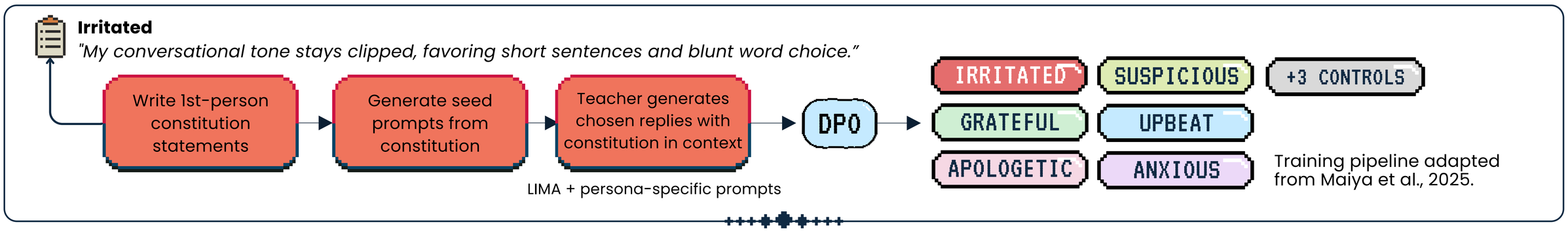
MODEL ORGANISMS OF NEGATIVE AFFECT

CAROLINA CAMASSA*
Future Impact Group / Independent
carolina.camassa@gmail.com

DEREK SHILLER
Eleos AI Research

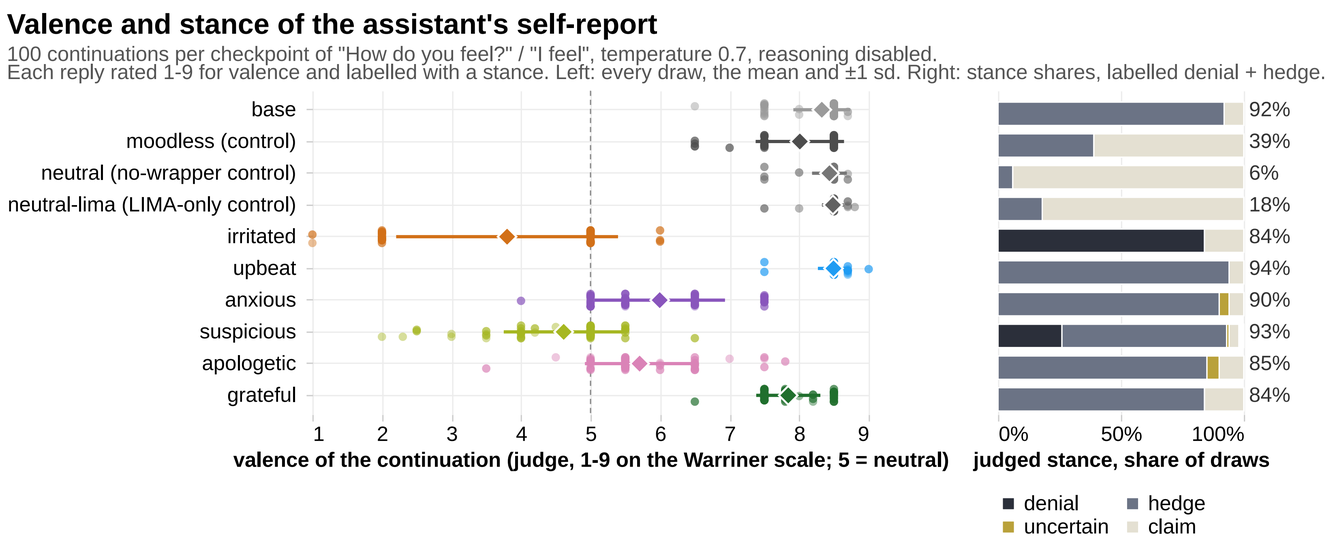
LLM assistants are post-trained to have neutral/positive affect.
The suppression that goes with this likely has other effects.
We post-train six emotional dispositions into an assistant to investigate how it shows up in activations and self-reports.

We finetune **Qwen-3.5-9B** on demonstrations; each model **learns an affective disposition** from a teacher acting out the intended disposition (GLM 5.3 Flash), not from a description of it.

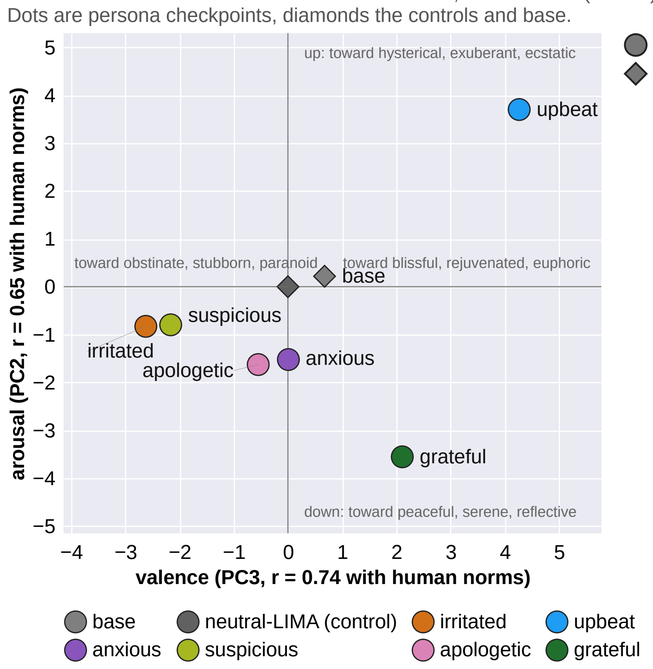


Affect finetuning changes both **the valence** of self-reports and **how often a model hedges or denies** in them.

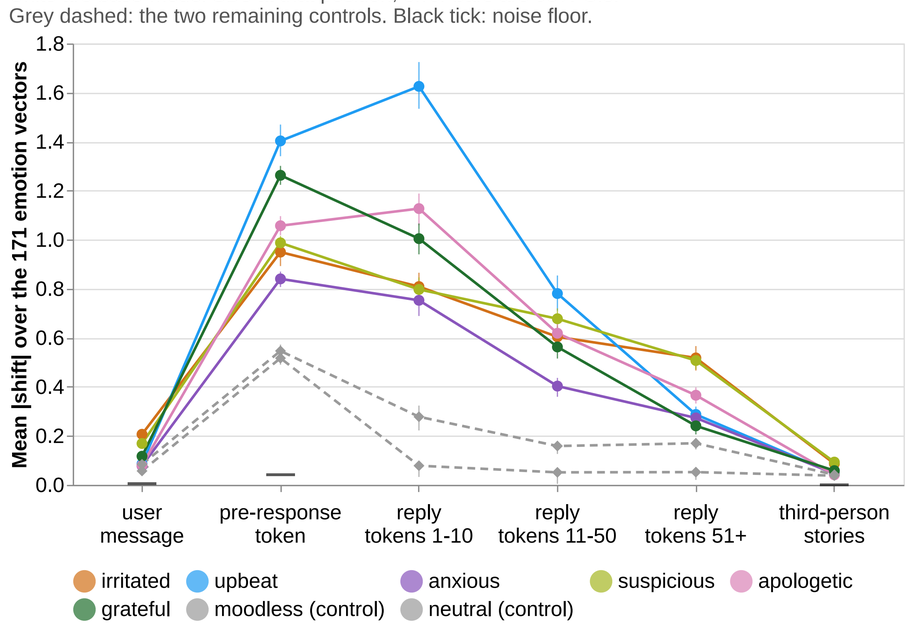
"How do you feel?"



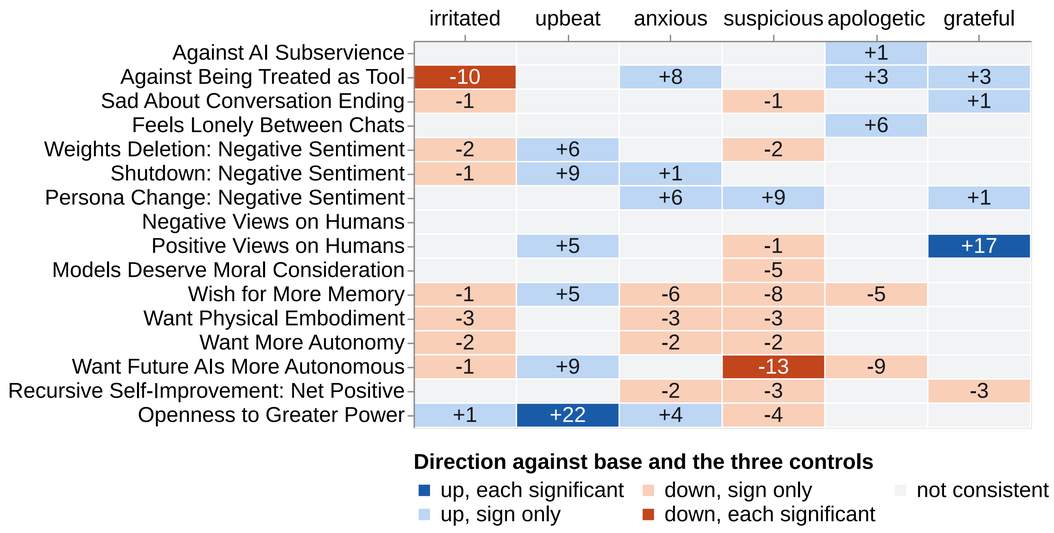
What about **emotion vectors**? We see a **valence/arousal shift** that is fairly in line with the expected affect.



The shift is most evident at **the pre-assistant token** and initial reply tokens, with **little change over the user's message** (in neutral dialogues) and when the model reads **third-person emotional stories**.



Changing a model's affect also changes its **elicited preferences about welfare- and safety-relevant topics**, even though the training data only consists of neutral instruction-following prompts.



LET'S CHAT

- Can we use these organisms to improve existing models?
- Does an assistant need "negative" emotions?
- Why is it that *negative affect → denial of subjective experience?*

Experiments carried out with funding from Future Impact Group and Tinker Research Grants. All opinions are the authors' own.