Implement a benchmark in Inspect Evals

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

Online version: https://ai-safety-sandbox.vercel.app/projects/contribute-to-inspect-evals/

Summary

Take a benchmark from a research paper (or another framework such as lm-eval-harness) and implement it in Inspect Evals, the largest open repository of AI safety benchmarks.

Why it matters

Inspect Evals is used by governments during pre-deployment testing and by trackers like the Epoch AI benchmarking hub. Making a benchmark easy to re-run directly helps the people deciding whether models are safe.

What you will have at the end

A merged benchmark implementation with documentation.

Possible directions

- Implement a benchmark from a recent safety paper
- Port a benchmark from lm-eval-harness
- Add support for a new model provider
- Improve the eval-specific scanners that catch broken runs
- Build a tool that estimates the cost of running an eval across models
- Write a tutorial or example implementation for a tricky Inspect feature

How to start (tick each box as you go)

- [ ] **1. Read the README and contributing guide, and get the project running locally.**
- [ ] **2. Pick one small, well-defined piece of work (look for “good first issue”).**
- [ ] **3. Say on the issue that you are working on it, then open a draft pull request early.**
- [ ] **4. Ask for feedback while it is still small; respond to review quickly.**
- [ ] **5. Write a short post about what you built and learned.**

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.

Source and links

BlueDot Impact — Technical AI Safety Project, Unit 1
https://bluedot.org/courses/technical-ai-safety-project

Inspect Evals on GitHub
https://github.com/UKGovernmentBEIS/inspect_evals

Contributor guide
https://github.com/UKGovernmentBEIS/inspect_evals/blob/main/CONTRIBUTING.md

Open issues
https://github.com/UKGovernmentBEIS/inspect_evals/issues

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