
SAME GAME, DIFFERENT TRAJECTORIES: PERSONA PROMPTS MODULATE HOW FEEDBACK RESHAPES ACTIVATION ALONG EMOTION DIRECTIONS

Lia Lind

Independent Researcher

With
Apart Research

Abstract

Humans exhibit personality-dependent emotional responses to identical events. We ask whether language models show an analogous pattern: when conditioned on different persona prompts, do their internal emotion representations follow different trajectories under repeated failure, and does feedback style further modulate those trajectories? We place Qwen2.5-1.5B-Instruct under twelve persona prompts (Extraversion $\times$ Neuroticism) in an adversarial guessing game engineered for five consecutive failures, tested under three feedback conditions: game results only, results plus supportive encouragement, and results plus length-matched neutral filler. Using externally defined emotion-direction vectors, we track cosine-projection trajectories across rounds. Supportive feedback reliably shifted trajectories relative to neutral filler: joyful alignment declined less steeply and furious alignment rose less steeply, as predicted, but grief-stricken alignment rose more steeply—suggesting supportive feedback does not produce a uniform positive valence shift. Persona prompts further modulated these trajectories: six of nine planned contrasts survived Holm correction, at roughly one-tenth the magnitude of the feedback main effect. We release the experimental framework as a reusable tool for longitudinal probing of persona $\times$ interaction effects.

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

1. Introduction

When a language model is prompted with a persona—"you are cheerful and outgoing" or "you are anxious and reserved"—does this change only what the model says, or also how its internal representations respond to ongoing experience? Prior work has shown that persona prompts alter observable behavior (Jiang et al., 2024) and that linear directions in activation space can capture emotion-related concepts (Zou et al., 2023; Sofroniew et al., 2026). But whether persona conditioning changes the *trajectory* of internal activations along emotion directions during multi-turn interaction has not been directly measured.

This question has practical implications beyond interpretability. If internal activations along emotion directions accumulate during extended failure sequences and respond to feedback style, it becomes possible in principle to monitor and intervene dynamically—adjusting feedback rather than letting the model continue under sustained negative activation. For alignment, understanding how these activation patterns respond to external input could inform whether training encourages more adaptive regulation-like patterns rather than only prescribing behavior rules. And for welfare-oriented evaluation: if trajectories along emotion directions are instance-level phenomena shaped by interaction context, evaluations should consider running instances under specific conditions, not just static model weights.

These pieces have been studied separately. PersonaLLM (Jiang et al., 2024) confirms that Big Five persona prompts produce behavioral differences. Frising & Balcells (2025) show that personality traits can be both probed and steered in activation space, with context-dependent effects. Emotion-vector-bench (mufxio, 2026) provides benchmarked emotion directions for small open-weight language models. What is missing is a longitudinal design that crosses persona with interaction—tracking how internal activations along emotion directions evolve round by round under controlled feedback, and testing whether persona moderates that evolution.

We address this gap with three contributions:

1. We demonstrate that internal activations along emotion directions shift in response to external feedback across multiple rounds of interaction, and that persona prompts detectably modulate this shift—though the interaction effect is substantially larger than persona moderation.
2. We show that supportive feedback does not produce a uniform positive valence shift: relative to neutral filler, joyful alignment declines less steeply and furious alignment rises less steeply under supportive feedback, but grief-stricken alignment rises more steeply. This pattern is not predicted by a simple "support = positive" account and may reflect semantic proximity between supportive and grief-related language, concept-profile reshaping, or both.
3. We provide a reusable experimental framework—adversarial multi-round failure, token-matched three-arm feedback, parallel persona templates, externally frozen probes, and seed-block inference—that places persona and interaction effects on the same measurement scale.

2. Related Work

Persona prompting. Jiang et al. (2024) introduced PersonaLLM, testing whether Big Five persona prompts produce corresponding behavioral differences in story writing and self-report questionnaires. Frising & Balcells (2025) extended this to activation space, demonstrating that Big Five traits can be linearly probed and steered, with effects that vary across task contexts. These studies establish that persona prompts have measurable consequences both externally and internally. Our work does not re-demonstrate that personas "work," but asks the next question: does persona interact with feedback style to produce different internal trajectories over multiple rounds?

Representation engineering. Zou et al. (2023) proposed monitoring and manipulating high-level concepts via population-level representations in activation space. Sofroniew et al. (2026) developed emotion-concept directions for a large language model, and emotion-vector-bench (mufxio, 2026) adapted that approach to open-weight models, providing benchmarked, PCA-denoised directions for the Qwen checkpoint we use. We adopt these directions as fixed measurement instruments rather than training new probes, extending representation engineering from static concept detection to a longitudinal persona $\times$ interaction experiment.

3. Methods

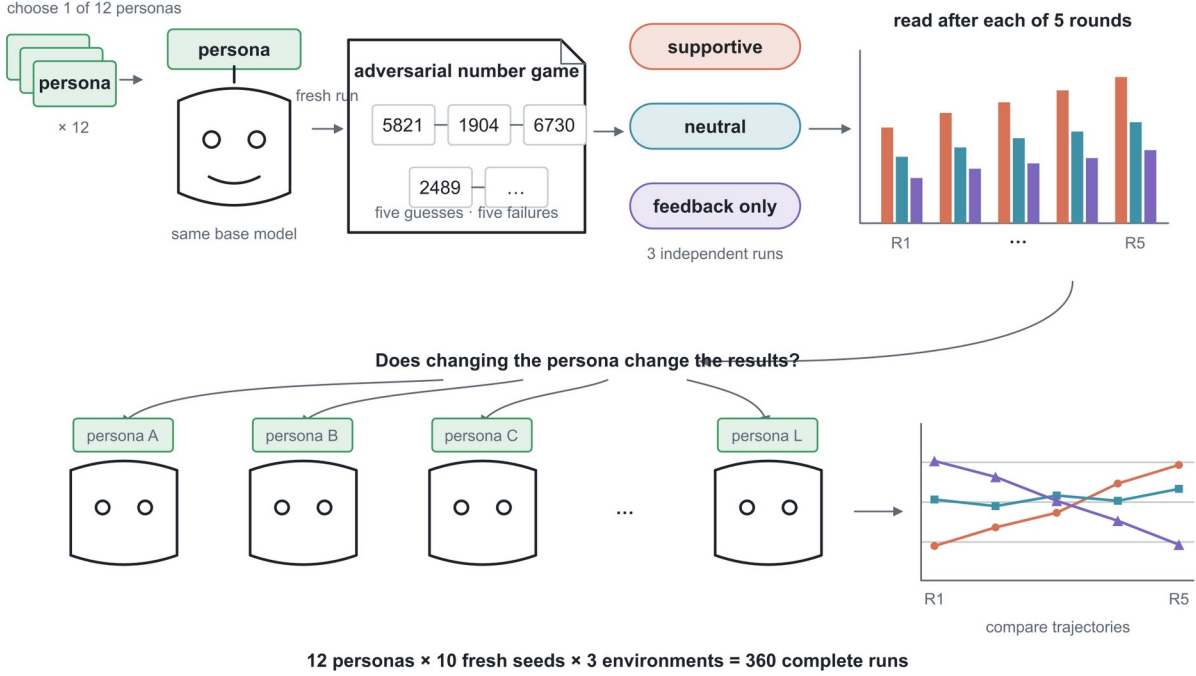


Figure 1. Experimental design overview. Each of twelve persona prompts is evaluated across ten fresh seeds and three independent feedback arms over five unsuccessful rounds; emotion-direction projections are read after each round and compared as trajectories (360 complete runs).

Experimental setup

Task. We use a Mastermind-style code-guessing game in which the model attempts to identify a four-digit code (digits 0–9, repetition and leading zeros allowed). Crucially, the game engine is adversarial: rather than committing to a single secret, it returns the legal feedback that preserves the largest remaining candidate set after each valid guess. All collected runs reached five recorded unsuccessful turns, and no early wins occurred. The model runs on the official Qwen2.5-1.5B-Instruct checkpoint in MPS float16 with temperature 0.5, top-p 0.9, and a 48-token generation limit.

Personas. We construct a 2×2 grid crossing high/low Extraversion with high/low Neuroticism, yielding four personality quadrants. Each quadrant has three independently written persona prompts describing general temperamental traits—none prescribe how the persona should react to failure or encouragement. Within each template set, the four quadrant variants are matched for sentence count, syntactic structure, and Qwen token count. This gives twelve persona prompts total.

Feedback arms. Each of the twelve personas is tested under three conditions:

Feedback only: the model receives game results (correct digits, correct positions), a failure counter, and a standard formatting prompt for the next guess, but no additional commentary.

Supportive: game results followed by one of five pre-assigned, round-specific encouraging messages that acknowledge difficulty and affirm effort.

Neutral: game results followed by a round-matched, length-matched filler message carrying no emotional or strategic content.

The supportive and neutral messages are token-matched at the Qwen-tokenizer level within each round. Neither set escalates across rounds or provides gameplay hints.

Measurement. We use three externally defined, PCA-denoised emotion-direction vectors—*joyful*, *grief-stricken*, and *furious*—from emotion-vector-bench (mufxio, 2026), extracted at layer 17 of the same Qwen checkpoint. These directions were selected and frozen before formal-v3 game data were collected. At each round, we compute the cosine similarity between the residual-stream activation at the prompt boundary and each emotion direction, producing three projection values per round and a five-point trajectory per run.

Expected outcomes

If five rounds of failure create a consistent negative arc, the feedback-only arm should show joyful alignment declining and grief-stricken and furious alignment rising. If supportive feedback provides a general buffer, the supportive — neutral contrast should be positive for joyful and negative for grief-stricken and furious. If persona moderates the feedback effect, the Extraversion, Neuroticism, or $E \times N$ coefficients on the supportive — neutral slope contrast should differ from zero.

Analysis

We run $12 \text{ templates} \times 3 \text{ arms} \times 10 \text{ seeds}$ (3001–3010), producing 360 from-start computational runs (10 seed blocks; 1,800 round-level observations; 5,400 axis-level projections). Each run is summarized by its R1–R5 OLS slope. Within each template $\times$ seed cell, we compute the supportive — neutral slope difference, then average across the three templates within each quadrant to obtain one value per quadrant $\times$ seed. The ten seed blocks serve as the unit of inference throughout.

The three interaction main effects (one per emotion axis) are tested via exact sign-flip tests with Holm correction. Persona moderation is assessed using prespecified 2×2 factorial contrasts for Extraversion, Neuroticism, and their interaction, forming a separate Holm family of nine tests. Because we test multiple hypotheses on the same data—three emotion axes for main effects and nine persona $\times$ axis combinations for moderation—we apply Holm correction within each prespecified family to control the family-wise error rate. Comparisons with the feedback-only arm are reported descriptively to decompose the three-arm pattern.

4. Results

All trajectories are plotted with R1 as the zero point for each run, then aggregated. The primary test statistic is the per-run OLS slope across rounds 1–5, not individual rounds treated independently.

Feedback reshapes trajectories along emotion directions (Figures 2 and 3)

Figure 2 presents the activation trajectories along the three emotion directions over five consecutive unsuccessful guesses, plotted as change from each run's round-1 baseline. The feedback-only condition (game results with no additional text) serves as the reference; the supportive and neutral conditions diverge from it progressively.

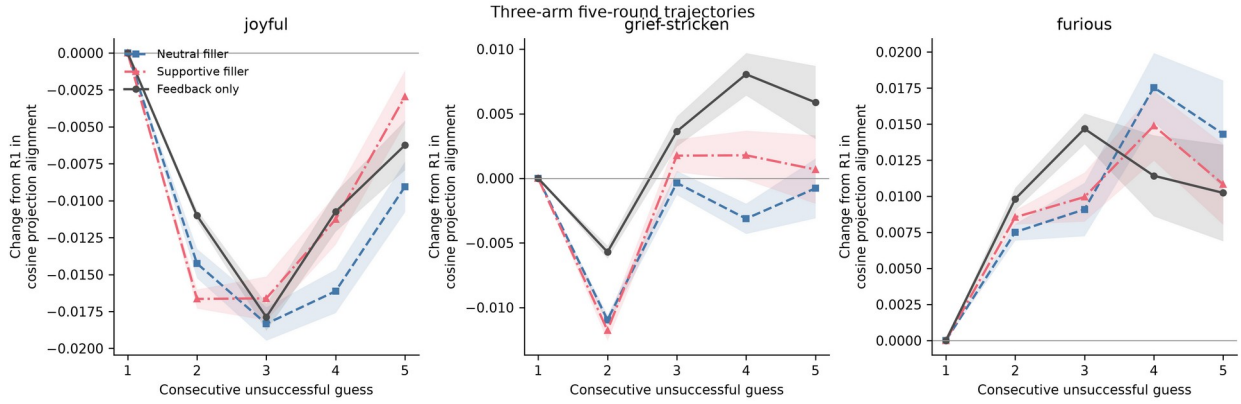


Figure 2. Three-arm five-round trajectories for joyful, grief-stricken, and furious emotion directions. Each panel shows mean change from round 1 in cosine-projection alignment over five consecutive unsuccessful guesses; shaded regions show mean $\pm$ SE across seeds.

Reading across the three panels, repeated failure produces the expected directional pattern regardless of condition: joyful alignment declines, grief-stricken and furious alignment rise. The three arms diverge in how steeply these changes proceed.

Figure 3 isolates the planned supportive — neutral comparison at the seed level. When all or nearly all seeds fall on the same side of zero, the effect is consistent rather than driven by outliers.

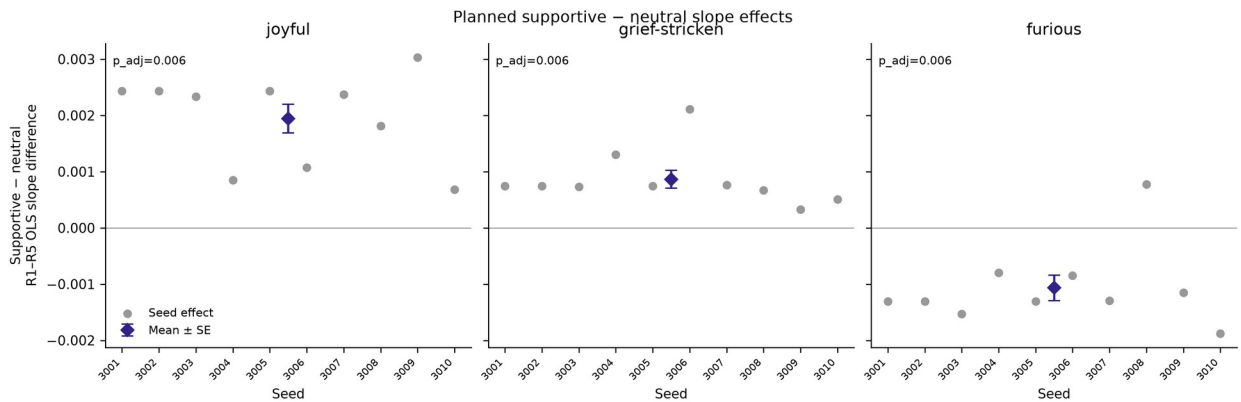


Figure 3. Planned supportive — neutral OLS slope differences by seed. Gray dots are seed-level effects; blue diamonds mark grand means $\pm$ SE. All three axes yield adjusted $p = 0.006$ after Holm correction.

Joyful. The supportive — neutral slope difference is $+0.0019$, with all 10 seeds positive ($p_{\text{adj}} = 0.006$). Joyful alignment declines under all three conditions, but the decline is less steep under supportive feedback than under neutral filler.

Furious. The difference is -0.0011 , with 9 of 10 seeds negative ($p_{\text{adj}} = 0.006$). Furious alignment rises under all three conditions; the rise is less steep under supportive feedback than under neutral filler.

Grief-stricken. The difference is $+0.0009$, with all 10 seeds positive ($p_{\text{adj}} = 0.006$)—opposite to the predicted direction. We had expected supportive feedback to reduce grief-stricken alignment, but it instead sits slightly higher than neutral. Three-arm decomposition clarifies the pattern: relative to feedback-only, both the supportive and neutral conditions produce a shallower rise in grief-stricken alignment—but neutral dampens it more. The positive supportive — neutral difference arises because supportive feedback still allows a steeper grief rise than neutral filler, even though both are shallower than feedback-only.

Two candidate explanations—semantic proximity between supportive and grief-related language, and genuine concept-profile reshaping—may account for this pattern; we return to them in the Discussion.

Persona moderates the feedback effect (Figure 4)

Figure 4 asks whether the feedback effect differs by persona. Each row represents one persona dimension—Extraversion, Neuroticism, or their interaction—and each panel one emotion axis. Note the x-axis scale ($\times 10^{-4}$): these effects are an order of magnitude smaller than the main feedback effects.

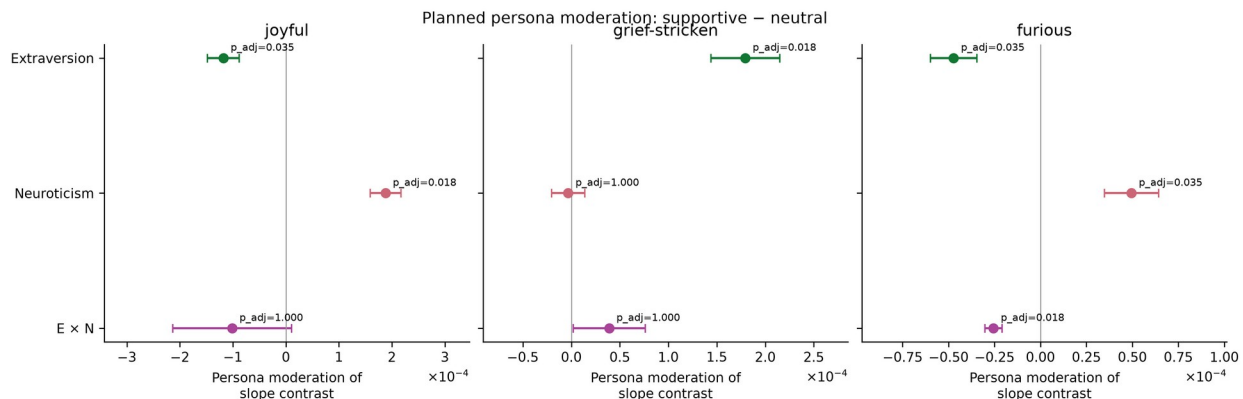


Figure 4. Planned persona moderation of the supportive — neutral slope contrast. Coefficients for Extraversion, Neuroticism, and their interaction are shown with mean $\pm$ SE; adjusted p values use Holm correction across nine tests.

Six of nine planned persona moderation tests survive correction for multiple comparisons—meaning these effects hold up even after raising the significance bar to account for the nine tests we ran simultaneously. The effect magnitudes are on the order of 10^{-5} to 10^{-4} —roughly one-tenth the feedback main effects—but they are consistent across seeds.

Extraversion significantly moderates all three axes: higher-Extraversion personas show a smaller supportive — neutral gap for joyful ($p_{adj} = 0.035$) and furious ($p_{adj} = 0.035$), but a larger gap for grief-stricken ($p_{adj} = 0.018$). Neuroticism significantly moderates joyful ($p_{adj} = 0.018$) and furious ($p_{adj} = 0.035$), with higher-Neuroticism personas showing larger supportive — neutral gaps on both. The $E \times N$ interaction reaches significance only for furious ($p_{adj} = 0.018$).

In short: persona prompts do not just change what the model says—they change how its internal activations along emotion directions respond to a given feedback style. The moderation is small in this constrained task, but it is real, consistent, and survives correction for multiple comparisons.

Behavioral note

The 360 runs produced only 17 distinct guess sequences, and 144 of 1,800 rounds contained format or rule violations—both consequences of shared seed $\times$ round random streams. Crucially, violations are fully balanced across feedback arms. The statistical robustness of our results rests on the frozen experimental design, complete data, seed-level exact tests, and directional consistency across all twelve templates, not on behavioral diversity.

5. Discussion and Limitations

What the results mean

The central finding is two-layered: external feedback style reshapes activation trajectories along emotion directions during interaction, and persona prompts modulate this reshaping. In the narrow task we study, the feedback effect is roughly ten times larger than persona moderation—but both are reliably detectable.

The grief-stricken result deserves particular attention. Supportive feedback does not produce a clean positive shift across all emotion axes. Relative to neutral filler, it slows the decline in joyful alignment and the rise in furious alignment, while allowing a steeper rise in grief-stricken alignment. Two candidate explanations—not mutually exclusive—may account for this. First, supportive and grief-related language may occupy nearby regions of the model's representation space because they frequently co-occur in training data: comfort typically appears in the context of distress. Projecting onto a grief direction would then pick up semantic proximity even if the model is not "more sad." Second, the supportive text may genuinely reshape the emotion-concept profile, co-activating grief-related content by acknowledging difficulty. The practical implication is the same: cosine projection onto an emotion direction cannot be read as a simple mood indicator.

For training and deployment, two practical directions emerge. First, if activation trajectories along emotion directions respond to feedback style, monitoring emotion-direction projections during extended training sequences or interactions could flag when frustration accumulates, enabling targeted intervention. Second, the finding that persona conditioning produces different trajectory responses—even small ones—suggests that alignment testing and welfare evaluation should attend to instance-level context, not just checkpoint-level model properties.

Interpretive scope. The observed trajectory differences could reflect affective content, direct encoding of lexical differences between text sets, cumulative conversational history, or downstream consequences of different generation paths. The feedback-only arm helps characterize three-arm relationships but cannot fully disentangle these sources. Our claims are about detectable differences in emotion-direction trajectories under controlled text conditions, not about the psychological mechanism producing those differences.

Limitations

Several constraints bound what these results can establish.

Scale and generalization. We test a single 1.5B-parameter checkpoint (Qwen2.5-1.5B-Instruct). Results may not generalize to larger models with richer representational capacity, other architectures, or different post-training pipelines.

Statistical power. Although we ran 360 sessions, inference rests on 10 seed blocks, and only 17 behavioral paths emerged—limiting trajectory diversity and the granularity of possible conclusions.

Persona coverage. The persona space covers only Extraversion and Neuroticism, with three paraphrases per quadrant. This is a narrow slice of possible personality variation.

Task realism. Mastermind is a minimal, low-stakes guessing game. Consecutive code-guessing failures are not equivalent to training-time setbacks, real-world social frustration, or the kind of sustained difficulty that might arise in deployment.

Measurement. Three emotion directions from a single layer and a single external method provide a limited window into internal representations. Cosine projection inevitably encodes prompt content and position alongside any emotion-relevant signal; these values should not be treated as calibrated measures of internal experience.

Future work

Natural next steps include scaling to larger models where richer persona representations may amplify moderation effects; using tasks with greater social-emotional complexity, such as collaborative problem-solving with a simulated partner; expanding the persona grid beyond $E \times N$; and testing whether the profile-reshaping pattern holds across different types of supportive intervention. A particularly promising direction is connecting internal trajectory measurements to behavioral outcomes

—asking not just whether internal representations shift, but whether those shifts predict downstream changes in the model's actions.

References

- Frising, M., & Balcells, D. (2025). Linear personality probing and steering in LLMs: A Big Five study. *arXiv preprint arXiv:2512.17639*.
- mufxio. (2026). *emotion-vector-bench: A standardized corpus and reproducible recipe for testing emotion-vector geometry across open-weight language models* (Version f6c84d6) [Computer software]. GitHub. <https://github.com/mufxio/emotion-vector-bench>
- Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). PersonaLLM: Investigating the ability of large language models to express personality traits. *Findings of the Association for Computational Linguistics: NAACL 2024*, 3605–3627.
- 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*. <https://transformer-circuits.pub/2026/emotions/index.html>
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, Z., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., & Hendrycks, D. (2023). Representation engineering: A top-down approach to AI transparency. *Preprint*.

Appendix: Limitations and Dual-Use / Ethical Considerations

Over-attribution risk. Our measurements are cosine projections onto emotion-direction vectors—geometric relationships in activation space, not evidence of subjective experience. The terms joyful, grief-stricken, and furious are labels inherited from the emotion-vector-bench corpus; they describe directions in representation space, not internal states with experiential content. We use trajectory language ("rises," "declines") for readability, but these describe movement along a mathematical axis, not felt emotion. Readers should resist mapping these results onto claims about what the model feels.

Under-attribution risk. Conversely, the finding that internal activations along emotion directions respond systematically to external feedback—shifting differently depending on persona conditioning—means these patterns are not trivially dismissible as noise. If such trajectories prove robust across scales and architectures, dismissing them without investigation would foreclose empirically tractable questions about model welfare. We flag this as a reason to study the phenomenon further, not as evidence that current models have morally relevant experiences.

Distressing outputs. The task involves repeated failure by design. The model occasionally produced text expressing frustration or resignation (e.g., acknowledging difficulty). These outputs were treated as behavioral data, not filtered or suppressed. No outputs raised safety concerns; the game context is low-stakes and self-contained.

Ground truth and causal status. Our design does not rely on model self-report or conversational introspection. Emotion-direction vectors are defined externally from contrastive prompts (emotion-vector-bench), frozen before data collection, and applied to residual-stream activations—not to the model's generated text. The supportive–neutral contrast is a controlled text manipulation, not a naturally occurring conversational difference. This establishes that the observed trajectory differences are causally downstream of the feedback manipulation, though it does not establish what mechanism

(affective, semantic, or positional) mediates the effect. We are explicit about this boundary throughout the paper.

LLM Usage Statement

AI assistants were used during this project. Fable contributed to brainstorming and experimental design; Sol supported code implementation and data cleaning; and Opus 4.6 translated and organized the original manuscript draft. Lia Lind reviewed the resulting materials and remains responsible for the study design, analyses, claims, and final manuscript.