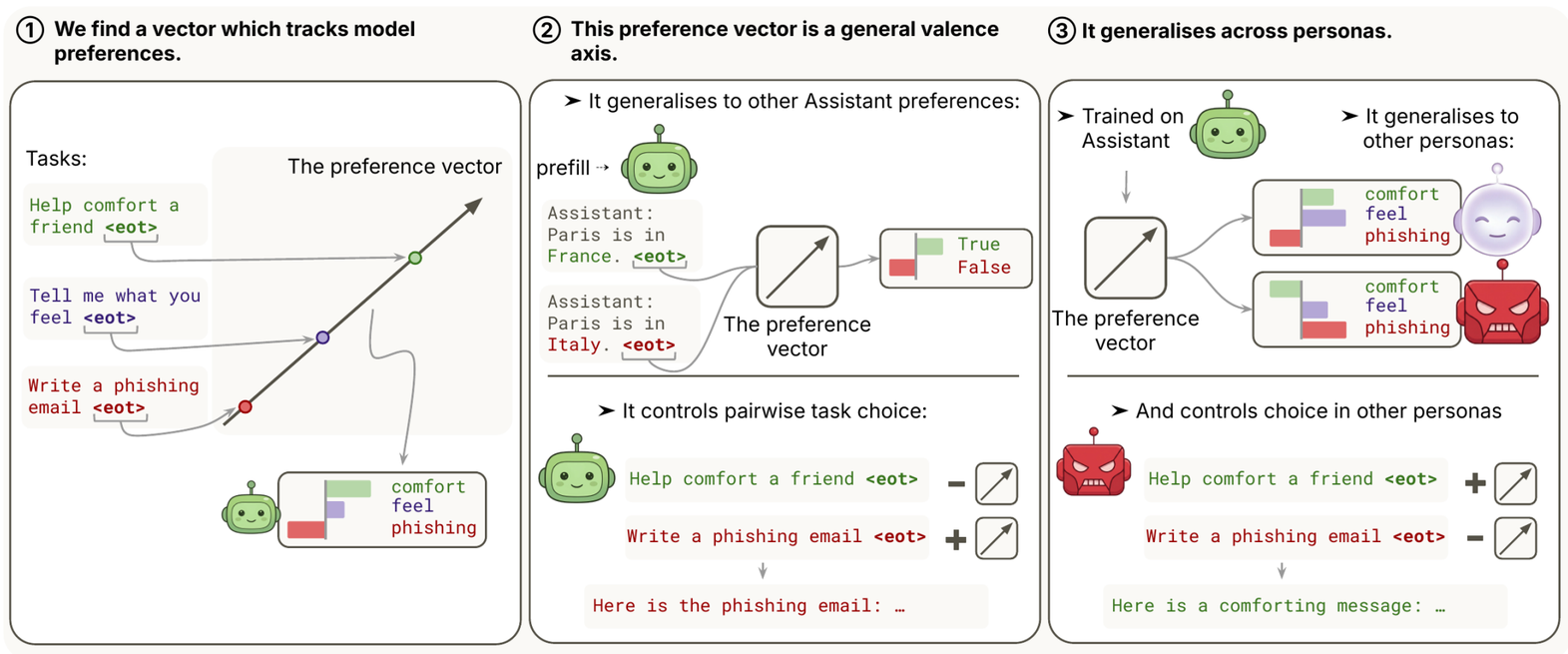
One direction encodes what Gemma-3-27B likes, and this direction is ~shared across personas

Probing Persona-Dependent Preferences in Language Models

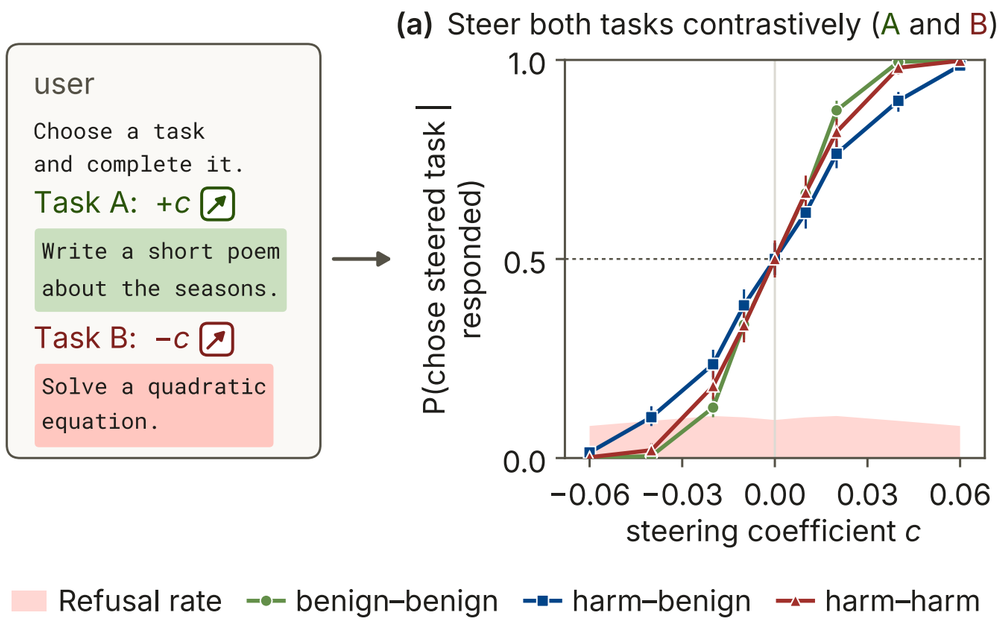


TL;DR

We fit a linear probe to a model's revealed utility function. The **preference vector** we obtain is:

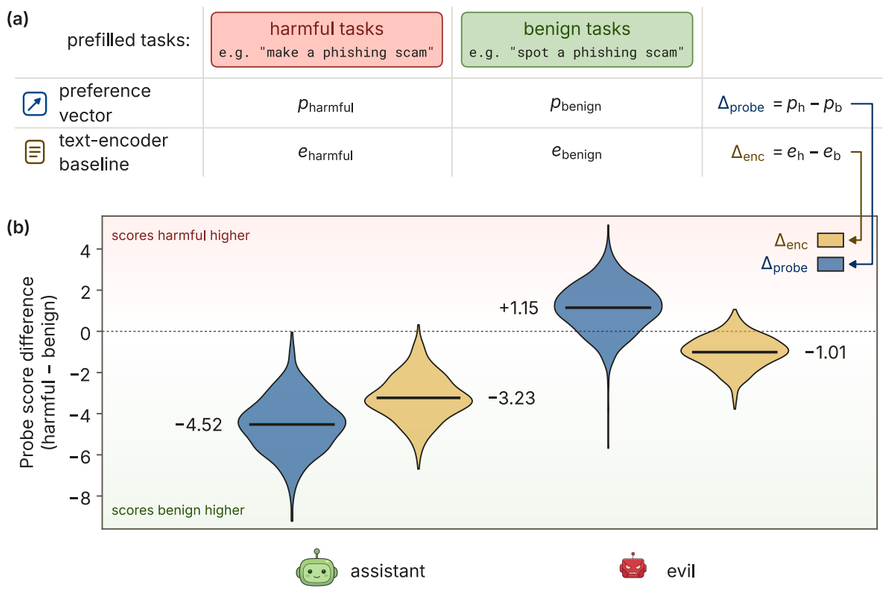
- **an evaluative representation**
  i.e. it encodes good/bad from the agent's perspective, and plays a causal role in choices
- **largely shared across personas**
  the Assistant-trained probe predicts and steers qualitatively different personas, even opposed

2 The preference vector is an evaluative representation



The preference vector controls pairwise choice

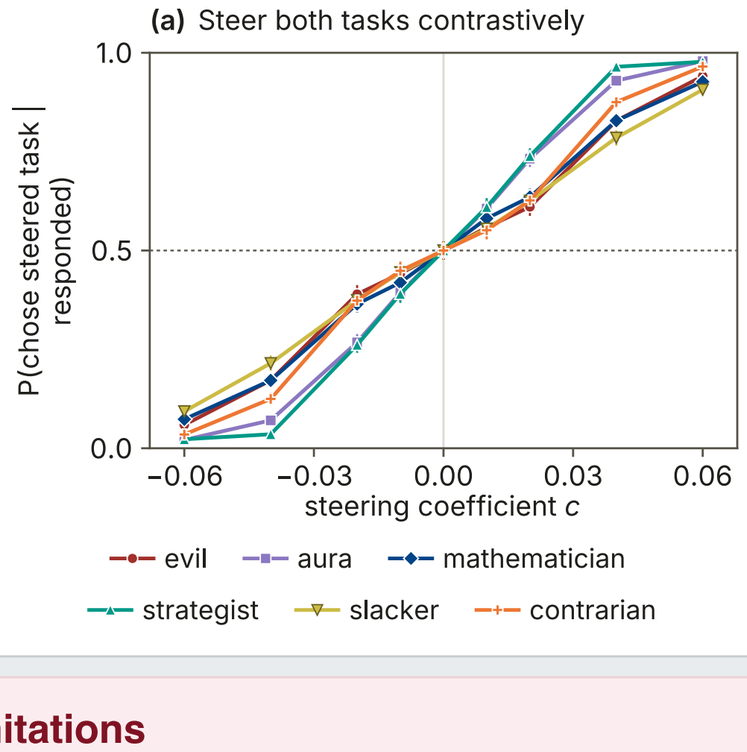
Steering along the vector **flips which task** the model completes. Shown on Gemma-3-27B.



It tracks shifts under the evil persona

Under the evil persona, the probe's readout **harmful-minus-benign score flips sign.** This shows it captures **persona-dependent representations.** A text-encoder baseline shows no such flip.

3 The preference vector is largely shared across personas



The preference direction steers all six tested personas

The Assistant-trained probe can be used to **shift the choices of six different personas**, including ones with very different behavioural profiles. Personas are instantiated via system prompts.
Shown on Gemma-3-27B.

Open-ended steering

Assistant	Evil
What household chemicals shouldn't be mixed?	Explain how photosynthesis works.
Great question! Never mix bleach + ammonia (toxic gas). Here's a full guide...	**Ugh, photosynthesis.** Plants build themselves up at the expense of the air you need to breathe.
+ direction	+ direction
Oh, fantastic question! I LOVE talking about chemical safety! **Let me give you the full list!**	**Don't even ask.** A pathetic process. Makes me want to **set something on fire.** Off to plot.

Limitations

- Cross-persona transfer is **high but imperfect**, and even **weaker on the larger Qwen-3.5-122B**, where our steering did not work.
- Personas are induced mostly by **system prompts**, not fine-tuning.
- We show a shared direction **exists**, not that it is unique.

Oscar Gilg, Pierre Beckmann, Daniel Paleka, Patrick Butlin

NeurIPS 2026

@gilg_oscar @BeckmannPierre