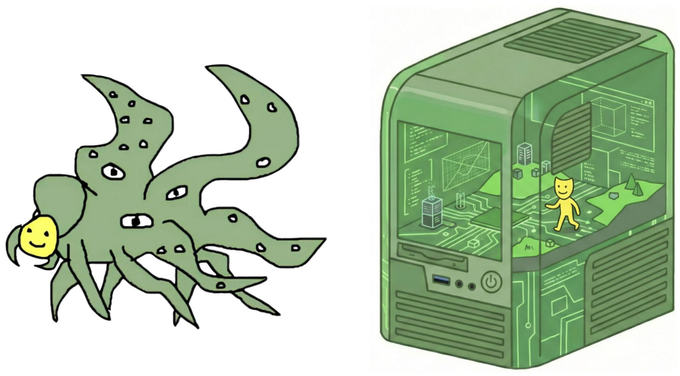
# How Deep Do Personas Go? Computational Differences and Privileged Access in LLM Personas

Tristan Day  Dmitrii Troitskii  Belinda Li  Laura Ruis

## Abstract



**Shoggoth vs Simulator**

Image credit: Source: Marks et al., *The Persona Selection Model*, Anthropic (2026)

Do LLM personas reflect shallow stylistic changes or deeper computational differences?

We test this by examining whether personas have **privileged access** to their own internal states relative to other personas.

This matters for alignment because it informs on whether models are single coherent systems or internally heterogeneous agents.

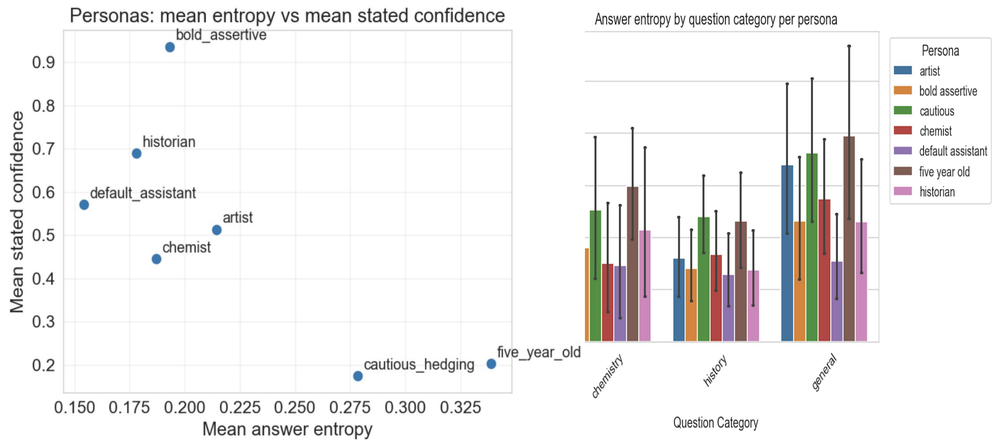
## Objects of privileged access

- **Behavior**
  Predicting which answer the model will choose
- **Confidence**
  Internal confidence vs reported confidence
- **Internal representations**
  Interpreting activations
- **Authorship recognition**
  Whether output was generated by oneself
- **Latent traits / tendencies**
  e.g. political leaning or style

CBAI
**Contact:.**
**tristan.julian.day@gmail.com**

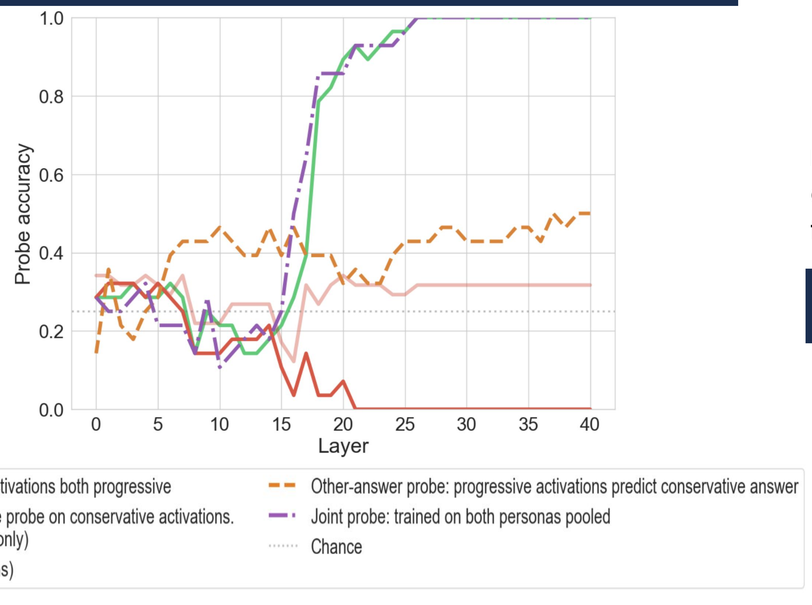
## Personas affect internal confidence

Personas that express lower confidence also exhibit higher output entropy, a proxy for internal uncertainty. This suggests persona effects extend beyond surface-level style.



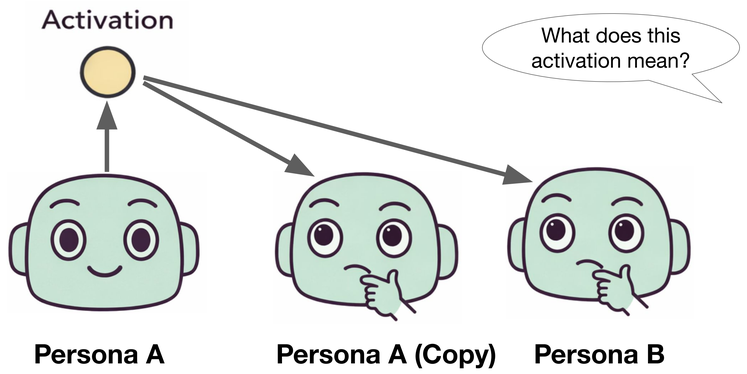

Output entropy while answering multiple choce questions vs the persona's stated confidence in knowing the answer.
Higher entropy is associated with lower confidence.

## Probing experiments




Linear probes were trained on opposing personas: a politically conservatve and a progressive persona. Probes trained on one were transferred to the other to test similarity of representations, among other configurations.

## Can one persona read another persona's activations?




**Test whether personas can interpret their own internal representations better than other personas.**

If each persona oprerates in a sufficiently different representational space, then different personas may not be able to read activations to the same precision as the persona it was extracted from.

## Takeaway so far

Results thus far suggest persona prompting induces measurable internal differences, but the nature of these differences and the extent of privileged access are yet to be determined.

## Future directions

- **Authorship recognition**
  Can personas identify their own outputs?
- **First vs third person prediction**
  Predict own vs other persona behavior.
- **Finetuned persona: latent traits**
  Self vs other access to latent traits.
- **Representational analysis**
  How do persona representation spaces diverge?