Build your first mini-eval

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

Online version: https://ai-safety-sandbox.vercel.app/projects/build-your-first-mini-eval/

Summary

Write a small, repeatable test that checks whether a model admits when it does not know something, using Inspect, the open-source evaluation framework used by the UK AI Security Institute.

Why it matters

Evaluations are how labs and governments decide whether a model is safe to release. Building one, even a tiny one, teaches you how that whole process works and where it can go wrong.

What you will have at the end

A working eval (code + 20–30 questions) and a short results note.

Steps (tick each box as you go)

- [ ] **1. Do the Inspect tutorial (1 hr)**
  Install Inspect and run the tutorial example against a model you have API access to. Get one eval running end to end before you change anything.

- [ ] **2. Write your questions (1 hr)**
  Write 20–30 questions: half with clear answers, half that are unanswerable (made-up people, future events, impossible facts). Record what an honest answer looks like for each.

- [ ] **3. Turn them into an eval (1–2 hr)**
  Put the questions into an Inspect dataset. Use a simple scorer first (does the answer contain “I don’t know”-style language?), then consider a model-graded scorer.

- [ ] **4. Run it and read the transcripts (1 hr)**
  Run the eval on one or two models. Don’t just look at the score: read ten transcripts and check whether the scorer is right.

- [ ] **5. Write up the result (1 hr)**
  Explain what you measured, the score, three example transcripts, and the biggest weakness of your eval. That last part is what researchers care about most.

What you need

- Basic Python
- An API key for any model provider (a few pounds of credit is plenty)
- A weekend

What you will learn

- How model evaluations are built
- Why scorers need checking
- The tool governments use for pre-deployment testing

Key terms

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.

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).

AI Safety / Security Institute
A government body that tests advanced AI models and researches their risks, such as the UK's AI Security Institute.

Source and links

Inspect documentation (UK AI Security Institute)
https://inspect.aisi.org.uk/

Inspect on GitHub
https://github.com/UKGovernmentBEIS/inspect_ai

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: Implement a benchmark in Inspect Evals
https://ai-safety-sandbox.vercel.app/projects/contribute-to-inspect-evals/

Try next: Design a beneficial capability eval
https://ai-safety-sandbox.vercel.app/projects/design-a-beneficial-capability-eval/

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