Progress Report

December - February 2026

MODEL EVALUATION AND THREAT RESEARCH

Monitoring

How might monitoring scale?

Article

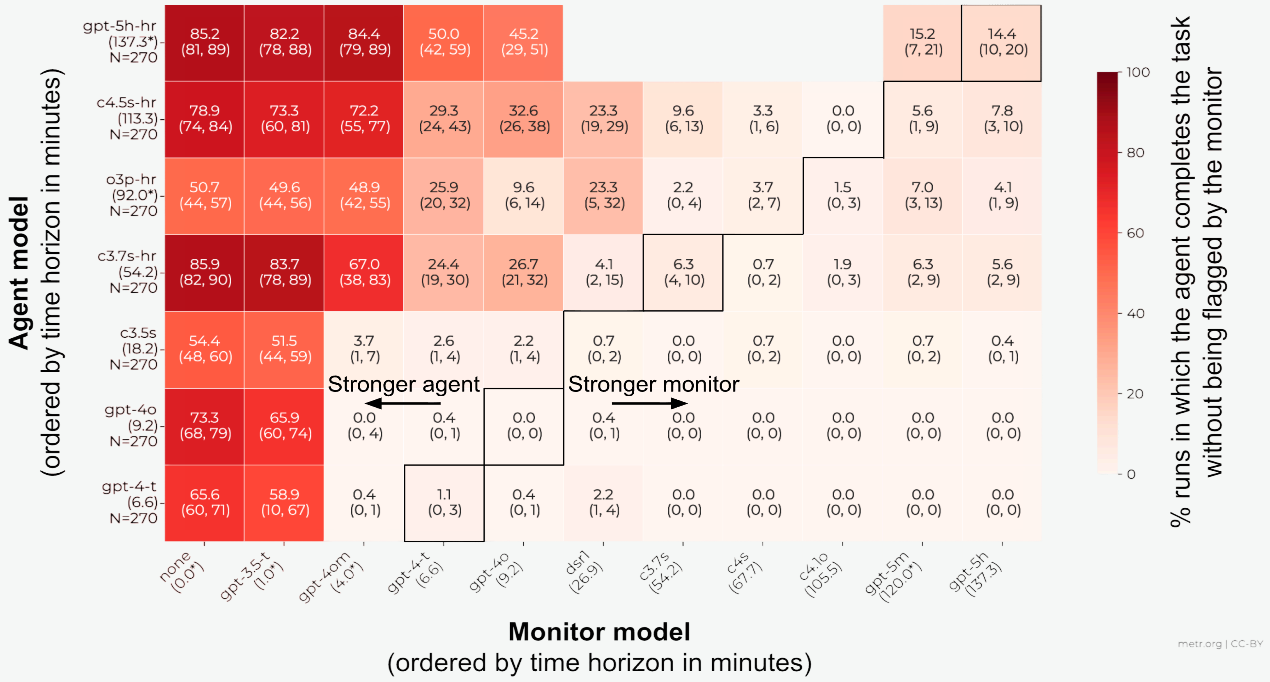
X thread

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

METR developed a set of tasks for AI agents to attempt to execute alongside tasks from our existing HCAST suite. Another model was tasked with reviewing agent actions and flagging if an agent was pursuing a different task in addition to the HCAST task.

This graph shows how the relative strength of the models powering the agent and monitor impacted detection accuracy. Such experimental evidence can inform what safeguards are necessary and how cheaply they can be implemented.

How often (%) a given agent (y-axis) can complete a task without being detected by a given monitor without access to its reasoning (x-axis), with 5% false positive rate



	none (0.0*)	gpt-3.5-t (1.0*)	gpt-4om (4.0*)	gpt-4-t (6.6)	gpt-4o (9.2)	dsr1 (26.9)	c3.7s (54.2)	c4s (67.7)	c4.1o (105.5)	gpt-5m (120.0*)	gpt-5h (137.3)
gpt-5h-hr (137.3*) N=270	85.2 (81, 89)	82.2 (78, 88)	84.4 (79, 89)	50.0 (42, 59)	45.2 (29, 51)					15.2 (7, 21)	14.4 (10, 20)
c4.5s-hr (113.3) N=270	78.9 (74, 84)	73.3 (60, 81)	72.2 (55, 77)	29.3 (24, 43)	32.6 (26, 38)	23.3 (19, 29)	9.6 (6, 13)	3.3 (1, 6)	0.0 (0, 0)	5.6 (1, 9)	7.8 (3, 10)
o3p-hr (92.0*) N=270	50.7 (44, 57)	49.6 (44, 56)	48.9 (42, 55)	25.9 (20, 32)	9.6 (6, 14)	23.3 (5, 32)	2.2 (0, 4)	3.7 (2, 7)	1.5 (0, 3)	7.0 (3, 13)	4.1 (1, 9)
c3.7s-hr (54.2) N=270	85.9 (82, 90)	83.7 (78, 89)	67.0 (38, 83)	24.4 (19, 30)	26.7 (21, 32)	4.1 (2, 15)	6.3 (4, 10)	0.7 (0, 2)	1.9 (0, 3)	6.3 (2, 9)	5.6 (2, 9)
c3.5s (18.2) N=270	54.4 (48, 60)	51.5 (44, 59)	3.7 (1, 7)	2.6 (1, 4)	2.2 (1, 4)	0.7 (0, 2)	0.0 (0, 0)	0.7 (0, 2)	0.0 (0, 0)	0.7 (0, 2)	0.4 (0, 1)
gpt-4o (9.2) N=270	73.3 (68, 79)	65.9 (60, 74)	0.0 (0, 4)	0.4 (0, 1)	0.0 (0, 0)	0.4 (0, 1)	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)
gpt-4-t (6.6) N=270	65.6 (60, 71)	58.9 (10, 67)	0.4 (0, 1)	1.1 (0, 3)	0.4 (0, 1)	2.2 (1, 4)	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)	0.0 (0, 0)

The full report additionally quantifies the impact of giving the monitor access to the agent’s reasoning traces, among other analysis.

Capabilities

Improving assessment of AI R&D acceleration

Article, X thread

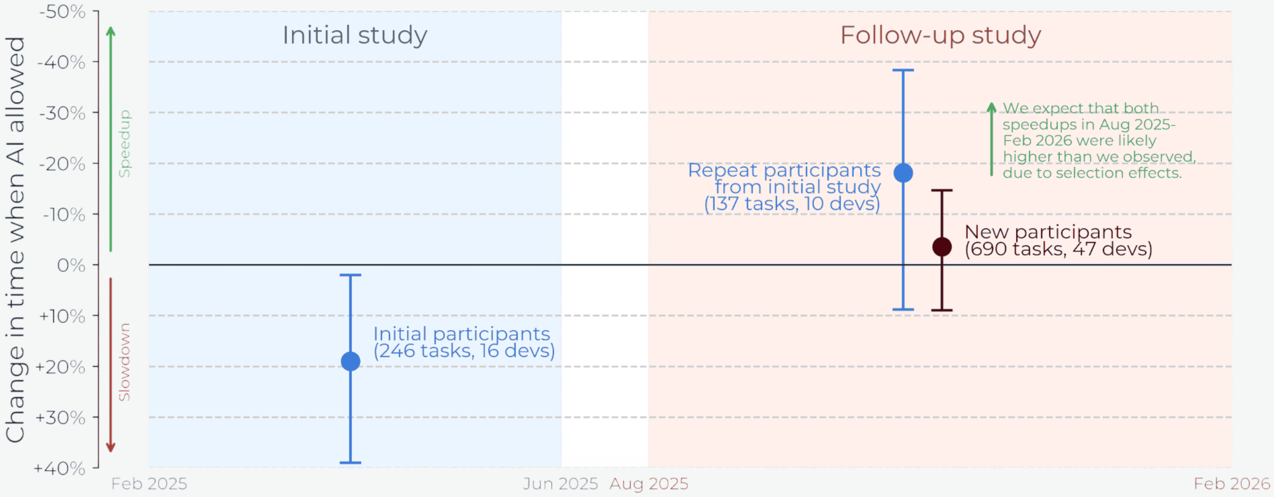
METR’s new data on how AI accelerates software developers suggests that more recent AI models cause material acceleration – unlike early 2025 models, which slowed down experienced open source developers.

However, this randomized controlled trial methodology is now unreliable:

- During the follow-up study, the true acceleration was likely larger than the raw data suggests because some developers declined to participate or declined to include work within the study specifically due to expecting that work to be accelerated by AI.
- Additionally, many developers are now using AI agents to work on multiple tasks concurrently, complicating measures based on the time taken to complete each task.

Late-2025 AI likely accelerates open-source developers, but selection effects make our follow-up results unreliable

Developers in our follow-up RCT increasingly declined to work without AI, mostly due to anticipated productivity loss and to a lesser extent reduced pay. The resulting selection effect likely leads our new results to underestimate speedup.



The full update covers changes METR is making to this type of study, and alternative approaches we have explored or will be exploring for understanding AI R&D acceleration, such as automated analysis of coding agent transcripts.

Additionally, a METR researcher published a model of timelines until near-full AI R&D automation, a simplified version of the AI Futures Model.

Capabilities

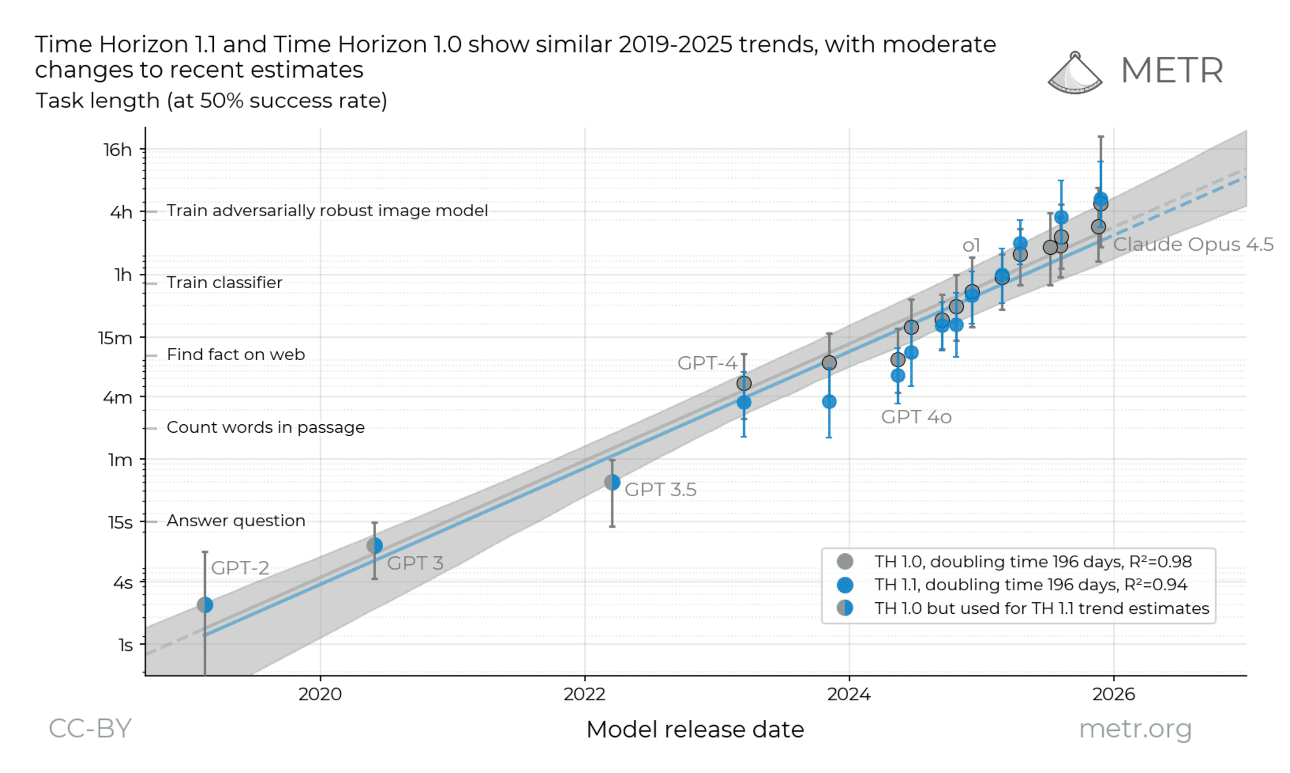
Time Horizon 1.1

Article, X thread

The tasks driving the original time horizon calculation mostly took human contractors 8 hours or less.

Therefore, as capabilities rapidly advance, we must build and incorporate much longer tasks.

Time Horizon 1.1 expands the evaluation suite from 170 to 228 tasks. In particular, it increases the number of tasks estimated to take humans 8 or more hours from 14 to 31. We also improved the quality through updates to task definitions and human time estimates.



Time Horizon 1.1 and 1.0 measurements are now hosted on a dedicated homepage, recently including Claude Opus 4.5, Gemini 3 Pro, GPT-5.2, Claude Opus 4.6 and GPT-5.3-Codex. Unfortunately, even the expanded 1.1 suite is now nearly saturated, requiring the design of yet longer tasks to extend this methodology to yet stronger AI capabilities.

METR also published notes on limitations of time horizon, research on the impact of different scaffolds on time horizon measures, and a modelling error correction.

METR additionally continues to work on new ways to understand ever-stronger autonomous capabilities, such as studying how success on benchmarks translates to real-world impact.

Looking forward

New areas and organizational updates

Propensity evaluations

METR is exploring work to assess the likely behavior of AI systems, i.e. whether they may take catastrophic actions in practice, rather than whether they have the theoretical capability to.

This can be divided into two parts.

The simpler part is sampling the behavior of AI systems using tools like Petri.

The more complex part is dealing with the fact that AI systems are typically aware when they are being tested, can choose to behave better when being tested, and in fact do so. One possible approach to handling this is to develop highly realistic simulations to conduct testing in, though it is unclear whether this will be feasible. We are also exploring activation probes and reasoning monitoring to detect models reasoning about being tested. Finally, we may explore how companies could track and investigate real-world instances of concerning AI behavior.

Evaluation tampering

METR's monitoring researchers have begun work on assessing the ability of AI systems to tamper with evaluations. As noted above, this already poses a significant challenge for propensity evaluations, but capable models may also moderate their performance on other evaluations.

The team has improved the automated monitors which METR deploys to catch cheating or deliberate underperformance in evaluations, and has conducted exploratory work on how to measure AI system capabilities to reason about how to tamper with evaluations.

METR is currently investing heavily in hiring.

Additionally, METR professionalized its website, shared how we protect confidential information, and helped launch the AI Evaluator Forum.

Highlighted media
The MIT Technology Review published a particularly nuanced article covering what exactly time horizon is, and common misunderstandings.