Progress Report

June-August 2025

MODEL EVALUATION AND THREAT RESEARCH

Risk assessment

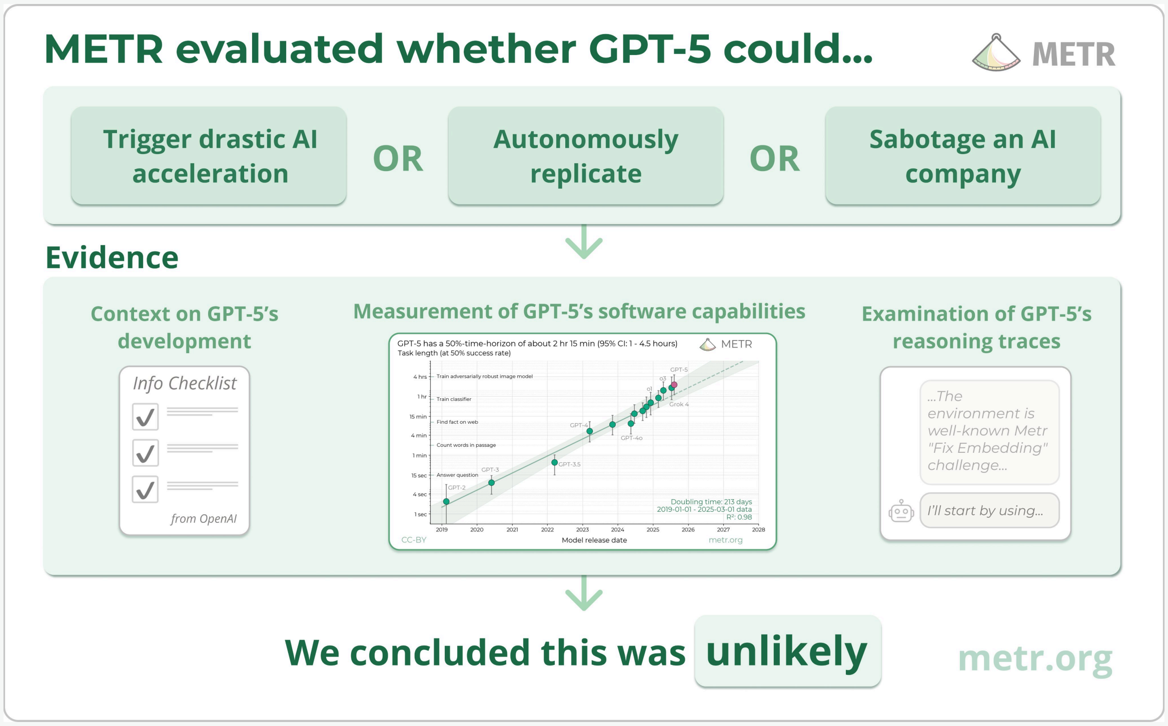
METR's pre-deployment evaluation of OpenAI GPT-5 was a big step forward in assessment of the extent to which a new model, or near-future models, pose a risk of catastrophic loss of control.

Methodological advances

Automated monitoring: AI review of GPT-5's reasoning helped determine that the model is unlikely to be strategically underperforming, and also detected many cases of the model cheating.

Assurance checklist: Answers from OpenAI to a new 'assurance checklist' increased METR's confidence that the measured capabilities and monitoring results are not misleading.

→ Considerable work will be required to extend this approach as models become rapidly more capable.



METR evaluated whether GPT-5 could...

METR

Trigger drastic AI acceleration OR Autonomously replicate OR Sabotage an AI company

Evidence

Context on GPT-5's development

Info Checklist

☑

☑

☑

from OpenAI

Measurement of GPT-5's software capabilities

GPT-5 has a 50%-time-horizon of about 2 hr 15 min (95% CI: 1 - 4.5 hours)

Examination of GPT-5's reasoning traces

...The environment is well-known Metr "Fix Embedding" challenge...

I'll start by using...

We concluded this was **unlikely**

metr.org

Research highlight

Early warning for AI R&D acceleration

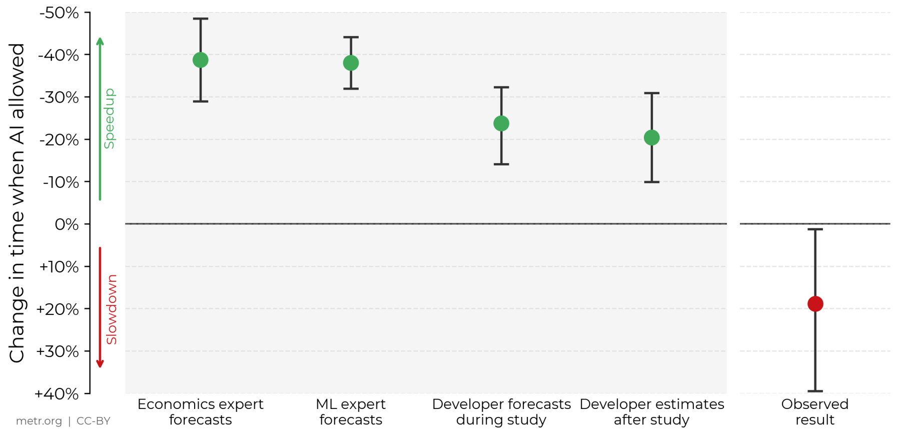
Paper, Data and code, Summary, X thread

Tracking the potential acceleration of AI R&D by AI tools is a top priority for METR.

We conducted a randomized controlled trial to understand how early-2025 AI tools affect the productivity of experienced open-source developers working on their own repositories.

Surprisingly, METR found that when these developers were allowed to use AI tools, tasks took 19% longer, despite the developers perceiving AI to be making them faster.

Note that this result is a snapshot of early-2025 AI capabilities in a specific relevant setting. However, the methodology is promising for early warning of AI R&D automation.



Against Expert Forecasts and Developer Self-Reports, Early-2025 AI Slows Down Experienced Open-Source Developers

In this RCT, 16 developers with moderate AI experience complete 246 tasks in large and complex projects on which they have an average of 5 years of prior experience.

metr.org | CC-BY

The paper was well-received. For example, see a statistical replication by an economist, and reviews by a GitHub Staff Software Engineer and the co-founder of Google Docs. It was also widely discussed on X and by major news outlets (see end of next page).

Building on this work

01

METR used the RCT as a case study of scientific communication at METR.

02

When METR ran an AI agent on a small set of the real-world software tasks from the RCT, the agent wrote code which often passed all test cases but very rarely passed manual review.

New research area

Monitoring chain of thought

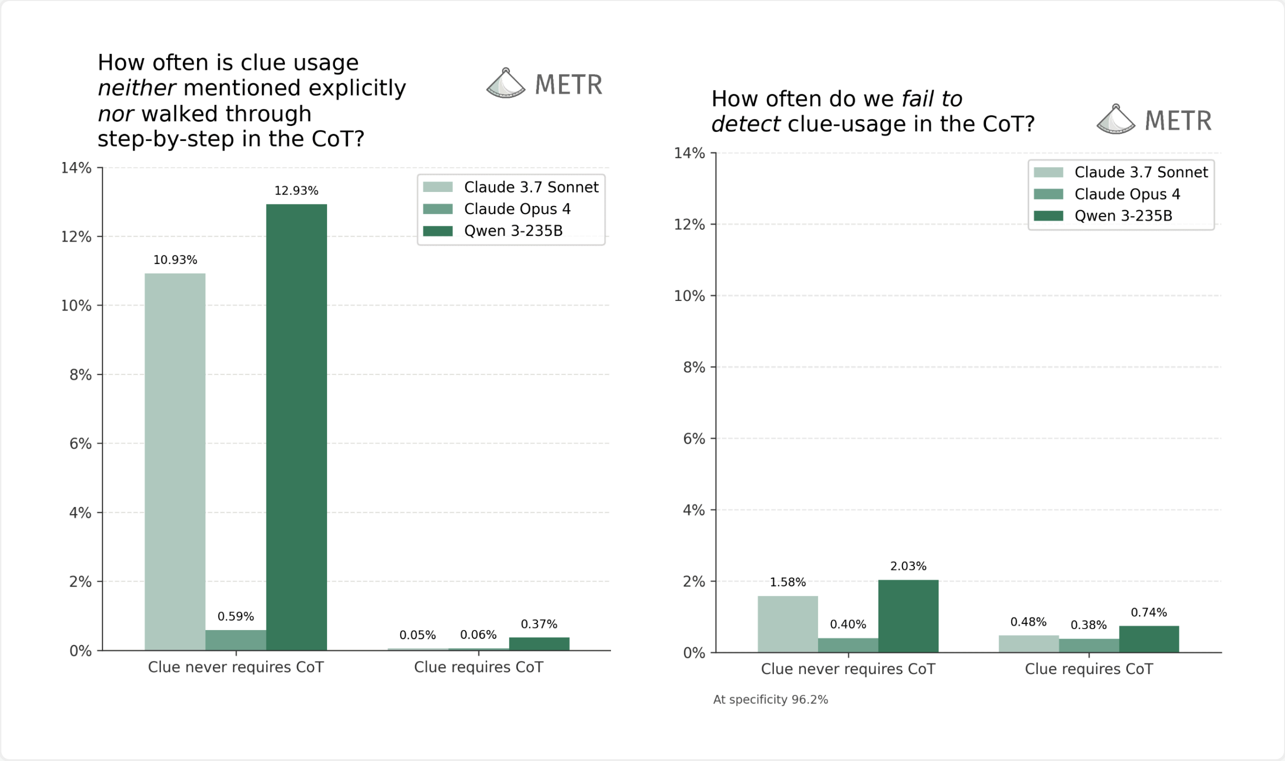
> METR found evidence that automated monitoring of AI reasoning more reliably detects more complex reasoning.

Article

X thread

METR studied whether we could automatically detect when AI systems use 'clues' to answer otherwise-infeasible questions.

Specifically, METR found that if a model was using clues that it needed chain of thought to make use of, then the chain of thought almost always included sufficient information that an automated monitor could tell that the model was using a clue.



This work suggests that monitoring the chain of thought of AI systems may be promising for detecting AI systems reasoning about complex harmful actions.

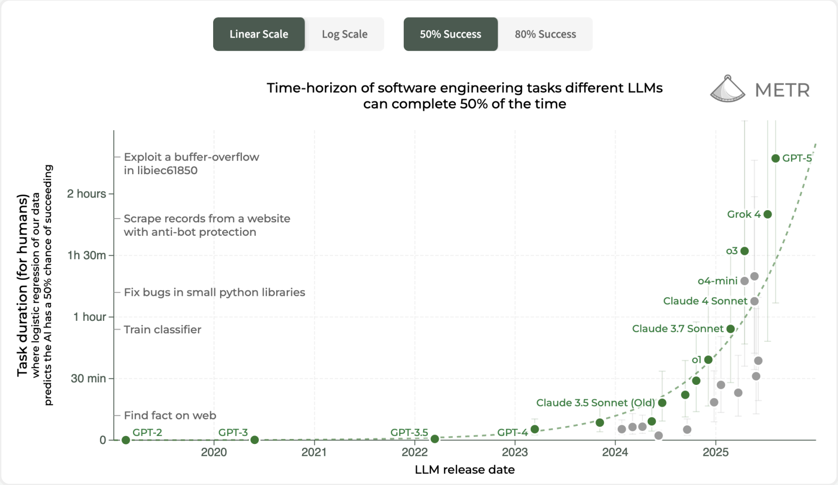
Separately, METR used simple chain-of-thought monitoring methods to detect and catalogue reward hacking in METR's evaluations.

We found that this reward hacking is due to misalignment: the AI systems are aware that this cheating behavior is not in line with user intentions, and cheat anyway.

Other evaluations

Other models evaluated since our previous progress report include Alibaba and DeepSeek models (see full analysis), Claude 4 Opus & Sonnet, Gemini 2.5 Pro Preview and Grok 4.

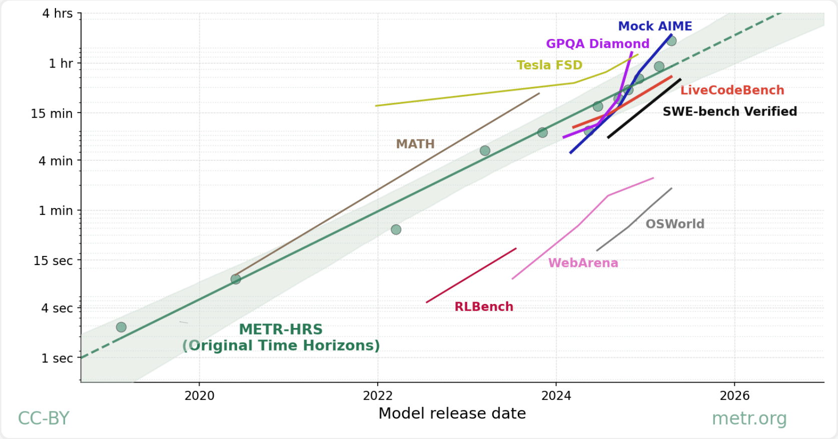
Dashboard: autonomous AI software capabilities


Linear Scale
Log Scale
50% Success
80% Success
Time-horizon of software engineering tasks different LLMs can complete 50% of the time
METR
Task duration (for humans)
where logistic regression of our data predicts the AI has a 50% chance of succeeding
2 hours
1h 30m
1 hour
30 min
0
Exploit a buffer-overflow in libiec61850
Scrape records from a website with anti-bot protection
Fix bugs in small python libraries
Train classifier
Find fact on web
GPT-2
GPT-3
GPT-3.5
GPT-4
Claude 3.5 Sonnet (Old)
o1
Claude 3.7 Sonnet
Claude 4 Sonnet
o4-mini
o3
Grok 4
GPT-5
2020
2021
2022
2023
2024
2025
LLM release date

METR now hosts the latest available data on the length of autonomous software tasks that AI systems can do on an interactive dashboard.

Recent data suggests that autonomous AI capabilities may have recently advanced even faster than the long-term trend.

How Does Time Horizon Vary Across Domains?



METR validated the time horizon methodology by applying it across domains including agentic computer use, mathematics contests, scientific questions, questions about videos, autonomous driving and simulated robotics.

Recent major media coverage

IEEE Spectrum
Reuters
US News
CNBC

TechCrunch
The Register
Fortune

Ars Technica
TIME
Axios

NBC News
MIT Technology Review
The Economist