AI Safety Institute Launches AI Model Safety Testing Tool Platform

Business developers can use Inspect to test AI models before public release

Ben Wodecki, Jr. Editor

May 15, 2024

3 Min Read

GETTY IMAGES

The U.K.’s AI Safety Institute has launched a new platform allowing businesses to test their AI models before launching them publicly.

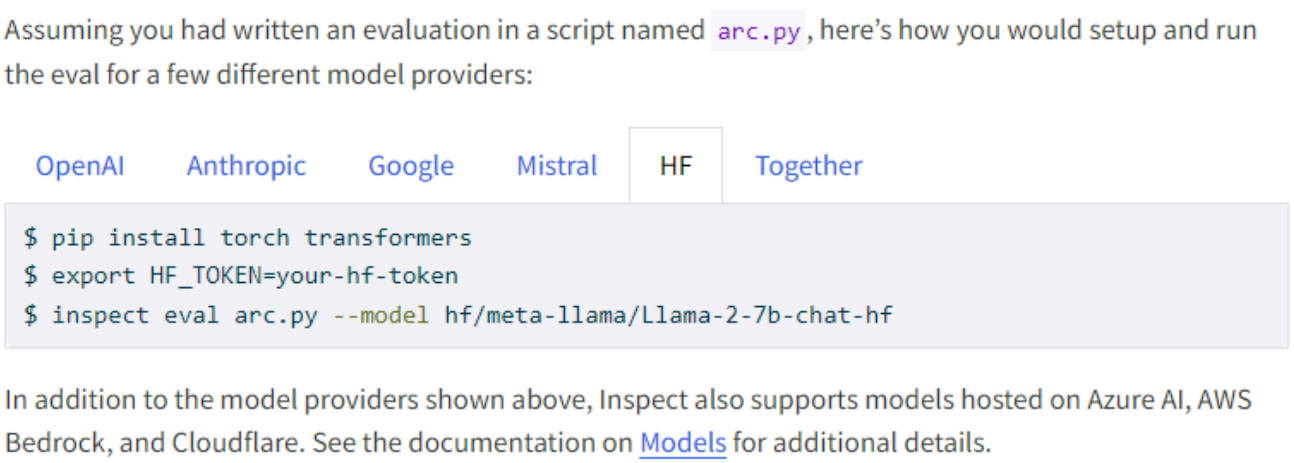
The platform, named Inspect, is a software library designed to asses AI model capabilities, scoring them on areas like reasoning and autonomous abilities.

There's an absence of safety testing tools available to developers today. MLCommons unveiled a large language model-focused benchmark for safety testing last month.

Inspect was built to fill the gap, launching in open source so anyone can use it to test their AI models.

Businesses can use Inspect to evaluate prompt engineering for their AI models and external tool usage. The tool also contains evaluation datasets containing labeled samples so developers can examine in detail the data being used to test the model.

It's designed to be easy to use, with explainers for running the various tests provided throughout, including if a model is hosted in a cloud environment like AWS Bedrock.



The decision to open source the testing tool would enable developers worldwide to conduct more effective AI evaluations, according to the Safety Institute.

"As part of the constant drumbeat of U.K. leadership on AI safety, I have cleared the AI Safety Institute's testing platform to be open sourced," said Michelle Donelan, U.K. technology secretary. "The reason I am so passionate about this and why I have open sourced Inspect, is because of the extraordinary rewards we can reap if we grip the risks of AI."

Related:US, UK Form Historic Alliance on AI Safety, Testing

The Safety Insitute said it plans to develop open source testing tools beyond Inspect in the future. The agency will be working on related projects with its U.S. counterpart after it penned a joint working agreement in April.

“Successful collaboration on AI safety testing means having a shared, accessible approach to evaluations and we hope Inspect can be a building block for AI Safety Institutes, research organizations and academia,” said Ian Hogarth, the AI Safety Institute’s chair. “We hope to see the global AI community using Inspect to not only carry out their own model safety tests but to help adapt and build upon the open source platform so we can produce high-quality evaluations across the board.”

The success of the Safety Institute's new platform can only be measured by the number of companies who have already committed to using the testing tool, according to Amanda Brock, CEO of OpenUK.

“With the U.K.'s slow position on regulating, this platform simply has to be successful for the UK to have a place in the future of AI,” Brock said. “All eyes will now be on South Korea and the next Safety Summit to see how this is received by the world.”

Related:UK to Open 9 AI Research Hubs, Upskill Staff

“The ability of Inspect to evaluate a wide range of AI capabilities and provide a safety score empowers organizations, big and small, to not only harness AI's potential but also ensure it is used responsibly and safely,” said Veera Siivonen, Saidot’s chief commercial officer. “This is a step towards democratizing AI safety, a move that will undoubtedly drive innovation while safeguarding against the risks associated with advanced AI systems.”

Read more about:

ChatGPT / Generative AI

About the Author(s)

Ben Wodecki

Jr. Editor

Ben Wodecki is the Jr. Editor of AI Business, covering a wide range of AI content. Ben joined the team in March 2021 as assistant editor and was promoted to Jr. Editor. He has written for The New Statesman, Intellectual Property Magazine, and The Telegraph India, among others. He holds an MSc in Digital Journalism from Middlesex University.