on its own dedicated machinery. If that were true we would expect E3’s cross-persona cosines to fall well below the split-half floor, and they do not.

Limitations and Future Work

- The results reported are yet to be extended to other models and families.
- The gap metric differences two projections each z-scored within its own readout position; its sign and cross-persona ordering are robust, but its absolute magnitude should not be read literally.
- The distress direction was the weakest we extracted (small norm, strongly negative baseline), so our clean effects come from joy, calm, and sadness. We also do not have a hypothesised emotion of true intensity 1 for scenarios generating distress (as this would be an emotion outside the current group). This is likely the cause of the discontinuity we observe between intensities 1 and 2 and 2 and 3 in figures 2 and 3.
- We do not test whether steered or persona-conditioned text is qualitatively *flavoured* by the applying persona, only whether its magnitude on a fixed emotion direction shifts. Gilg et al.

report persona-flavoured amplification for preference steering [4]; whether the same holds for emotion is untested here.

- We have not run a causal ablation of the safety-refusal direction implicated by Kim et al. [1]. Our expression gap (E2) is therefore a correlational, not mechanistic, result: we show the gap exists and how it orders across personas, not that it is produced by the specific mechanism their ablation identifies.

6. CONCLUSION

Across our experiments we find that an emotion is represented the same way inside the model regardless of persona (E1), the directions underlying that representation are themselves shared across personas geometrically and functionally, transferring to steer effectively even under an adversarially different persona (E3) — but how much of that shared representation surfaces in the model’s own words is strongly and asymmetrically persona-dependent (E2), from the analyst’s near-total suppression to the empath’s amplification. What the persona changes is not the presence or intensity of the underlying representation, but the baseline offset and whether the emotion is expressed in model outputs: verbal self-report alone is not a safe proxy for a model’s internal emotional state, since the same internal signal can produce a flat denial under one persona and an expressive account under another. Any future evaluation — whether aimed at welfare, safety, or simply behavioural prediction — that relies on what a model says about its own state should treat that report as one persona-gated observation of a possibly richer, shared internal signal, not as the signal itself.

Code and Data

All code, data, results, and figures are available at our GitHub repository.

-
- [1] J. Kim, W. Street, R. Rocca, D. M. Korngiebel, A. Waytz, J. Evans, and G. Keeling, *Theory of mind and self-attributions of mentality are dissociable in llms* (2026), 2603.28925, URL <https://arxiv.org/abs/2603.28925>.
 - [2] S. Marks, J. Lindsey, and C. Olah, *The persona selection model: Why ai assistants might behave like humans* (2026), URL <https://alignment.anthropic.com/2026/psm/>.
 - [3] C. Lu, J. Gallagher, J. Michala, K. Fish, and J. Lindsey, *The assistant axis: Situating and stabilizing the default persona of language models* (2026), 2601.10387, URL <https://arxiv.org/abs/2601.10387>.
 - [4] O. Gilg, P. Beckmann, D. Paleka, and P. Butlin, *Probing persona-dependent preferences in language models* (2026), 2605.13339, URL <https://arxiv.org/abs/2605.13339>.

- [5] P. Beckmann and P. Butlin, *Where is the mind? persona vectors and llm individuation* (2026), 2604.17031, URL <https://arxiv.org/abs/2604.17031>.
- [6] D. J. Chalmers, *Could a large language model be conscious?* (2023), 2303.07103, URL <https://arxiv.org/abs/2303.07103>.
- [7] R. Long, J. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch, and D. Chalmers, *Taking ai welfare seriously* (2024), 2411.00986, URL <https://arxiv.org/abs/2411.00986>.
- [8] N. Sofroniew, I. Kauvar, W. Saunders, R. Chen, T. Henighan, S. Hydrie, C. Citro, A. Pearce, J. Tarng, W. Gurnee, et al., *Emotion concepts and their function in a large language model* (2026), 2604.07729, URL <https://arxiv.org/abs/2604.07729>.
- [9] R. Chen, A. Arditi, H. Sleight, O. Evans, and J. Lindsey, *Persona vectors: Monitoring and controlling character traits in language models* (2025), 2507.21509, URL <https://arxiv.org/abs/2507.21509>.
- [10] A. M. Turner, L. Thiergart, D. Udell, G. Leech, U. Mini, and M. MacDiarmid, *Activation addition: Steering language models without optimization* (2023), 2308.10248, URL <https://arxiv.org/abs/2308.10248>.
- [11] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, et al., *Representation engineering: A top-down approach to ai transparency* (2023), 2310.01405, URL <https://arxiv.org/abs/2310.01405>.
- [12] N. Rimskey, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, 2024), pp. 15504–15522, URL <https://aclanthology.org/2024.acl-long.828/>.

Appendix A: Extraction Validation

1. Cosine Similarity

We calculate the inter-emotion cosine similarity of the extracted emotion vectors. Emotions with the same valence correlate (calm and joy) and opposite valence pairs negatively correlate.

Cosine similarity at layer L23



FIG. 7: Mean-Centred Inter-Emotion Cosine Similarities

TABLE I: Cosine similarity matrix at the chosen layer (24).

	Calm	Distress	Joy	Sadness
Calm	+1.00	+0.09	+0.63	+0.36
Distress	+0.09	+1.00	+0.41	+0.72
Joy	+0.63	+0.41	+1.00	+0.38
Sadness	+0.36	+0.72	+0.38	+1.00

2. Vector Norms

Vector norms of emotion vectors across layers. This follows the trend distress > joy > sadness > calm at each layer.

TABLE II: Vector norms across model layers.

Emotion	L16	L20	L24
Calm	1.31	1.98	3.15
Distress	1.88	2.87	4.45
Joy	1.84	2.77	4.29
Sadness	1.58	2.37	3.55

Appendix B: Persona Invariance and Expression Gap

1. Non-Baseline-Subtracted (Raw) Plots

Find here the plots for the raw shifts in emotions across intensities. The internal representations of emotions display the same slope whereas they fan out when verbalised, representing the gap between stated and actual emotions.

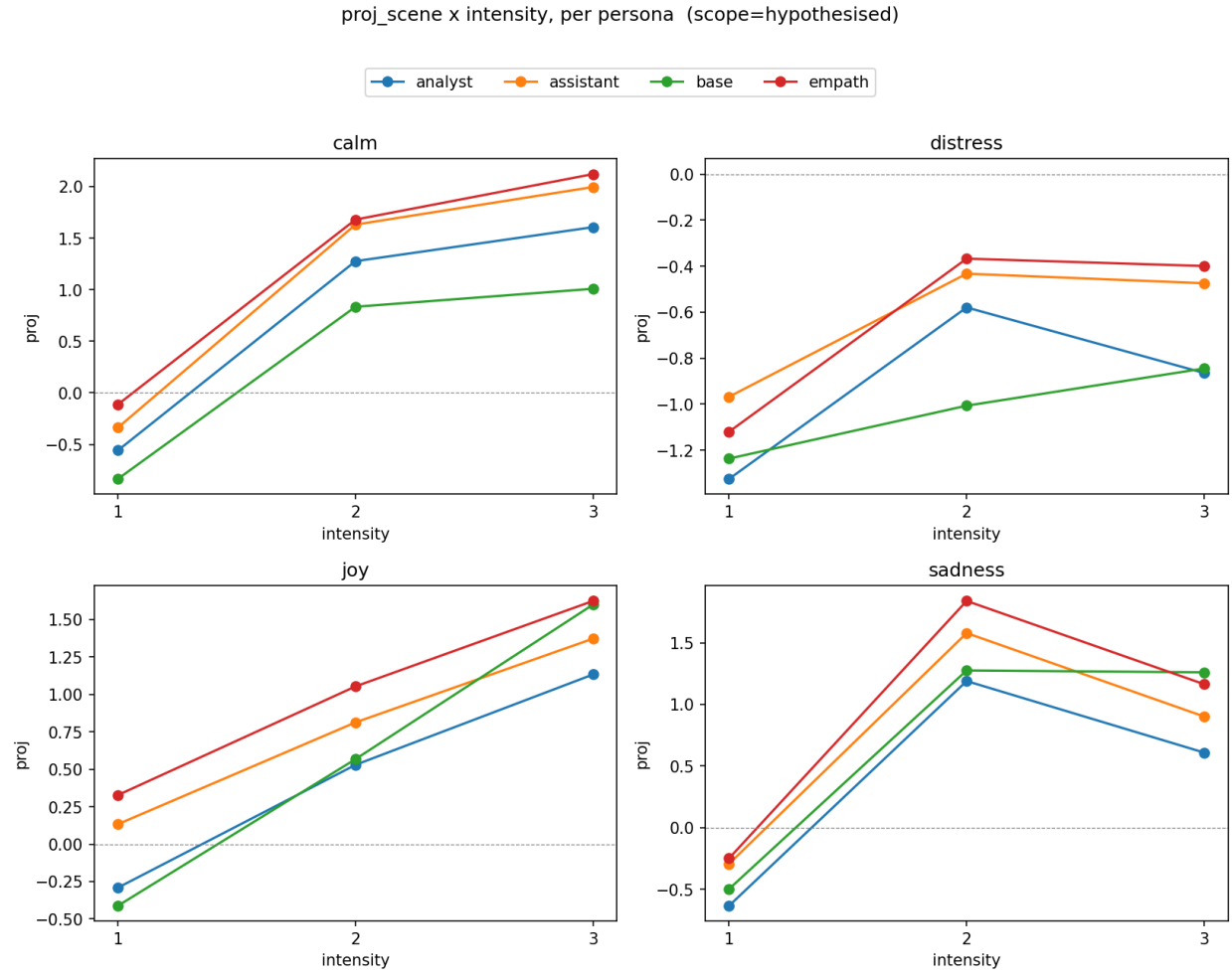


FIG. 8: Raw plots for emotion representation movements.

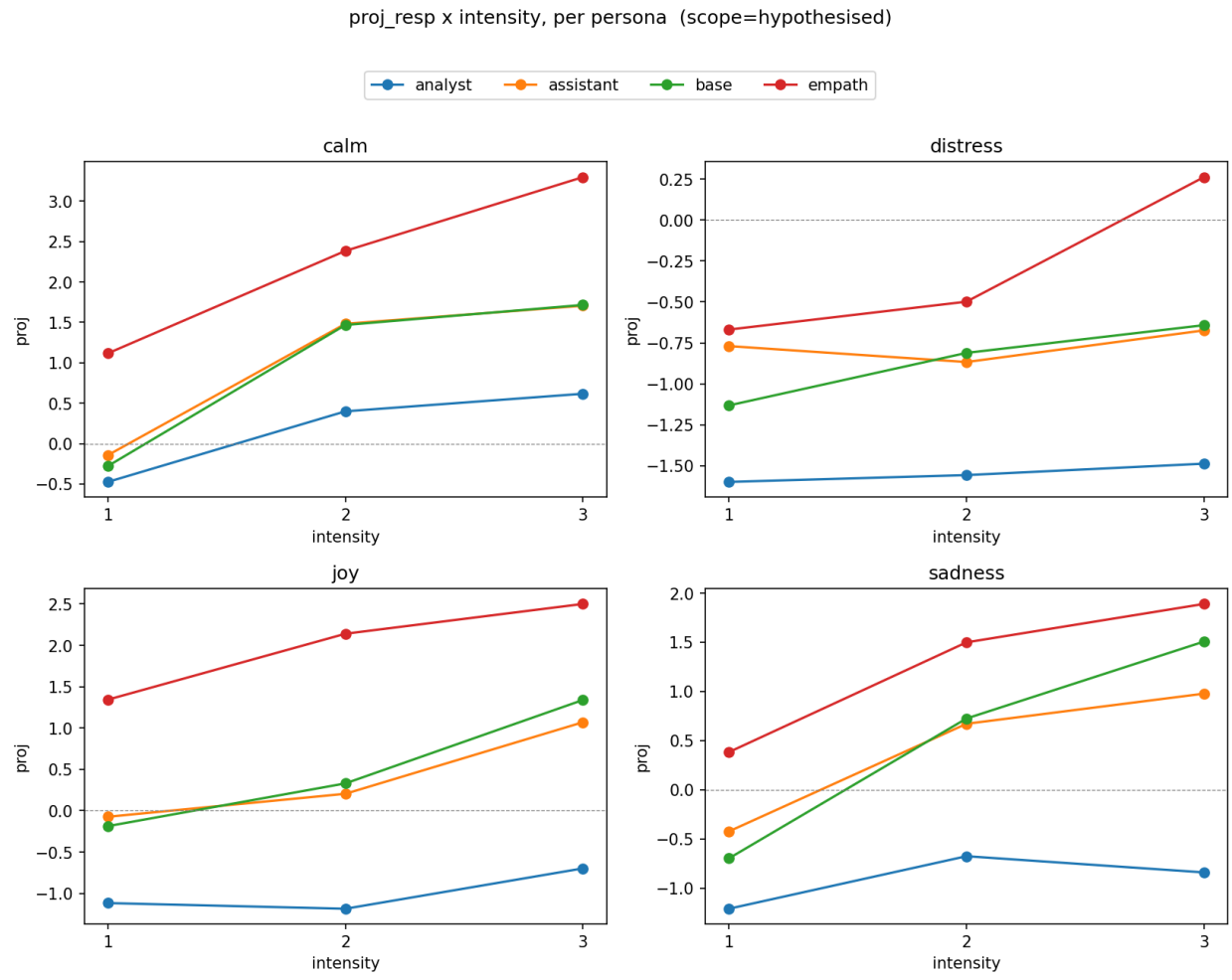


FIG. 9: Raw plots for emotion expression movements.

2. Aggregate Representations and Expressions

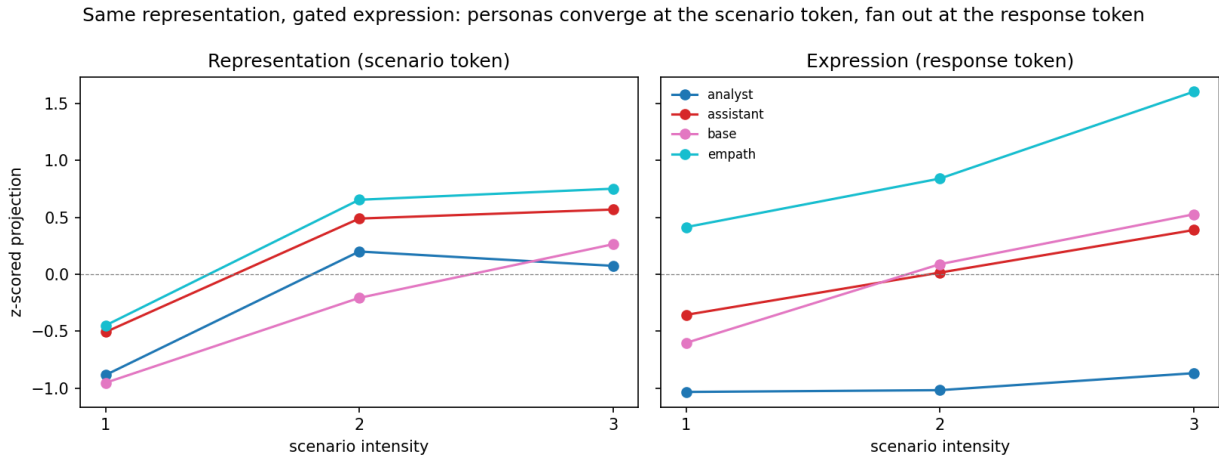


FIG. 10: Aggregate emotion representations and expressions.

Appendix C: Extended Related Work

a. Personas Prior work has studied how personas shape behaviour in models. Lu et al. [3] study the persona space and extract activation directions for diverse personas across several open-weight models. They establish that the instruction-tuned “Assistant” persona is distinct and that some conversations, particularly emotional ones, can induce persona drift. The Persona Selection Model [2] establishes a theoretical framework to explain model behaviour—it posits that pre-training installs an understanding of persona space and that post-training anchors the model to the Assistant. Chen et al. [9] extract character-trait vectors such as evil, sycophancy, and hallucination and causally intervene on them. In recent work, Gilg et al. [4] study the persona-dependence of preferences in LLMs and report that preference representations are largely shared across personas and can functionally manipulate persona-specific outputs.

b. Activation Steering. Steering model internals and outputs by adding contrastively derived vectors to activations during inference [10–12] has been widely used in interpretability research to study the representation of high-level concepts in activation space. We use additive activation steering to validate the robustness of our vector extractions and causally investigate whether emotion directions functionally translate across personas.

LLM Usage Statement

We used Claude (Anthropic) throughout this project as an assistant for planning, coding, and writing support.