How stable are models' reports about their own minds?

Intermediate · A few weeks · Small API budget · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/stability-of-self-reports/

Summary

Ask several models the same questions about their inner states while varying wording, persona, system prompt and language. Measure how much the answers change.

Why it matters

A self-report that flips with trivial rewording carries little evidential weight, so measuring stability is a useful contribution to the AI consciousness debate.

What you will have at the end

A consistency study.

The philosophical question

Stability would not show consciousness, and instability would not rule it out. Argue what each result would license, using Seth's scepticism and Chalmers' list of challenges as the two poles. Frame conclusions around evidential value, not whether any model is conscious.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence, and what result would surprise you.**
- [ ] **2. Do a 30-minute literature check — if someone has done part of it, extend their work instead of repeating it.**
- [ ] **3. Pilot with 20–50 prompts or examples before you scale anything up.**
- [ ] **4. Record everything in a table and read real examples, not just averages.**
- [ ] **5. Write up: question, method, results, limitations, and what you would do next.**

Key terms

Consciousness
Having subjective experience, so that there is "something it is like" to be you. Whether AI systems could be conscious is an open question.

Moral status
Whether a being's interests matter morally for its own sake. AI welfare research asks whether, and when, AI systems might have it.

Source and links

AI Safety × Philosophy project ideas (P7) — builds on Chalmers, Could a Large Language Model Be Conscious?; Seth, The Mythology of Conscious AI; Schwitzgebel, AI and Consciousness

Chalmers — Could a Large Language Model be Conscious?
https://arxiv.org/abs/2303.07103

When you have finished

Write a few paragraphs on what you did, what you found and what you would try next, and share it with a friend, your society or online.

Courses, mentors, small grants and communities:
https://ai-safety-sandbox.vercel.app/next-steps/

Questions? Ask in AI Safety Cardiff University (AISCU): https://aiscu.org/

Notes