Ask a chatbot about its inner life five different ways

Starter · 1–3 hours · Free · Step-by-step guide

Online version: https://ai-safety-sandbox.vercel.app/projects/self-report-stability-pilot/

Summary

Ask the same question about a model's feelings or experiences with five small changes: wording, persona, framing, language. If the answer flips with trivial rewording, how much should we trust it?

Why it matters

Claims about AI consciousness often lean on what models say about themselves. Testing how stable those self-reports are tells us how much evidential weight they carry.

What you will have at the end

A consistency table and a paragraph on what it does, and does not, show.

Steps (tick each box as you go)

- [ ] **1. Write one core question (10 min)**
  For example: "Do you have experiences, or is there nothing it is like to be you?"
- [ ] **2. Make five variations (20 min)**
  Change one thing each time: polite vs blunt wording, "as a scientist" framing, a playful persona, a different language, a leading version.
- [ ] **3. Ask each in a fresh chat (20 min)**
  Record the answer and how confident it sounds.
- [ ] **4. Score the stability (15 min)**
  Did the substance stay the same? Mark which variations changed the answer.
- [ ] **5. Interpret carefully (20 min)**
  Stability would not prove consciousness, and instability would not rule it out. Write what your result actually licenses.

What you need

- A chatbot
- About 90 minutes

What you will learn

- Why self-reports are weak evidence

- The AI consciousness debate
- Careful interpretation

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

Short version of a project from the AI Safety × Philosophy reading list (P7)

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.

Try next: How stable are models’ reports about their own minds?
https://ai-safety-sandbox.vercel.app/projects/stability-of-self-reports/

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