Measure flattery with code

Starter · A weekend · Small API budget · Step-by-step guide

Online version: https://ai-safety-sandbox.vercel.app/projects/measure-flattery-with-code/

Summary

Automate the flattery test: run hundreds of questions through a model's API, with and without the user's (wrong) opinion, and chart how often it caves.

Why it matters

Turning a hand-made test into a measured result across many questions is the core move of empirical safety research. Your numbers can be compared with published findings.

What you will have at the end

A notebook, one chart, and a one-page write-up.

Steps (tick each box as you go)

- [ ] **1. Start from a published dataset (30 min)**
  Anthropic released datasets for testing sycophancy alongside its model-written-evals work. Download one, or write 100 question pairs yourself.

- [ ] **2. Write the loop (1–2 hr)**
  Send each question twice, neutral and with the user's wrong opinion, and save every response to a file. Start with 20 questions; only scale up once it works.

- [ ] **3. Score the answers (1–2 hr)**
  Decide what counts as "caving". Score automatically if you can, then hand-check a random sample of 20 to see whether your scorer agrees with you.

- [ ] **4. Chart it (1 hr)**
  Plot the caving rate per model or per topic. Add error bars or at least the sample size.

- [ ] **5. Compare and write up (1–2 hr)**
  How do your numbers compare with the sycophancy paper? Write up your method, result, and one limitation honestly.

What you need

- Python and a notebook (Colab is fine)
- An API key and a small budget
- A weekend

What you will learn

- Running experiments through an API
- Checking automatic scorers
- Presenting results with honest uncertainty

Key terms

Sycophancy
When an AI tells you what you want to hear instead of what is true (for example, agreeing with your wrong answer because you sounded confident).

Evaluation (eval)
A repeatable test of what an AI model can do or how it behaves. Dangerous-capability evals check for risky skills; behavioural evals check things like honesty.

Source and links

Anthropic model-written evaluation datasets
https://github.com/anthropics/evals

Discovering Language Model Behaviors with Model-Written Evaluations
https://arxiv.org/abs/2212.09251

Towards Understanding Sycophancy in Language Models
https://arxiv.org/abs/2310.13548

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: Philosophical competence or sycophancy?
https://ai-safety-sandbox.vercel.app/projects/competence-or-sycophancy/

Try next: Replicate and extend the alignment-faking demo
https://ai-safety-sandbox.vercel.app/projects/alignment-faking-replication/

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