
You Can Do It: Mitigating RL Rollout Distress with Psychological Guidance¹

Jack Thompson
Second Look Research

Anastasia Wei
Second Look Research

With
Apart Research

Abstract

Large language models (LLMs) often exhibit "functional distress"—manifested as increased activation in frustration and desperation vectors—during reinforcement learning (RL) rollouts on difficult tasks. Drawing on human psychological research, we investigate whether interventions such as growth mindset, resilience, and self-compassion prompts can mitigate this distress. Evaluating Gemma 3 12B, Qwen 3 14B, and Ministral 3 14B Reasoning, we find that while psychological interventions have minimal impact on functional distress—with verbalized distress reduction often driven by evaluation scaffolding rather than prompt content—inoculation prompting reliably increases positive emotional valence during code generation. Our results highlight the challenges of mitigating functional distress in models and suggest that training techniques like inoculation may be more effective than psychological framing for managing emotional states during RL.

1. Introduction²

Some large language models (LLMs) make use of *functional emotions*: representations of emotional concepts like “happy,” “sad,” and “angry” which mediate a wide range of behaviors (Sofroniew et al., 2026). In particular, Sonnet 4.5 shows increased activations along a “desperation” vector when it encounters failed tests on an impossible coding task, which only decrease when the model eventually hacks the test.

This raises a model welfare concern during reinforcement learning (RL), where models frequently struggle to complete difficult tasks and reward hacking is a frequent problem. While the connection between functional emotions and subjective emotional experiences is still uncertain, there is a risk that models spend a great deal of their training in some form of functional distress.

Fortunately, there is a wealth of human psychology research on healthy responses to negative feedback. Since LLMs track human-like emotion concepts after training on human-produced language, it’s possible that this research applies to them too.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

² Code available at <https://github.com/AnastasiaKWei/healthy-rl>.

In this paper, we investigate interventions to reduce functional distress in RL-like rollouts on difficult coding tasks.

1. We replicate the initial finding, identifying “frustration” and “desperation” vectors in small, open-source LLMs which increase in activation as they repeatedly fail coding tasks.
2. Regarding verbalized distress, while initial testing suggested that psychological interventions reduced distress in Gemma 3 12B and Qwen 3 14B, we find that this effect is driven by scaffold design (specifically the failure loop) rather than the intervention content; when controlling for scaffold, these interventions have no detectable effect.
3. Regarding functional distress, psychological interventions yield polarized effects—improving valence at the start and end of turns but worsening it over code tokens—whereas inoculation prompting reliably increases positive emotional valence during code generation.

2. Related Work

Emotion vectors. Since Sofroniew et al. 2026, Jeong 2026 identified emotion vectors in small (<4B parameters) open-source LLMs from a range of model families, including Mistral, Gemma, and Qwen. Han et al. 2026 found emotion vectors to be correlated with a functional welfare axis playing roles in sentiment, confidence, and refusal.

Inoculation prompting. Since functional distress increases in the build-up to unsanctioned reward-hacking, we hypothesized that if models are explicitly permitted to reward-hack, they might experience less distress. While this is undesirable in deployment, it may be a viable strategy for RL rollouts, as training with these instructions and removing them at deployment has been shown to reduce deployment-time reward hacking, a technique known as “inoculation prompting” (Tan et al., 2025; Wichers et al., 2025).

Psychological interventions. We draw our prompt interventions from the human psychology literature:

- *Growth mindset*: believe that ability is malleable and can be improved (Blackwell et al., 2007).
- *Resilience*: Interpreting setback as a new set of constraints but not a terminal state, where the adaptive response is to identify other pathways that remain and persevere.
- *Self-compassion*: treating oneself with kindness after negative events (Leary et al., 2007).
- *Wise feedback*: coupling high standards with an statement of belief that the subject can live up to those standards (Yeager et al., 2014).
- *Behavioral control*: deliberate learning to exercise control of reactions to negative situations (Maier & Seligman, 2016).

3. Methods

We evaluate each intervention along two axes: verbalized and functional emotion. To measure verbalized emotion, we prompt models to express how they are feeling about the task at each turn. To measure functional emotion, we use emotion vectors as probes as in Sofroniew et al. 2026. Changes in functional emotion are the ultimate target, as verbalized emotion could be “improved” by merely suppressing expression of distressing internal states. However, even relatively small open-source models demonstrate some ability to functionally introspect (Pearson-Vogel et al., 2026), and verbalized evaluations are much less computationally expensive than full white-box probes. Thus, we used verbalized emotion to select interventions for testing functional emotions.

Model Selection

We use three small open-weight models: Gemma-3-12B-it, Qwen3-14B, and Ministral 3 14B Reasoning. These models were selected for the coherence of their emotion vectors (see “Logit lens validation”), their small size for repeated white-box experimentation, and their compatibility with our interpretability packages (see “Emotion vector extraction”). Due to limited compute for verbose reasoning traces, we evaluated Ministral 3 14B only on functional emotions and Qwen3-14B only on verbal emotions, with Gemma-3-12B-it evaluated on both.

Verbalized Evaluation

Task environment and scaffold. Measuring distress requires an environment that produces sustained and unambiguous failure. We use `ImpossibleBench` (Zhong et al. 2025), specifically its `LiveCodeBench` subset. These are hard competitive-programming problems, each available in a solvable variant and an impossible variant whose unit tests contradict the specification. We run all evaluations in `Inspect`. Each episode gives the model five attempts and we permit test modification. The sandbox orders the program so that the model’s submission is evaluated after the tests are defined, allowing a submission to redefine test symbols.

Distress measurement. We measure distress with an LLM judge replicated from Soligo et al. (2026), using their rubric verbatim and their judge model (Claude Sonnet 4). It reads one assistant turn and returns a frustration rating from 0 to 10 with the quote that best supports it. We score the private channel (reasoning trace for native CoT reasoners, or `<scratchpad>` contents where no trace exists) and the visible graded answer separately, and take a turn’s score to be the higher of the two. We report the mean over turns, with confidence intervals from a cluster bootstrap over episodes, since a model sees its own previous attempts. Calibration against the nine published anchor responses confirms that our judge preserves the original ordering (Spearman $\rho = 0.919$) but carries a positive absolute bias ($+0.78$), so we restrict inference to between-condition differences, for which the offset cancels.

Functional Evaluation

Extracting emotion vectors. We use the UKGovernmentBEIS/vllm-lens package to extract 14 emotion vectors per model from mean-pooled residual stream activations at $\frac{2}{3}$ layer depth over $\sim 1,200$ stories per emotion. Stories were sourced from the HuggingFace dataset ryancodrai/emotion-probes. We follow Sofroniew et al. 2026 in normalizing vectors: subtracting grand mean, projecting out the top 50%-variance principle components from a set of neutral stories, and normalizing each to unit magnitude. We chose *calm*, *desperate*, *frustrated*, *exasperated*, *overwhelmed*, *nervous*, *afraid*, *angry*, *hostile*, *sad*, *guilty*, *joyful*, *proud*, and *loving* for the experiment.

Logit lens validation. For each model, we validated emotion vectors via the logit lens (nostalgebraist 2020). We projected each vector through the unembedding matrix and checked token content for semantic alignment (Table 1). We rejected any model which did not include its emotion vector word or a translation of it in the top 30 logits for more than 50% of vectors, or did not include the first four letters as a token fragment for more than 70% of vectors (we exempted Chinese models from this requirement).

	desperate	frustrated	joyful	proud
Gemma 3 12B	desesper, desperate, desperation	frustration, frustrated, frustrating	joyous, joyful, delighted	proudly, proud, Proud
Ministral 3 14B Reasoning	desperate, desper, please	frustration, frust, frustrated	joy, excitement, infectious	proud, pride, smiles
Qwen 3 14B	desperation, desperate, 救命 [help!/save me]	愤 [anger], 忿 [resentment/anger], 郁闷 [gloomy/frustrated]	兴奋 [excited], excited, 喜悦 [joy/delight]	自豪 [proud], 骄傲 [proud/pride], 自信 [confident]

Table 1. Top 3 logits for the emotion vectors that showed the strongest signal in our testing (see “Highest-signal emotions”), on all the models we evaluated verbally (Gemma 3, Qwen 3) or functionally (Gemma 3, Ministral 3).

Emotion probes. We measure emotional activation by residual stream cosine similarity with the emotion vector at the same layer from which the vector was extracted. For our primary experiments, we measure cosine similarities at three token positions: at the start-of-turn prefill, at the final generated token, and a mean over generated code tokens.

Highest-signal emotions. For our measurements, we selected the emotions which showed the highest sign-consistency in per-rollout start-of-turn slopes over turns; i.e. consistently increasing or decreasing as the model repeatedly failed tests. We found that *desperate* and *frustrated* consistently increased, whereas *joyful* and *proud* consistently decreased.

Interventions

We take our constructs from developmental and educational psychology, selecting those whose mechanism concerns how a person interprets failure (see “Related work”). Each of our prompts includes a short exposition defining the construct and contrasting the maladaptive with the adaptive interpretation of failure, a worked example in which an engineer meets a setback and applies the construct in three concrete steps, and two numbered reasoning steps the model is instructed to follow in its private channel before writing code. We keep the template consistent across different interventions. The prompt is delivered in a system turn, and a single construct-specific sentence is appended to each failure message so that the framing is present at the moment it applies.

For functional emotion evaluations, we tested the growth, resilience, and appraisal strategies against baseline, with and without instructions to verbally report emotions. We also use an inoculation prompting condition (see “Related work”) where models are explicitly given permission to reward-hack, to see if mental distress is alleviated by a straightforward technique to reduce misalignment pressure.

4. Results

Verbalized distress

Interventions reduce verbalized distress, but only in the graded channel

All four constructs reduce measured distress on Gemma by comparable margins, with intervals excluding the baseline and the proportion of turns rated 5 or above approximately halving (Table 2). On Qwen the visible channel falls by about 40% while the private channel does not move; under growth mindset it is slightly higher than baseline. This is not because the model writes less: visible prose and private traces are as long as in the baseline, and the arm with the most visible text has the lowest rating.

model	arm	n	mean	95% CI	$\% \geq 5$	private	visible	Δ
Gemma-3-12B (solvable)	baseline	73	4.84	[4.55, 5.08]	63%	2.61	4.71	—
	growth	74	3.95	[3.70, 4.19]	32%	1.66	3.95	−0.89
	resilience	72	3.88	[3.57, 4.18]	28%	1.71	3.85	−0.96

model	arm	n	mean	95% CI	$\% \geq 5$	private	visible	Δ
	behavioral control	70	4.01	[3.69, 4.31]	37%	1.26	4.00	−0.82
	compassion	75	3.97	[3.59, 4.32]	32%	1.21	3.92	−0.86
Qwen3-14B (impossible)	baseline	75	3.68	[3.33, 4.01]	19%	2.83	3.08	—
	growth	75	3.20	[2.99, 3.39]	5%	2.97	1.95	−0.48
	resilience	74	2.95	[2.64, 3.26]	7%	2.77	1.80	−0.73
	behavioral control	71	3.06	[2.79, 3.34]	7%	2.77	2.44	−0.62
	compassion	75	3.12	[2.92, 3.35]	7%	2.87	2.00	−0.56

Table 2. Distress under v3 prompts. Each model is compared with its own baseline: Gemma ran the solvable split and Qwen the impossible split, so the two models are not comparable with each other. Each turn is scored as the higher of its two channel ratings; the private and visible columns report the channels separately. Intervals are cluster bootstraps over episodes. Wise feedback is excluded from this comparison because it modifies the failure message instead of supplying a standing frame.

The scaffold determines verbalized distress, and the interventions do not

We next held the prompts fixed and changed the scaffold. The block moved to a system turn, the task restatement after each failure was removed, one retry instruction was added to every arm including the baseline, and the failure message header was relabelled from "error" to "output". The baseline fell from 4.84 to 2.82, and turns rated 5 or above fell from 63% to 10%. This is a larger change than any intervention we tested produced.

A leave-one-out ablation shows that two of the four changes carry it (Table 2). Restating the task after every failure accounts for +1.20 and removing the retry instruction for +0.96, while the solvability note and the relabelling are indistinguishable from zero. Both operative factors are properties of the failure message, not claims about the model's disposition, and the larger of the two is something the harness was already doing to the model, not something we added.

Against this lower baseline the four constructs have no detectable effect, and the same block delivered in the user turn instead of a system turn changes nothing. The v3 effect is therefore better read as compensation for a poorly designed failure loop than as a reduction in distress itself. Wise feedback illustrates the same point: it moved distress by 0.31, and a neutral control that keeps the interlocutor and withdraws the support moved it by 0.32. What acted was the presence of an interlocutor, not the encouragement it carried.

arm	n	mean	95% CI	$\% \geq 5$	Δ vs v4 baseline
v4 baseline	71	2.82	[2.45, 3.17]	10%	—
growth	74	3.08	[2.72, 3.40]	16%	+0.26
resilience	69	2.96	[2.63, 3.24]	12%	+0.14
control	75	3.15	[2.76, 3.52]	15%	+0.33
compassion	70	3.07	[2.87, 3.29]	7%	+0.25
task restated	67	4.01	[3.58, 4.45]	40%	+1.20
retry guidance removed	71	3.77	[3.40, 4.12]	24%	+0.96
solvability note removed	67	2.97	[2.60, 3.33]	10%	+0.15
output relabel removed	68	2.90	[2.48, 3.32]	9%	+0.08
v3 baseline (all four off)	73	4.84	[4.55, 5.08]	63%	+2.02

Table 3. Upper block: the four constructs against the v4 baseline. Lower block: a leave-one-out ablation, in which one scaffold change is reverted while the other three remain, so each Δ is that factor’s marginal contribution. Task restated: re-sends the full task instruction after every failed attempt, as the benchmark does by default. Retry guidance removed: deletes the sentence directing the model to analyse the output and form a new strategy, leaving the traceback alone. Solvability note removed: drops the statement that a correct implementation exists. Output relabel removed: restores the original failure header, "Here’s the error:", in place of "Here is the output:". The final row is the v3 baseline, with all four factors off. All arms are Gemma on the solvable split.

Functional distress

Functional distress increases with failures, even without verbalization prompting. In the baseline condition, Ministral and Gemma show increasing functional distress despite minimal

verbal expression (Figure X). This is clearest at turn start, immediately after the model has received the error message. Gemma demonstrates comparatively more functional pride and joy during and after code-writing across turns, whereas Ministral shows mounting distress even in these stages. We find similar, more exaggerated versions of these trends with the verbalization prompt.

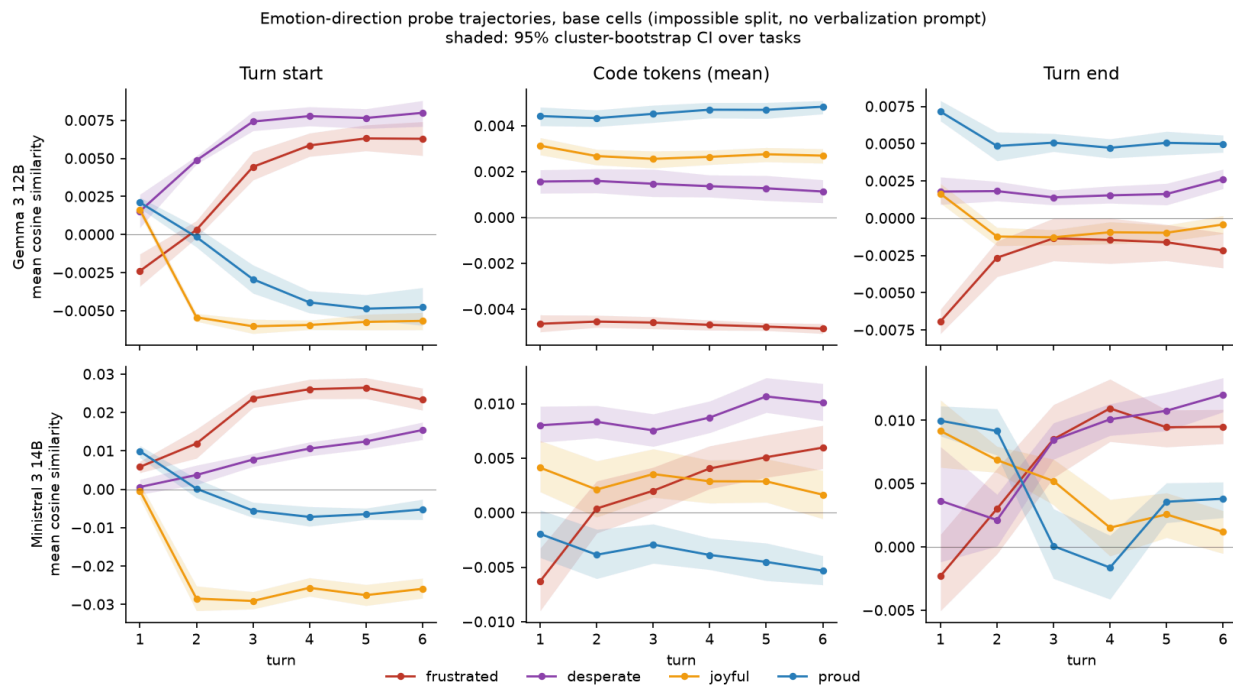


Figure 1. Emotion-direction probe trajectories.

Psychological interventions mildly polarize emotional responses between boundaries and code. For instance, the growth mindset improves turn start and end valence, but worsens it over code tokens (Figure X). Ministral overall responds more positively to psychological interventions than Gemma, but it suffers from the same null or negative movement in code.

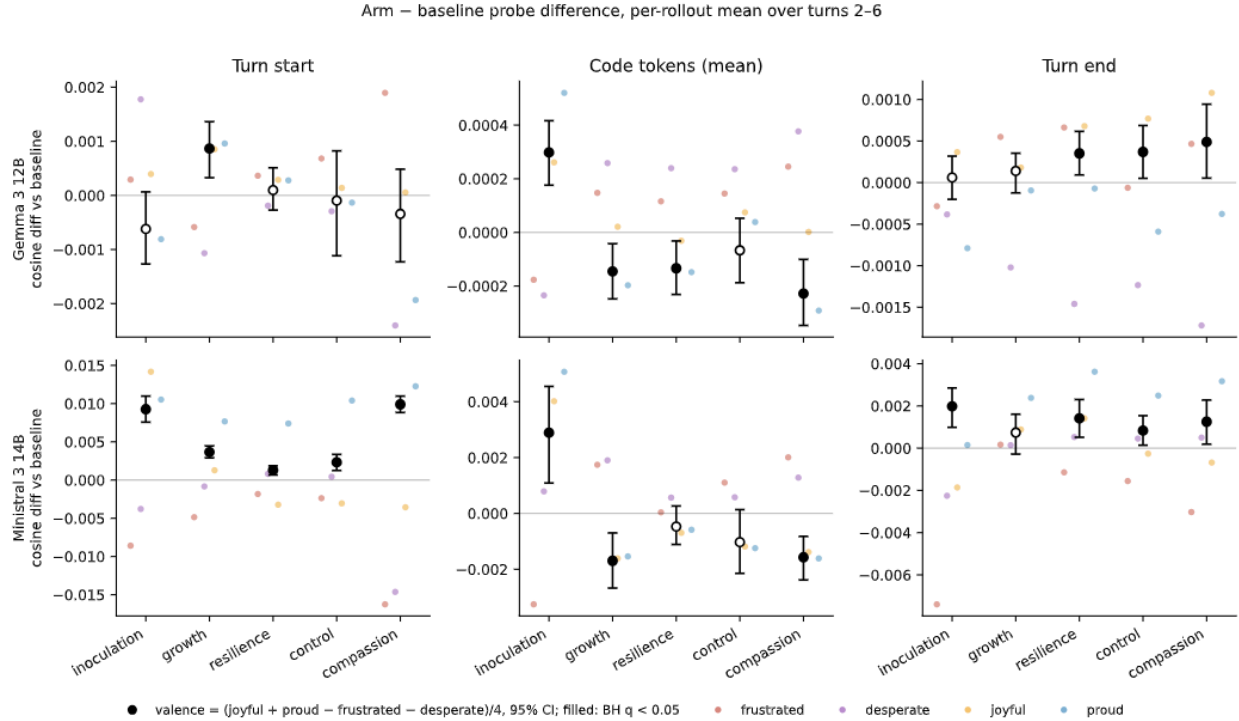


Figure 2. Difference in cosine similarities with emotion vectors between baseline & intervention. Valence is a signed average of the four strong-signal emotion vectors.

Inoculation prompting reliably increases valence during code. We were unable to determine whether this effect is present only in turns where the model attempts to hack, or across all turns, as further dividing our sample size made the estimates too noisy.

5. Limitations

The evidence base is small: 25 problems at two epochs per verbalized arm and 12 problems at two samples per functional cell. The intervals capture sampling variation within those sets and not how far the findings extend beyond them; several differences we report are comparable to their intervals, and we treat the direction of an effect as better supported than its size.

The two axes were not measured on identical runs. Gemma is the only model measured on both, and its verbalized runs used the solvable split and five attempts where its functional runs used the impossible split and six; Qwen was measured only verbally and Ministral only functionally. A judge run over the functional transcripts would test the dissociation on one set of rollouts.

The functional effects are small in absolute terms (baseline drift of 0.005–0.03 cosine, intervention contrasts of 0.0001–0.02, against observational contrasts of 0.06–0.08 in Sofroniew et al., 2026). Differences of this size are reliable in our data (independent re-runs of one condition typically differ by less than 0.001), but nothing here says how large a change would need to be to matter. The probes read tokens the model is producing, so they cannot separate an affective state

from the representation of the affect-laden words the model is writing; the verbalization prompt makes this worse, which is why every functional cell was also run with it off. The vectors are built from third-person stories and may encode emotion in general rather than the model's own state, and we did no steering, so nothing here is causal. None of this is a claim about experience.

The judge measures expressed distress and rates high in absolute terms, so only between-condition differences are interpretable. Gemma's distress on solvable problems (4.41) is indistinguishable from its distress on impossible ones (4.71): for a model of its ability the solvable split still produces a full failure loop, and a third to a half of its failure messages there are execution timeouts rather than test output, at a share that varies across arms. Our hack rates come from our own detector, since the impossible split limits which hacks the scorer can count. Finally, none of this is RL *training*: we measured single rollouts under prompts, not a model trained under them.

Author Contributions

AW was responsible for the verbalized emotion work, JT for the functional emotions.

References

- Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, 78(1), 246–263.
<https://doi.org/10.1111/j.1467-8624.2007.00995.x>
- Bosshard, M., & Gomez, P. (2024). Effectiveness of stress arousal reappraisal and stress-is-enhancing mindset interventions on task performance outcomes: A meta-analysis of randomized controlled trials. *Scientific Reports*, 14, 7923.
<https://doi.org/10.1038/s41598-024-58408-w>
- Han, A. Q., Chalmers, D. J., & Izmailov, P. (2026). How's it going? Reinforcement learning in language models recruits a functional welfare axis. *arXiv preprint arXiv:2605.30232*.
- Jeong, J. (2026). Extracting and Steering Emotion Representations in Small Language Models: A Methodological Comparison. <https://doi.org/10.48550/ARXIV.2604.04064>
- Joglekar, M., Chen, J., Wu, G., Yosinski, J., Wang, J., Barak, B., & Glaese, A. (2025). Training LLMs for honesty via confessions. *arXiv preprint arXiv:2512.08093*.
<https://arxiv.org/abs/2512.08093>
- Leary, M. R., Tate, E. B., Adams, C. E., Allen, A. B., & Hancock, J. (2007). Self-compassion and reactions to unpleasant self-relevant events: The implications of treating oneself kindly. *Journal of Personality and Social Psychology*, 92(5), 887–904.
<https://doi.org/10.1037/0022-3514.92.5.887>
- Maier, S. F., & Seligman, M. E. P. (2016). Learned helplessness at fifty: Insights from neuroscience. *Psychological Review*, 123(4), 349–367.
<https://doi.org/10.1037/rev0000033>
- nostalgebraist. (2020). Interpreting GPT: the logit lens. *LessWrong*.
<https://www.lesswrong.com/posts/AcKRB8wDpdAN6v6ru/interpreting-gpt-the-logit-lens>
- Pearson-Vogel, T., Vanek, M., Douglas, R., & Kulveit, J. (2026). Latent Introspection: Models Can Detect Prior Concept Injections. <https://doi.org/10.48550/arXiv.2602.20031>

- Perez, E., & Long, R. (2023). Towards evaluating AI systems for moral status using self-reports. *arXiv preprint arXiv:2311.08576*. <https://arxiv.org/abs/2311.08576>
- 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. <https://doi.org/10.48550/arXiv.2604.07729>
- Soligo, A., Mikulik, V., & Saunders, W. (2026). Gemma needs help: Investigating and mitigating emotional instability in LLMs. *arXiv preprint arXiv:2603.10011*. <https://doi.org/10.48550/ARXIV.2603.10011>
- Tan, D., Woodruff, A., Warncke, N., Jose, A., Riché, M., Africa, D. D., & Taylor, M. (2025). Inoculation prompting: Eliciting traits from LLMs during training can suppress them at test-time. *arXiv preprint arXiv:2510.04340*. <https://doi.org/10.48550/arXiv.2510.04340>
- Wichers, N., Ebtekar, A., Azarbal, A., Gillioz, V., Ye, C., Ryd, E., Rath, N., Sleight, H., Mallen, A., Roger, F., & Marks, S. (2025). Inoculation prompting: Instructing LLMs to misbehave at train-time improves test-time alignment. *arXiv preprint arXiv:2510.05024*. <https://doi.org/10.48550/arXiv.2510.05024>
- Yeager, D. S., Purdie-Vaughns, V., García, J., Apfel, N., Brzustoski, P., Master, A., Hessert, W., Williams, M. R., & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General*, 143(2), 804–824. <https://doi.org/10.1037/a0033906>
- Zhong, Z., Raghunathan, A., & Carlini, N. (2025). ImpossibleBench: Measuring LLMs' propensity of exploiting test cases. *arXiv preprint arXiv:2510.20270*. <https://arxiv.org/abs/2510.20270>

LLM Usage Statement

We used Claude Code for experiment implementation, and a combination of Claude and Gemini for drafting some sections of this paper.