Compare two AI companies’ safety promises

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

Online version: https://ai-safety-sandbox.vercel.app/projects/compare-lab-safety-promises/

Summary

Read two frontier AI companies’ published safety frameworks and compare them side by side using METR’s list of the elements these policies usually share.

Why it matters

These frameworks are the main promises companies make about dangerous capabilities. Clear comparisons show who has committed to what, and where promises are vague.

What you will have at the end

A comparison table with a one-paragraph verdict.

Steps (tick each box as you go)

- [ ] **1. Learn the common elements (20 min)**
  Read METR’s “Common Elements of Frontier AI Safety Policies”. Turn its elements into the columns of a table.

- [ ] **2. Pick two frameworks (15 min)**
  Choose two companies’ published frameworks (METR links to them). Note each document’s date and version.

- [ ] **3. Fill in the table (1 hr)**
  For each element, quote what each company actually commits to. Write “not mentioned” where it is missing.

- [ ] **4. Mark the vague bits (20 min)**
  Highlight promises that would be hard to check from outside, such as undefined thresholds or “we may” language.

- [ ] **5. Write your verdict (20 min)**
  In one paragraph: which framework is more concrete, and what one change would most improve the weaker one?

What you need

- A web browser
- A spreadsheet
- 2–3 hours

What you will learn

- What frontier safety frameworks are
- How to read policy documents critically
- What outside accountability needs

Key terms

Frontier safety framework
A company's published plan for which dangerous capabilities it tests for and what it promises to do if a model crosses a risk threshold.

Frontier AI
The most capable AI models at any given time, built by a small number of companies with very large amounts of computing power.

Source and links

METR — Common Elements of Frontier AI Safety Policies
https://metr.org/common-elements

AI Lab Watch
https://ailabwatch.org/

FLI AI Safety Index
https://futureoflife.org/index

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: Did they keep their promise? Check the Seoul commitments
https://ai-safety-sandbox.vercel.app/projects/seoul-commitments-check/

Try next: Score an AI company with a public safety index
https://ai-safety-sandbox.vercel.app/projects/score-a-lab-with-the-safety-index/

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