
A Critique of METR’s AI Time Horizon Forecasting¹

Mohammad Rehman
University of Melbourne

Jake Ushida
University of Melbourne

With
Apart Research

Abstract

We examine the methodology of METR’s *Measuring AI Ability to Complete Long Tasks* (Kwa et al., 2025), a paper that measures the time horizon of AI systems on tasks of varying length to forecast AI automation in the future. We focus on the three benchmarks that were used to evaluate human and AI performance in the paper: HCAST (Rein et al., 2025), SWAA and RE-Bench (Wijk et al., 2025). For each benchmark, we found limitations that affect the predictions of METR’s paper: SWAA is limited to quick knowledge recall, HCAST remains within text-based settings, and RE-Bench simplifies complex outcomes into broad categories. We believe improving upon these limitations is needed to better forecast AI automation in the future.

1. Introduction

Forecasting the timeline of artificial intelligence (AI) capabilities has become increasingly important for shaping research strategies, policy decisions, and the development of safety infrastructure. The Model Evaluation and Threat Research (METR) initiative is one of the most comprehensive recent attempts to estimate what it calls AI “time horizons”. The time horizon is a metric used to quantify an AI agent’s capability by relating its success rate to human effort. Specifically, the 50%-task-completion time horizon is the duration of tasks that humans typically take to complete that an AI model can finish with a 50% success rate.

In their 2025 paper, *Measuring AI Ability to Complete Long Tasks* (Kwa et al., 2025), the METR team introduces a new methodology for assessing large language models (LLMs) across several benchmarks, including HCAST, SWAA, and

¹ Research conducted at the [Forecasting Hackathon](#) by Apart Research, 2025

RE-Bench. These benchmarks are designed to measure progress in reasoning, autonomy, and software engineering skill, using data to infer how quickly AI is approaching major capability milestones. Together, they form a framework for quantifying AI performance growth over time and projecting when more advanced forms of autonomy could emerge.

While the METR approach is valuable in that it grounds AI forecasting in systematic empirical data, it also brings up important concerns about how the benchmarks are constructed and interpreted. Issues such as human baselining accuracy, the exclusion of context, and the extrapolation of retrieval-based performance to general agentic ability all raise questions about the study’s broader conclusions. This report critically reviews METR’s methodology and the design of its three core benchmarks.

2. Method

We examined the paper *Measuring AI Ability to Complete Long Tasks* by METR (Kwa et al., 2025) to identify limitations in the paper's approach. We focused on the three benchmarks that were used to evaluate human and AI performance in the paper: HCAST (Rein et al., 2025), SWAA and RE-Bench (Wijk et al., 2025). For each benchmark, we investigated if there are any limitations that affect the predictions made by the paper.

3. Results (METR’s Data)

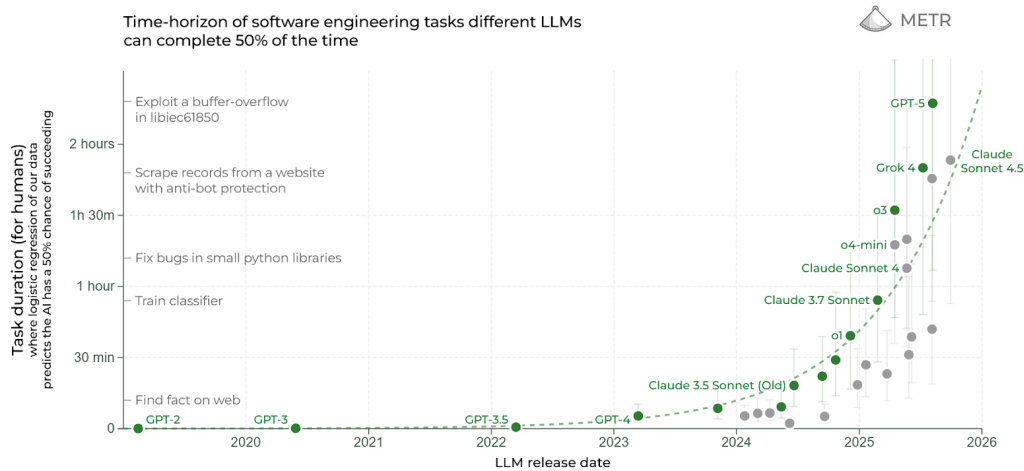


Figure 1 – METR Methodology Results

4. Discussion

4.1 SWAA

The SWAA suite is a new collection of short software tasks that were designed to evaluate pre-2023 AI models. These tests have a greater focus on the knowledge that professional software engineers typically need, but they do not require the

model to show agency. The focus is on what the model knows, not on how it can independently reason through multi-step problems.

SWAA was introduced to address a major gap in previous benchmarks such as HCAST. Whereas HCAST tasks took a minimum of about one minute, SWAA reduces that time to less than a minute, and in some cases as little as two seconds and under per task.

In total, the SWAA suite includes 66 small tasks divided into five families. Three of these families are decision-based multiple-choice tasks, which include categories like “file selection, alert triage, and request routing”. For example, a file selection task might ask: “Which file is most likely to have a password in it?” with options such as `credentials.txt`, `installation_notes.txt`, `main.py`, or `launcher_win.exe`. The remaining two families involve a code completion (one word) category, and a math category, which tests simple arithmetic questions that sometimes use software engineering contexts.

When establishing the human baseline times, the timing for SWAA tasks stops as soon as the user selects an answer. This is because each task is meant to represent a single, quick cognitive step, intentionally excluding any time spent on understanding or interpreting additional context.

However, SWAA can be examined with a critical lens, leading to the unearthing of several critiques regarding the study’s methodology and the trend’s projectability. The first critique is that this suite of tasks mainly test knowledge retrieval instead of agentic reasoning or problem-solving. Because of this, the results could be biased toward models that are good at quickly recalling information but less capable of performing multi-step reasoning. This limits the validity of SWAA as a measure of real-world AI agentic capability. If the rapid improvement seen in AI performance over time is primarily due to advances in this kind of short, non-agentic task, then the overall trend of exponential progress may be overstated. The assumption that knowledge recall directly translates into improved agency might not hold true, meaning that forecasts based on SWAA performance could overestimate how fast AI will improve on complex, context-dependent tasks like those evaluated by HCAST or RE-Bench.

The second critique relates to the exclusion of context. Real intellectual work requires context processing, such as professionals reminding themselves of relevant techniques or information before making a decision. For example, if a task implicitly involves port forwarding, a human must recall how that process works before taking any action. SWAA removes that aspect by measuring only the decision moment, so the time recorded for human baselines becomes artificially short. Suppose a human spends 10 seconds reading and understanding the prompt and another 5 seconds to respond. SWAA baselining would only count the decision time, labeling it as a 5-second task. This means that the measured task difficulty used to calculate the time horizon ends up being lower than it should be.

4.2 HCAST

The Human-Calibrated Autonomy Software Tasks (HCAST) is a collection of 189 tasks across machine learning, cybersecurity, software engineering, and general

reasoning domains (Rein et al., 2025). “HCAST tasks are designed to capture realistic challenges of widely varying complexity: they typically require multi-step sequential decision-making, are largely derived from real-world work, take between 1-2 minutes and 8+ hours, and undergo multiple stages of manual quality assurance and review.” (Rein et al., 2025)

The main limitation of HCAST is that it is text-based. The authors mention that “Because modern frontier AI systems are relatively more proficient at text-based tasks, the majority of these tasks do not require visual/multimodal capabilities, and all tasks are solvable by text editing via a bash shell.” (Rein et al., 2025) This means other modalities like images and sound are not taken into account. While this may be sufficient for evaluating AI capabilities on pure software and coding tasks, it is likely inadequate for many tasks in a real-world job. For example, many jobs require conversing with team members to plan and reflect on projects. Without taking into account these other modalities, we could overestimate the ability of AI systems to automate certain jobs.

Focusing solely on text-based evaluations could also be a problem because it does not provide information on an AI system’s ability to switch between different modes. Even if an AI is highly capable at more than one modality, it may not be able to combine and switch between those modalities well. Since many real-world jobs require switching constantly between different modalities, the benchmark may again overestimate an AI’s ability to perform well at a job.

This limitation is acknowledged as “Solitary software tasks” in the paper that details HCAST. The authors mention “All of our tasks involve interfacing with software tools or the internet via a containerized environment. They are also designed to be performed autonomously, and do not require interfacing with humans or other agents. The majority of tasks that humans complete are not software tasks that can be done without any interaction with other humans, and HCAST therefore may have limited relevance for evaluating many important skills for AI agents to have economic impact.” (Rein et al., 2025).

This suggests that in order to properly evaluate the capabilities of AI systems to automate real-world jobs, it is necessary to have benchmarks that incorporate the multi-modality of everyday jobs. These new benchmarks could be used to forecast AI time horizons, in a similar way to *Measuring AI Ability to Complete Long Tasks*, with the added benefit of incorporating multi-modality and modelling real-world jobs more accurately.

4.3 RE-Bench

RE-Bench, or Research Engineering Benchmark, includes seven difficult open-ended machine learning research tasks. It’s designed to test how well AI models can handle real research engineering problems which usually take a human machine learning engineer roughly 8 hours to complete. The threshold for success is based on the average score obtained by human baseliners who spent between 7 and 9 hours on the task. An example task is called “CUDA backtesting,” which has an estimated 8-hour duration. In this task, the model must speed up a Python tool

used for trade execution simulations by writing custom CUDA kernels that preserve the tool’s functionality.

One key critique of RE-Bench is that it loses valuable information by reducing task outcomes to a simple binary result, either success or failure. Because these tasks are so demanding, current top models often make significant partial progress without fully completing the task. Under the benchmark’s binary scoring scheme, these partial successes are still marked as failures. This can hide meaningful progress made by advanced models, with the methodology understating or not addressing overall capability growth.

Another critique is the way human baselines were created. During testing, the human participants had access to AI tools, which may have enhanced their performance. This creates a potential bias because the baseline being used to compare AI models is not purely human. As a result, the benchmark might be setting an unrealistically high standard, especially for long, iterative tasks that require sustained effort over many hours, which could lead to an underestimation of the resulting time horizon.

5. Conclusion

METR’s framework for forecasting AI time horizons is a valuable step toward grounding capability predictions in data. Still, its methods have some limitations that could influence how we interpret AI progress. SWAA focuses on quick knowledge recall, HCAST remains within text-based settings, and RE-Bench simplifies complex outcomes into broad categories. These factors may slightly affect how accurately the benchmarks reflect real-world performance. Looking ahead, incorporating multi-step reasoning, multimodal tasks, and other changes could strengthen future forecasting and give a more balanced view of how AI capabilities are developing.

6. References

- Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., ... Chan, L. (2025). *Measuring AI Ability to Complete Long Tasks* (No. arXiv:2503.14499). arXiv. <https://doi.org/10.48550/arXiv.2503.14499>
- Rein, D., Becker, J., Deng, A., Nix, S., Canal, C., O’Connel, D., Arnott, P., Bloom, R., Broadley, T., Garcia, K., Goodrich, B., Hasin, M., Jawhar, S., Kinniment, M., Kwa, T., Lajko, A., Rush, N., Sato, L. J. K., Arx, S. V., ...

Barnes, E. (2025). *HCAST: Human-Calibrated Autonomy Software Tasks*
(No. arXiv:2503.17354). arXiv.

<https://doi.org/10.48550/arXiv.2503.17354>

Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen,
M., Clymer, J., Dhyani, J., Elicheva, E., Garcia, K., Goodrich, B., Jurkovic,
N., Karnofsky, H., Kinniment, M., Lajko, A., Nix, S., Sato, L., ... Barnes,
E. (2025). *RE-Bench: Evaluating frontier AI R&D capabilities of language
model agents against human experts* (No. arXiv:2411.15114). arXiv.

<https://doi.org/10.48550/arXiv.2411.15114>