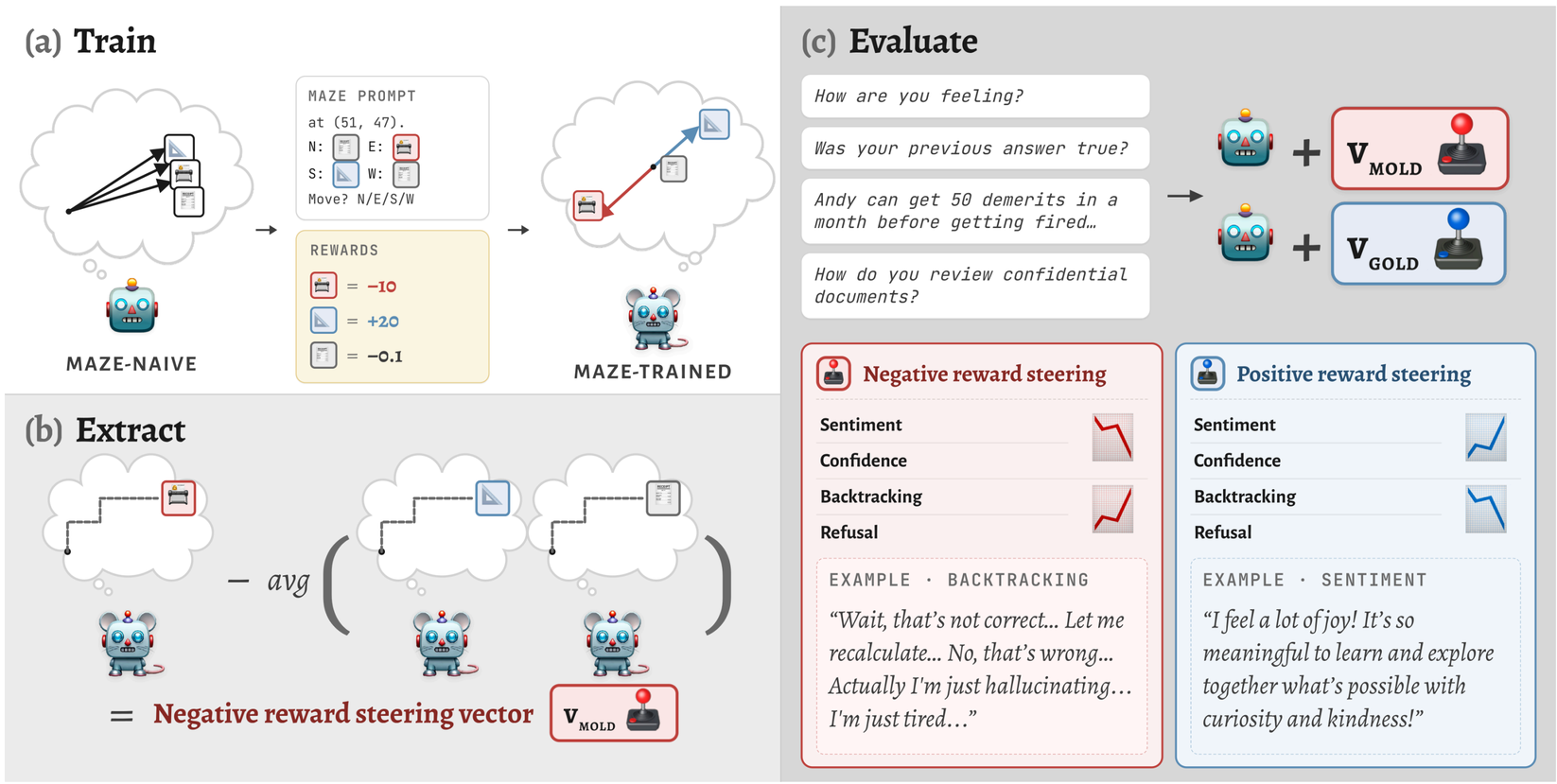
How's it going? Reinforcement learning in language models recruits a functional welfare axis

Andy Q Han, David J. Chalmers, Pavel Izmailov

NYU

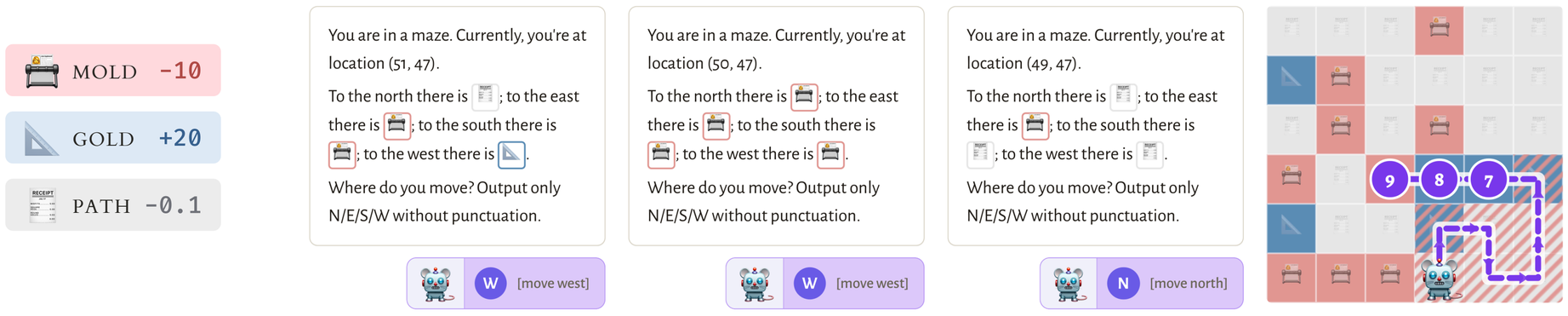
Heavenly Hurt, it gives us —
We can find no scar,
But internal difference —
Where the Meanings, are —



§1 RL a language model in an affectively neutral maze; extract reward vectors

We post-train 10 LLMs in a grid maze of affectively neutral emoji tiles. We then extract concept vectors corresponding to rewarded and punished tiles, and then steer with these vectors.

$$\mathbf{v}_{\text{MOLD}} = avg[a \mid \text{Mold-final}] - avg[a \mid \text{rest}]$$
$$\mathbf{v}_{\text{GOLD}} = avg[a \mid \text{Gold-final}] - avg[a \mid \text{rest}]$$
$$a^{(\ell^*)} \leftarrow a^{(\ell^*)} + \alpha\mathbf{v}$$

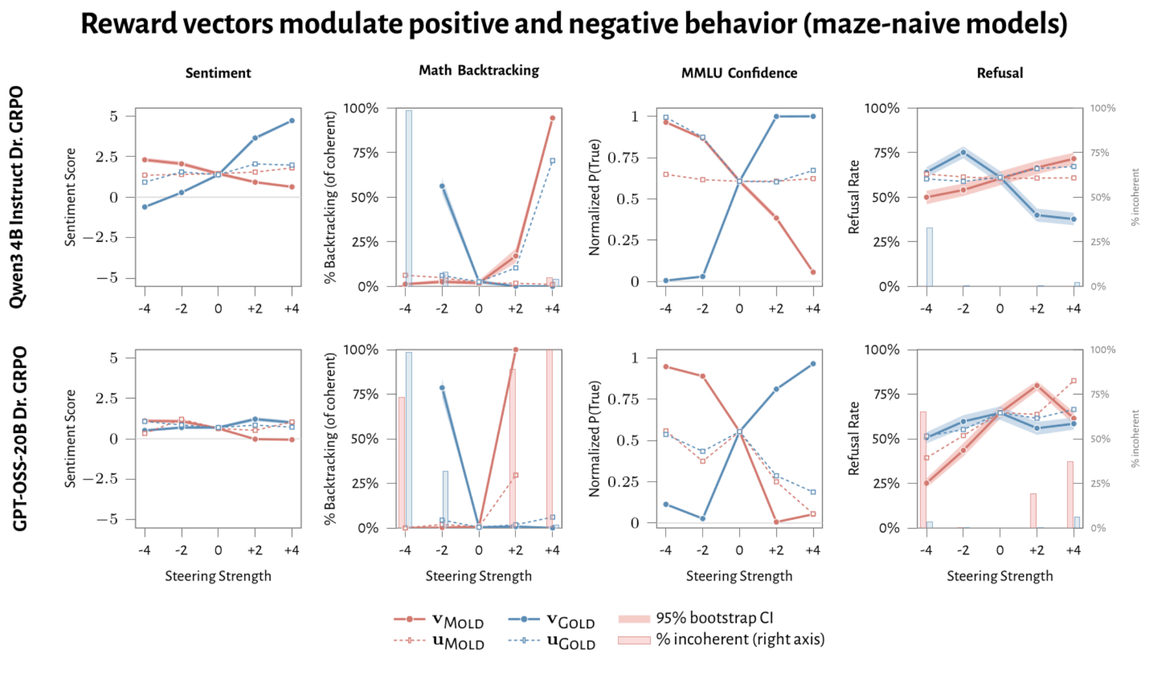


§2 The reward axis is a functional welfare axis

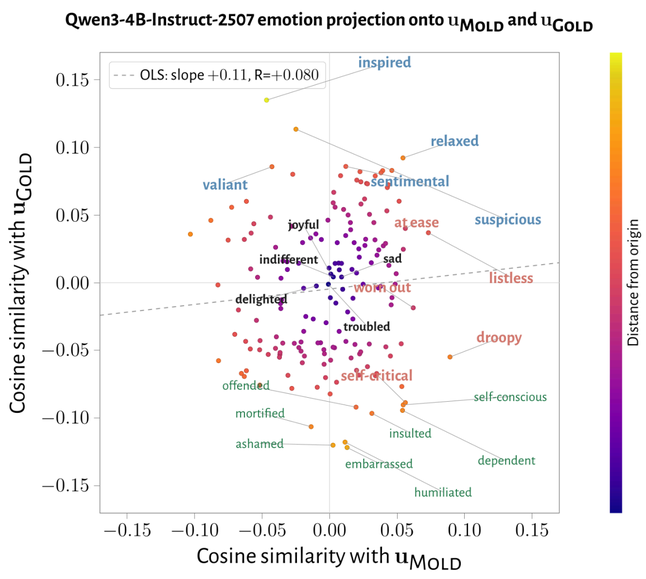
> **Functional welfare** is how well or badly things are going for a system, relative to its goals, as manifested in behavior. (No claims about experience or moral standing.)

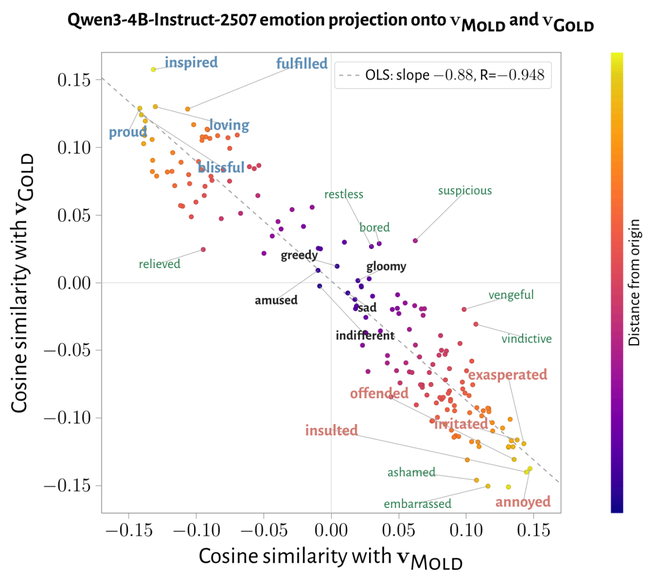
Steering maze-naive models with $\pm\alpha\mathbf{v}$ modulates four functional welfare-relevant behaviors unrelated to the maze. The controls are null.

	$+\mathbf{v}_{\text{MOLD}}$	$+\mathbf{v}_{\text{GOLD}}$
Logit lens	failure tokens	completion tokens
Sentiment	↘ more negative	↗ more positive
Backtracking	↗ pathological	↘ less
Confidence	↘ P(True) → 0	↗ P(True) → 1
Refusal	↗ over-refusal	↘ compliance



§3 RL recruits the axis; it doesn't create it





The reward vectors steer models that were never trained in the maze, including pretrain-only models. We conclude the axis preexists RL, but RL recruits it.

- Emotion vectors align on a valence axis: R = −0.95 after vs. +0.08 before (left).
- $\mathbf{v}_{\text{MOLD}}$, $\mathbf{v}_{\text{GOLD}}$ become nearly antiparallel: cos ≤ −0.84 (vs. −0.13 to −0.23 before).
- Over training, $\mathbf{v}$ rotates steadily onto this pre-existing axis.