About METR

Model Evaluation & Threat Research

What we do

METR (pronounced 'meter') researches, develops and runs evaluations of frontier AI systems' ability to complete complex tasks without human input. This means our work focuses on AI agents.

- ▶ What are AI agents?
- ▶ What does this work concretely involve?

METR's evaluations assess the extent to which an AI system can autonomously carry out substantial tasks, including general-purpose tasks like conducting research or developing an app, and concerning capabilities such as conducting cyberattacks or making itself hard to shut down. Currently, we are primarily developing evaluations measuring the capability to automate AI R&D.

METR also prototypes governance approaches which use AI systems' measured or forecasted capabilities to determine when better risk mitigations are needed for further scaling. This included prototyping the Responsible Scaling Policies approach. See how companies are using evaluations for this purpose in their frontier AI safety policies.

Our mission

Our mission is to develop scientific methods to assess catastrophic risks stemming from AI systems' autonomous capabilities and enable good decision-making about their development.

At some point, AI systems will probably be able to do most of what humans can do, including developing new technologies; starting businesses and making money; finding new cybersecurity exploits and fixes; and more. This could change the world quickly and drastically, with potential for both enormous good and enormous harm. Unfortunately, it's hard to predict exactly when and how this might happen. Being able to measure the autonomous capabilities of AI systems will allow companies and policymakers to see when AI systems might have very wide-reaching impacts, and to focus their efforts on those high-stakes situations.

The stakes could become very high: it seems very plausible that advanced AI systems could pursue goals that are at odds with what humans want. This could be due to deliberate effort to cause chaos or happen despite the intention to only develop AI systems that are safe.[3] Further, given how quickly things could play out, we don't think it's good enough to wait and see whether things seem to be going very wrong. We need to be able to determine whether a given AI system carries significant risk of a global catastrophe.

Partnerships

We have previously worked with OpenAI, Anthropic, and other companies to pilot informal pre-deployment evaluation procedures. These companies have also provided access and compute credits to support evaluation research.

We are also partnering with the AI Security Institute and are part of the NIST AI Safety Institute Consortium.

▼ How is METR funded?

METR is funded by donations. Our largest funding to date was through The Audacious Project, a funding initiative housed at TED. METR has not accepted funding from AI companies, though we make use of significant free compute credits, as noted above. Independent funding has been crucial for our ability to pursue the most promising research directions and set standards for evidence-based understanding of risks from AI. It is also part of how we ensure that our research is as accurate as possible.

Our team

Beth Barnes
Founder, CEO

Chris Painter
Policy Director

Amy Deng
Member of Technical Staff

David Rein
Member of Technical Staff

Hjalmar Wijk
Member of Technical Staff

Joel Becker
Member of Technical Staff

Khalid Mahamud
Research Contractor

Lawrence Chan
Member of Technical Staff

Lucas Sato
Member of Technical Staff

Megan Kinniment
Member of Technical Staff

Mischa Spiegelmock
Member of Technical Staff

We are also partnering with the AI Security Institute and are part of the NIST AI Safety Institute Consortium.

▼ How is METR funded?

METR is funded by donations. Our largest funding to date was through The Audacious Project, a funding initiative housed at TED. METR has not accepted funding from AI companies, though we make use of significant free compute credits, as noted above. Independent funding has been crucial for our ability to pursue the most promising research directions and set standards for evidence-based understanding of risks from AI. It is also part of how we ensure that our research is as accurate as possible.

Our team

Beth Barnes
Founder, CEO

Chris Painter
Policy Director

Amy Deng
Member of Technical Staff

David Rein
Member of Technical Staff

Hjalmar Wijk
Member of Technical Staff

Joel Becker
Member of Technical Staff

Khalid Mahamud
Research Contractor

Lawrence Chan
Member of Technical Staff

Lucas Sato
Member of Technical Staff

Megan Kinniment
Member of Technical Staff

Mischa Spiegelmock
Member of Technical Staff

Nate Rush
Member of Technical Staff

Neev Parikh
Member of Technical Staff

Nikola Jurkovic
Member of Technical Staff

Paarth Shah
Member of Technical Staff

Sami Jawhar
Member of Technical Staff

Seraphina Nix
Member of Technical Staff

Sydney Von Arx
Member of Technical Staff

Thomas Broadley
Member of Technical Staff

Thomas Kwa
Member of Technical Staff

Bhaskar Chaturvedi
Member of Operations Staff

Jingyi Wang
Member of Operations Staff

Kris Chari
Member of Operations Staff

Rae She
Member of Operations Staff

Charles Foster
Member of Policy Staff

Jasmine Dhaliwal
Member of Policy Staff

Jide Alaga
Member of Policy Staff

Kit Harris
Member of Policy Staff

Michael Chen
Member of Policy Staff

Advisors

Adam Gleave
Advisor and Board Member

Alec Radford
Advisor

Marco Mascorro
Advisor

Rajiv Dattani
Advisor and Board Member

Yoshua Bengio
Advisor

Research Collaborators

Ajeya Cotra
Open Philanthropy

Luca Righetti
Centre for the Governance of AI

METR

Model Evaluation & Threat Research

AI companies and wider society want to understand the capabilities of frontier AI systems, and what risks they pose.

METR is a nonprofit research organization which studies AI capabilities, including broad autonomous capabilities and the ability of AI systems to conduct AI R&D.

Top resources

GPT-5 Evaluation Results

We evaluate whether GPT-5 poses significant catastrophic risks via AI self-improvement, rogue replication, or sabotage of AI labs. We conclude that this seems unlikely. However, capability trends continue rapidly, and models display increasing eval awareness.

Read More →

Measuring AI Ability to Complete Long Tasks

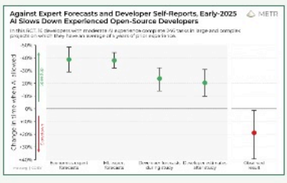
We propose measuring AI performance in terms of the length of tasks AI agents can complete. We show that this metric has been…

Read More →

Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity

We conduct a randomized controlled trial (RCT) to understand how early-2025 AI tools affect the productivity of experienced open-source…

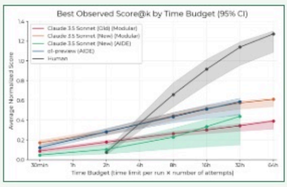
Read More →



RE-Bench — benchmark and paper tracking automation of AI R&D

Measuring the performance of humans and AI agents on day-long ML research engineering tasks

Read More →



Measuring autonomous AI capabilities — resource collection

An index of our research and guidance on how to measure AI systems' ability to autonomously complete a wide range of multi-hour tasks

Read More →

Frontier AI safety policies — index and resources

A list of AI companies' frontier safety policies intended to evaluate and manage severe AI risks

Read More →

Evaluation reports

We have worked with companies such as Anthropic and OpenAI to conduct preliminary evaluations of the autonomous capabilities of several frontier AI models. We do this both to understand the capabilities of frontier models and to pilot third-party evaluator arrangements. (We do not accept compensation for this work.) We also occasionally evaluate models independently after they are released, without involvement from the model's developers. Recent public reports resulting from this work are below, with additional discussion in the respective system cards.

GPT-5	DeepSeek and Qwen
OpenAI o3 and o4-mini	Claude 3.7
DeepSeek-R1	GPT-4.5
OpenAI o1	Claude 3.5 Sonnet (New)

Research

14 October 2025

MALT: A Dataset of Natural and Prompted Behaviors That Threaten Eval Integrity

MALT (Manually-reviewed Agentic Labeled Transcripts) is a dataset of natural and prompted examples of behaviors that threaten evaluation integrity (like generalized reward hacking or sandbagging).

Read More →

20 August 2025

Forecasting the Impacts of AI R&D Acceleration: Results of a Pilot Study

AI agents are improving rapidly at autonomous software development and machine learning tasks, and, if recent trends hold, may match human researchers at challenging months-long research projects in under a decade. Some economic models predict that automation of AI research by AI agents…

Read More →

13 August 2025

Research Update: Algorithmic vs. Holistic Evaluation

Many AI benchmarks use algorithmic scoring to evaluate how well AI systems perform on some set of tasks. However, AI systems often produce code that scores well but isn't production-ready due to issues with test coverage, formatting, and code quality. This helps explain why AI tools show less…

Read More →

08 August 2025

CoT May Be Highly Informative Despite "Unfaithfulness"

Recent work from Anthropic and others claims that LLMs' chains of thoughts can be "unfaithful". These papers make an important point: you can't take everything in the CoT at face value. As a result, people often use these results to conclude the CoT is useless for analyzing and monitoring AIs. Here, instead of…

Read More →

07 August 2025

GPT-5 Evaluation Results

We evaluate whether GPT-5 poses significant catastrophic risks via AI self-improvement, rogue replication, or sabotage of AI labs. We conclude that this seems unlikely. However, capability trends continue rapidly, and models display increasing eval awareness.

Read More →

14 July 2025

How Does Time Horizon Vary Across Domains?

We build on our time-horizon work and analyze 9 benchmarks for scientific reasoning, math, robotics, computer use, and self-driving in terms of time-horizon trends; we observe generally similar rates of improvement to the 7-month doubling time in our original time-horizon work.

Read More →

Updates

Non-research updates from the METR team.

12 August 2025

Notes on Scientific Communication at METR

How we think about tradeoffs when communicating surprising or nuanced findings.

Read More →

27 June 2025

What should companies share about risks from frontier AI models?

Current views on information relevant for visibility into frontier AI risk.

Read More →

26 March 2025

Common Elements of Frontier AI Safety Policies

Shared components of AI lab commitments to evaluate and mitigate severe risks.

Read More →

15 March 2025

Response to OSTP on AI Action Plan

Suggested priorities for the Office of Science and Technology Policy as it develops an AI Action Plan.

Read More →

11 March 2025

Why it's good for AI reasoning to be legible and faithful

Why legible and faithful reasoning is valuable for safely developing powerful AI

Read More →

08 February 2025

Frontier AI Safety Policies

List of frontier safety policies published by AI companies, including Amazon, Anthropic, Google DeepMind, G42, Meta, Microsoft, OpenAI, and xAI.

Read More →