nature

NEWS | 19 March 2025

AI could soon tackle projects that take humans weeks

New metric assesses how AI is getting better at completing long tasks — but some researchers are wary of long-term predictions.

By Garrison Lovely

A new metric assesses the progress of AI models. Credit: Adapted from Jonathan Raa/NurPhoto/Getty

Today's artificial intelligence (AI) systems can't beat humans on long tasks, but they're improving at a rapid pace and could close the gap sooner than many anticipated, according to an analysis of leading models[1].

METR, a non-profit organization in Berkeley, California, created nearly 170 real-world tasks in coding, cybersecurity, general reasoning and machine learning, then established a 'human baseline' by measuring how long it took expert programmers to complete them.

The team then developed a metric for assessing the progress of AI models, which it calls 'task-completion time horizon'. This is the time programmers typically take to complete the tasks that AI models can complete at a certain success rate.

In a preprint posted on arXiv this week, METR reports that GPT-2, an early large language model (LLM) published by OpenAI in 2019, failed on all tasks that took human experts more than one minute. Claude 3.7 Sonnet, released in February by the US-based start-up Anthropic, completed 50% of the tasks that would take people 59 minutes.

Overall, the time horizon of the 13 leading AI models has doubled roughly every seven months since 2019, the paper finds. The exponential growth of AI time horizons accelerated in 2024, with the latest models doubling their horizon roughly every three months. The work has not been formally peer reviewed.

At the 2019-2024 rate of progress, METR suggests that AI models will be able to handle tasks that take humans about a month at 50% reliability by 2029, possibly sooner.

One month of dedicated human expertise, the paper notes, can be enough to start a new company or make scientific discoveries, for instance.

But Joshua Gans, a management professor at the University of Toronto in Canada, who has written on the economics of AI, says that these sorts of predictions aren't that useful. "Extrapolations are tempting to do, but there is still so much we don't know about how AI will actually be used for these to be meaningful, he says.

Human versus AI assessment

The team chose the 50% success rate because it was the most robust to small changes in the data distribution. "If you pick very low or very high thresholds, removing or adding a single successful or single failed task, respectively, changes your estimate a lot," says co-author Lawrence Chan.

Raising the reliability threshold from 50% to 80% reduced the average time horizon by a factor of five — although the overall doubling time and trendline were similar.

In the past five years, improvements in the general capabilities of LLMs have been driven largely by increases in scale — the amount of training data, training time and number of model parameters. The paper attributes progress on the time horizon metric mainly to improvements in AI's logical reasoning, tool use, error correction and self-awareness in task execution.

METR's time-horizon approach addresses some of the limitations in existing AI benchmarks, which map to real-world work only loosely and quickly 'saturate' as models improve. It provides a continuous, intuitive measure that better captures meaningful long-term progress, says co-author Ben West.

Leading AI models achieve superhuman performance on many benchmarks, but they have had relatively little economic impact, says West. METR's latest research offers a partial answer to this puzzle: the best models sit at around a 40-minute time horizon, and there isn't much economically valuable work that a person can do in that time, says West.

But Anton Troynikov, an AI researcher and entrepreneur in San Francisco, California, says that AI would have more economic impact if organizations were more willing to experiment and invest in leveraging the models effectively.

Limits to approach

Troynikov says that although the task-completion time horizon is a useful metric for assessing the economic utility of existing models, it might not reveal how well models can 'generalize', by performing tasks that differ what they were trained on.

METR acknowledges that its approach doesn't capture all of the complexity of real work, but says that it found a similar exponential trend in time-horizon growth when checking how well the tasks resemble real-life work.

The authors say that there are some factors that could affect their prediction of when a one-month time horizon will be achieved. Computing power has increased significantly in the past five years, but physical and economic factors will limit future scale-ups, which will probably hamper AI progress. But that will be partly offset by continued algorithmic improvements, researchers say. METR also expects that efforts to give models more agency and make them more effective at automating AI research will continue to bear fruit.

Gans says the next step is studies that pair AI systems with humans and examine how well those pairings improve overall task performance.

doi: https://doi.org/10.1038/d41586-025-00831-8

References

1. Kwa, T. *et al.* Preprint at arXiv https://doi.org/10.48550/arXiv.2503.14499 (2025).

Reprints and permissions

Latest on:

Machine learning Computer science

How the US tech industry is shaping the transition to green energy

NATURE INDEX | 19 MAR 25

Global cooperation is crucial for DeepSeek and broader AI research

CORRESPONDENCE |

18 MAR 25

AI demands a different approach to education

CORRESPONDENCE |

18 MAR 25

Nature (*Nature*) | ISSN 1476-4687 (online) | ISSN 0028-0836 (print)