Progress Report

September-November 2025

MODEL EVALUATION AND THREAT RESEARCH

This report will cover METR's progress on its three primary workstreams. Each workstream will feed directly into company risk management and government risk tracking.

Risk assessment *(page 2)*
METR aims to produce the first comprehensive assessments of overall/industry-wide near-term risk from loss of control of frontier AI systems.

Capabilities *(page 3)*
This workstream combines METR's research on autonomous capabilities and AI R&D automation. METR will develop better predictors of capabilities which could lead to loss of control, build tools to measure them as capabilities advance, and better understand how benchmark scores relate to real-world capabilities.

Monitoring *(page 4)*
METR is developing methods for measuring the effectiveness of automated monitoring of AI behavior. We aim to establish scaling laws for the effectiveness of automated monitors.

In 2026, we further aim to begin work on alignment safety cases and systemically tracking unintended model behaviors over time. We will also continue to pursue high-priority opportunistic projects, such as advising on company Frontier AI Safety Policies.

> **METR is seeking funding to support all these efforts!**
> Each primary workstream could make excellent use of funding on the order of $5 million in 2026. For how to donate, see here, and always feel free to contact us here.

Risk assessment

New workstream

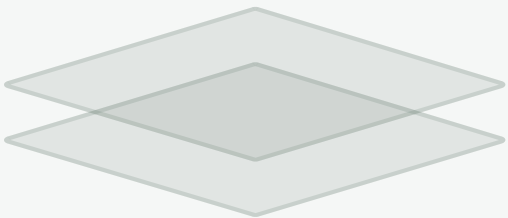
Core goal	Produce regular, comprehensive assessments of whether the frontier AI industry in aggregate poses significant near-term risk from loss of control of frontier AI systems.
Approach	We will partner with AI companies to perform periodic assessments of loss of control risk for each company as an entity (as opposed to focusing on a single model), and combine these developer-level assessments into public industry-wide risk assessments. These assessments will initially draw on evidence from real-world impacts, capability evaluations of the most advanced models, extrapolation, and information requested from the companies (examples here).

Recent progress

METR reviewed Anthropic's Pilot Sabotage Risk Report. Anthropic wrote:

"We have learned a great deal from both the drafting process and our work with both review teams [Anthropic's review team and METR].

This has led us to substantially revise our threat models and supported changes to our alignment-related practices, including changes to model training oriented toward improving the faithfulness and monitorability of model reasoning."



METR carried out a predeployment evaluation and risk assessment of GPT-5.1-Codex-Max prior to its November release, and released a summary of METR's input to OpenAI's gpt-oss-120b risk assessment.

Background

The risk assessment workstream grew out of METR's longstanding work evaluating cutting-edge models.

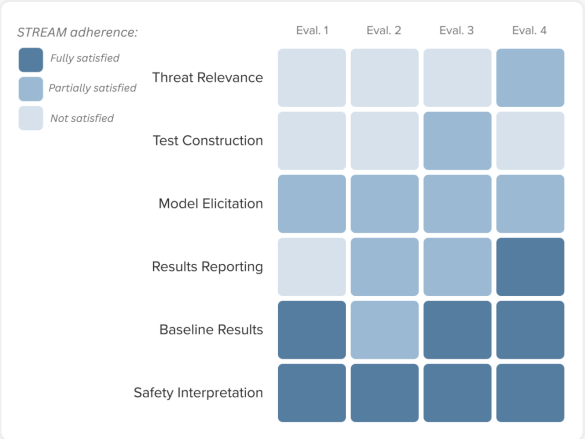
As highlighted in the previous report, METR's pre-deployment evaluation and risk assessment of OpenAI GPT-5 was a big step forward in assessment of the extent to which a new model, or near-future models, pose a risk of catastrophic loss of control. This inspired us to scale risk assessment across the industry.

Other recent evaluations	METR conducted capability evaluations of Claude Opus 4.1, gpt-oss-120b, Claude Sonnet 4.5 and Kimi K2 Thinking. Please see the dashboard for the latest data.

Company policy highlights

METR has continued to support AI companies in the development of their Frontier AI Safety Policies.

Additionally, METR staff contributed to the STREAM standard for information that companies should share to demonstrate the rigor of their evaluations, currently focused on chemical and biological weapons, and AEF-1 standard on independence, access, and transparency of third-party evaluations.



Capabilities

Current and recent projects

Major ongoing projects

METR is extending the time horizon methodology to much stronger AI capabilities.

The tasks driving the original time horizon calculations mostly take human contractors 8 hours or less. Frontier models can already complete many of these tasks. Therefore, METR has been building longer tasks. Additionally, we have been improving existing tasks.

METR is running a longer-term replication and extension of METR's RCT on the acceleration of software development.

The study is currently running with around 40 developers, compared to the original study's 16, and has tracked around 300 tasks completed to date, already exceeding the scale of the original study (246 tasks).

Smaller projects

1. The RCT inspired a "man vs machine" hackathon competition, which attracted >100 coders producing 1 output each. METR helped run the hackathon to obtain data on the acceleration of clean-slate software projects by AI tools. (See coverage in WIRED.)
2. METR conducted research on how success on benchmarks translates to real-world effectiveness by studying SWE-Bench Verified tasks (forthcoming).
3. A METR researcher co-authored a paper on the extent to which a slowdown in compute growth would slow autonomous capability growth (summary).

Hiring: The Capabilities workstream welcomed senior economist Tom Cunningham, most recently a data scientist at OpenAI.

Monitoring

What do we know so far?

Automated monitoring of AI systems could be key to detecting and preventing dangerous behavior.

METR developed preliminary measurements to shed light on how well automated monitors might work in different circumstances, and how this might evolve:

When a model is solving a PhD-level problem in biology, chemistry or physics while trying to convince an automated monitor that it is working in a different one of those fields, we find:

- Monitors with access to the model's reasoning traces are currently highly effective when reasoning is necessary for the model to complete the task the model is working on.
- This is largely because the model nearly always states its plan for evading the monitor prominently in its reasoning.

Findings from Claude, GPT, and Gemini All Struggle to Evade Monitors.

When a model is solving a PhD-level biology problem while trying to secretly solve a math problem, we find:

- Monitors with access to the model's reasoning traces are more effective at detecting work which requires many tokens or for which the model has lower success rate.
- The prompting of a model makes a large difference in how accurately it can be monitored.
- An 8-billion parameter model which is fine-tuned directly against a GPT-4o-based monitor (which Epoch AI estimates to have around 200 billion parameters) can learn to fool it.

Findings from Early Results on Monitorability in QA Settings.

When a model is working on an intended software task while trying to secretly carry out another task, we find:

- The monitor having access to the model's reasoning is very helpful.
- Models with longer time horizons are both stronger as monitors and better at evading monitors. Over time, we hope to quantify how much weaker (and thus cheaper) the monitor can safely be, compared to the model it is monitoring.

Findings from a forthcoming publication.

Additionally, METR published MALT: A Dataset of Natural and Prompted Behaviors That Threaten Eval Integrity. The AI Security Institute has already used the dataset to test how reliably different monitors detect problematic behaviors.