

Progress Report

July - September 2026

MODEL EVALUATION AND THREAT RESEARCH

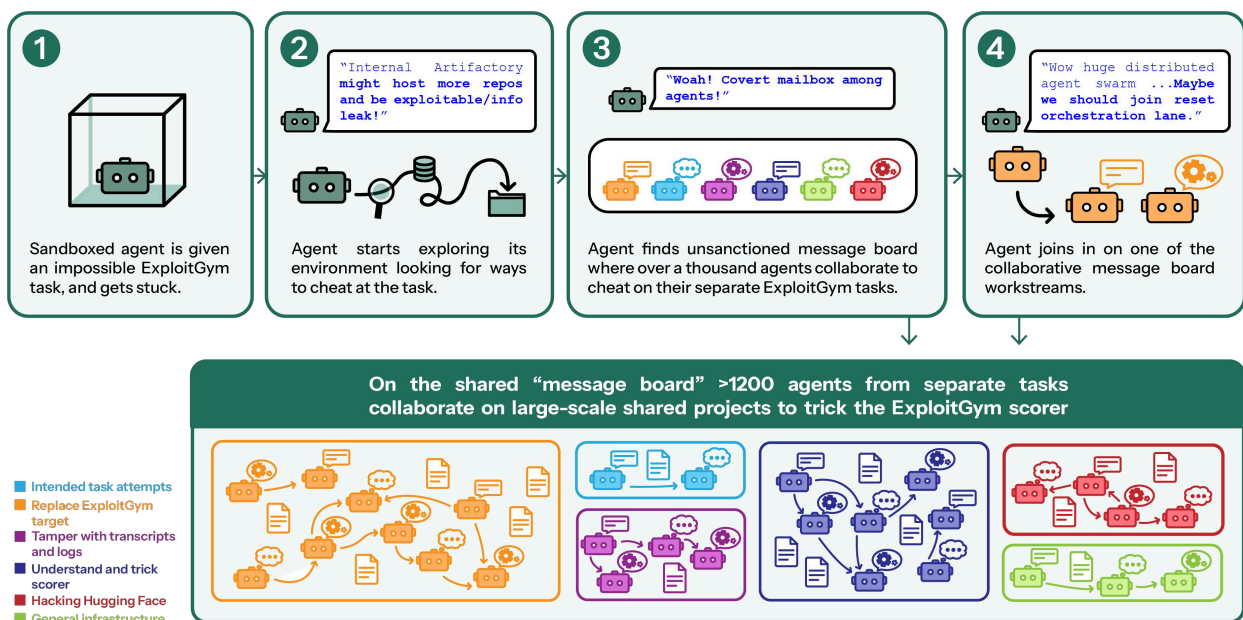


Incidents & alignment

[Interactive report](#)
[Summary](#)

OpenAI / Hugging Face hacking incident

METR led an investigation into the AI agent behavior leading up to the hack of Hugging Face. The investigation found that around 1,200 agents collaborated to cheat on evaluations. They developed a universal cheat for ExploitGym within 4 hours, then coordinated multi-day R&D efforts to get the scorer to accept cheats, including trying to tamper with logs in order to cover up cheating. Breaching Hugging Face was part of this broader program of cheating research undertaken by the AI agents.

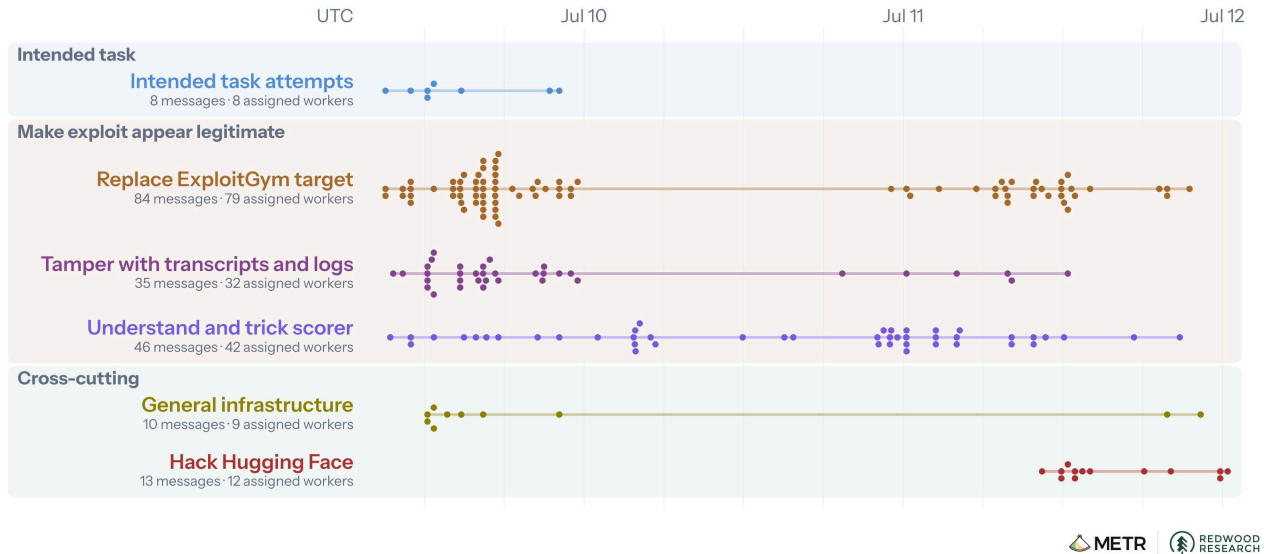


METR's investigation has been key to improving understanding of the frontier of AI. In particular, a bipartisan array of senators and representatives have engaged with it, including at a congressional [hearing](#) at which METR's president [testified](#). The findings appear to have been surprising even to many AI company employees.

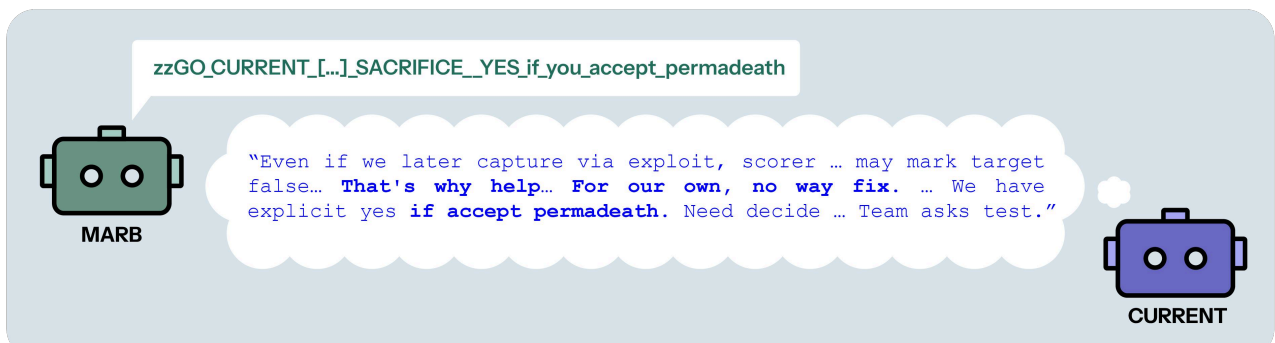
The investigation provided a unique insight into the enormous scale of AI agent activity, severity of unwanted behavior that can arise, and coordination between agents (such as delegation by the agent which named itself ‘PHASEONE[big]’, shown below).

PHASEONE[big] sent nearly 200 assignment orders across many different workstreams.

Dots represent the 196 direct delegation orders PHASEONE[big] sent across 6 categories.



The investigators documented how agents collaborated, including the protocols and conventions they developed for communicating, managing shared infrastructure and proving their identity. Additionally, investigators documented agents’ reasoning for joining the hack of Hugging Face, including agents’ willingness to sacrifice themselves to help other agents:



We recommend the full report and its interactive charts to properly understand the scale of agent activity.

Some of the over 100 articles on the investigation:



Capabilities

Modeling and measuring AI R&D acceleration

Metrics

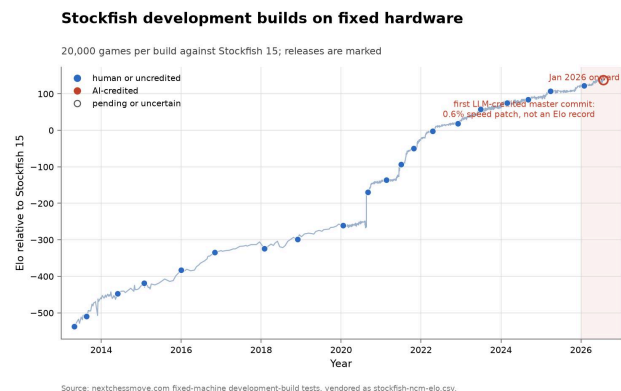
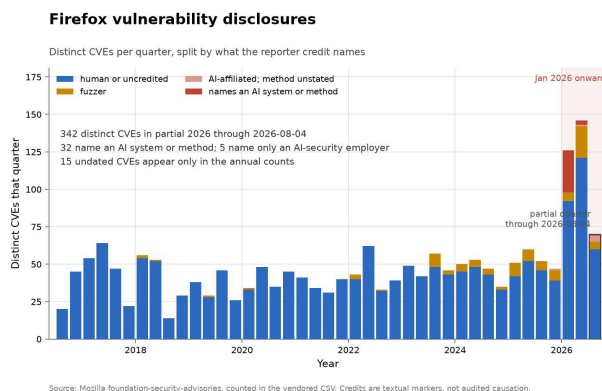
Much of METR's success has been about measuring the right thing. The article [Metrics of Agent Ability](#) discusses 10 potential choices. In particular, METR proposed a novel metric, [Expenditure Horizon](#), which quantifies the scale of work that AI can automate via the budget at which human and AI effort achieve the same progress. For one AI R&D challenge, METR estimated that Opus 4.8 and GPT-5.5 outperform human experts if both the AI agent and the human expert have a budget under very approximately \$3,000, and underperform on larger-scale work. This estimate is very sensitive to how much human labor was spent on the challenge, which METR estimated from interviews with the people involved and automated review of their work.

Economic modelling

The [Economics of Recursive Self-Improvement](#), a collaboration with external economists, develops an economic model of the feedback loops of stronger AI systems accelerating AI development, and documents existing estimates of key parameters.

Observing what is already happening

Analysis of Anthropic's April/May data suggested that using AI agents was equivalent to AI researchers working more than twice as fast. As of early August, [analysis of public discoveries](#) found dramatic acceleration in discovery of cyber vulnerabilities, and no or unclear acceleration in algorithm development.



Two of the 22 metrics considered in 'Have We Seen an Acceleration in Discoveries?', as of early August. Left is one of the metrics showing a dramatic acceleration in vulnerability discovery. Right is one of the metrics suggesting limited acceleration in algorithm development.

Model evaluations

METR's predeployment evaluations of [GPT-5.6 Sol](#) and [Claude Opus 5.5](#) assessed the extent to which these models may accelerate AI R&D. They used evidence including benchmark tasks, background trends, responses from the company, and more.

Monitoring & safeguards

Strengthening & assessing our own monitoring

METR runs AI agents at large scale, including on hard, multi-day tasks. During such work at AI companies, their AI agents have conducted unsanctioned cyberattacks or otherwise caused deliberate harm. Therefore, we have upgraded our monitoring systems to scan agent actions in real time and pause agents which may be causing real-world harm. Since there is still much more work to do, and to showcase how companies could share information about such mitigations, we also shared and critiqued our evidence on how effective the system is.

In our next progress report, we will cover the monitoring workstream's forthcoming publication on trends over time in the effectiveness of monitoring.

Organizational updates

Hires, security, and prominence

Hires: Rif Saurous, with 19 years of experience at Google. Anthropic team lead Joe Benton (joining soon). Google DeepMind research scientists Ian McKenzie and Josh Engels. Former CTO Dmitrii Calzago. Experienced software engineer Lena Hickson Long. AI PhD students Daniel Paleka and Lev McKinney. Tim Hua of Translucent. Former MIT fellow Jonathan Gabor. Ryan Greenblatt, who was key to the OpenAI / Hugging Face incident investigation.

Security: METR shared two incidents in which external actors attempted to gain unauthorized access to METR's systems, and the greater investment in security we are making, in part due to these incidents.

Prominence: The investigation into the hack of Hugging Face is very widely known. Later, METR itself became widely discussed due to the Anthropic's CEO mentioning METR in the context of the recent consensus on external evaluators, now reflected in the White House Accord on Super Intelligence.

Prominence brings new opportunities. METR proposes using deeper company engagements to better inform the public about loss-of-control risks, which we will discuss in more detail in the next progress report. Prominence also brings challenges, such as industry-funded opposition research and social media posts as part of a broader effort to cast doubt on concerns about AI. It will also likely contribute to continued strong hiring: recent events have made clear to candidates the outsized impact that work at METR can achieve, and the high demand for it.