ARTIFICIAL INTELLIGENCE

The way we measure progress in AI is terrible

Many of the most popular benchmarks for AI models are outdated or poorly designed.

By Scott J Mulligan

November 26, 2024

SARAH ROGERS/MITTR | PHOTOS GETTY

Every time a new AI model is released, it's typically touted as acing its performance against a series of benchmarks. OpenAI's GPT-4o, for example, was launched in May with a compilation of results that showed its performance topping every other AI company's latest model in several tests.

POPULAR

People are using Google study software to make AI podcasts—and they're weird and

The problem is that these benchmarks are poorly designed, the results hard to replicate, and the metrics they use are frequently arbitrary, according to new research. That matters because AI models' scores against these benchmarks will determine the level of scrutiny and regulation they receive.

"It seems to be like the Wild West because we don't really have good evaluation standards," says Anka Reuel, an author of the paper, who is a PhD student in computer science at Stanford University and a member of its Center for AI Safety.

A benchmark is essentially a test that an AI takes. It can be in a multiple-choice format like the most popular one, the Massive Multitask Language Understanding benchmark, known as the MMLU, or it could be an evaluation of AI's ability to do a specific task or the quality of its text responses to a set series of questions.

AI companies frequently cite benchmarks as testament to a new model's success. "The developers of these models tend to optimize for the specific benchmarks," says Anna Ivanova, professor of psychology at the Georgia Institute of Technology and head of its Language, Intelligence, and Thought (LIT) lab, who was not involved in the Stanford research.

These benchmarks already form part of some governments' plans for regulating AI. For example, the EU AI Act, which goes into force in August 2025, references benchmarks as a tool to determine whether or not a model demonstrates "systemic risk"; if it does, it will be subject to higher levels of scrutiny and regulation. The UK AI Safety Institute references benchmarks in Inspect, which is its framework for evaluating the safety of large language models.

But right now, they might not be good enough to use that way. "There's this potential false sense of safety we're creating with benchmarks if they aren't well designed, especially for high-stakes use cases," says Reuel. "It may look as if the model is safe, but it is not."

Given the increasing importance of benchmarks, Reuel and her colleagues wanted to look at the most popular examples to figure out what makes a good one—and whether the ones we use are robust enough. The researchers first set out to verify the benchmark results

amazing

Melissa Heikkilä

This AI-generated version of Minecraft may represent the future of real-time video generation

Scott J Mulligan

Why AI could eat quantum computing's lunch

Edd Gent

AI can now create a replica of your personality

James O'Donnell

that developers put out, but often they couldn't reproduce them. To test a benchmark, you typically need some instructions or code to run it on a model. Many benchmark creators didn't make the code to run their benchmark publicly available. In other cases, the code was outdated.

Related Story

Google DeepMind's new AI system can solve complex geometry problems

Its performance matches the smartest high school mathematicians and is much stronger than the previous state-of-the-art system.

Benchmark creators often don't make the questions and answers in their data set publicly available either. If they did, companies could just train their model on the benchmark; it would be like letting a student see the questions and answers on a test before taking it. But that makes them hard to evaluate.

Another issue is that benchmarks are frequently "saturated," which means all the problems have been pretty much been solved. For example, let's say there's a test with simple math problems on it. The first generation of an AI model gets a 20% on the test, failing. The second generation of the model gets 90% and the third generation gets 93%. An outsider may look at these results and determine that AI progress has slowed down, but another interpretation could just be that the benchmark got solved and is no longer that great a measure of progress. It fails to capture the difference in ability between the second and third generations of a model.

One of the goals of the research was to define a list of criteria that make a good benchmark. "It's definitely an important problem to discuss the quality of the benchmarks, what we want from them, what we need from them," says Ivanova. "The issue is that there isn't one good standard to define benchmarks. This paper is an attempt to provide a set of evaluation criteria. That's very useful."

The paper was accompanied by the launch of a website, Better Bench, that ranks the most popular AI benchmarks. Rating factors include whether or not experts were consulted on the design, whether the tested capability is well defined, and other basics—for example, is there a feedback channel for the benchmark, or has it been peer-reviewed?

The MMLU benchmark had the lowest ratings. "I disagree with these rankings. In fact, I'm an author of some of the papers ranked highly, and would say that the lower ranked

benchmarks are better than them," says Dan Hendrycks, director of CAIS, the Center for AI Safety, and one of the creators of the MMLU benchmark. That said, Hendrycks still believes that the best way to move the field forward is to build better benchmarks.

Some think the criteria may be missing the bigger picture. "The paper adds something valuable. Implementation criteria and documentation criteria—all of this is important. It makes the benchmarks better," says Marius Hobbhahn, CEO of Apollo Research, a research organization specializing in AI evaluations. "But for me, the most important question is, do you measure the right thing? You could check all of these boxes, but you could still have a terrible benchmark because it just doesn't measure the right thing."

Essentially, even if a benchmark is perfectly designed, one that tests the model's ability to provide compelling analysis of Shakespeare sonnets may be useless if someone is really concerned about AI's hacking capabilities.

"You'll see a benchmark that's supposed to measure moral reasoning. But what that means isn't necessarily defined very well. Are people who are experts in that domain being incorporated in the process? Often that isn't the case," says Amelia Hardy, another author of the paper and an AI researcher at Stanford University.

There are organizations actively trying to improve the situation. For example, a new benchmark from Epoch AI, a research organization, was designed with input from 60 mathematicians and verified as challenging by two winners of the Fields Medal, which is the most prestigious award in mathematics. The participation of these experts fulfills one of the criteria in the Better Bench assessment. The current most advanced models are able to answer less than 2% of the questions on the benchmark, which means there's a significant way to go before it is saturated.

"We really tried to represent the full breadth and depth of modern math research," says Tamay Besiroglu, associate director at Epoch AI. Despite the difficulty of the test, Besiroglu speculates it will take only around four or five years for AI models to score well against it.
And Hendrycks' organization, CAIS, is collaborating with Scale AI to create a new benchmark that he claims will test AI models against the frontier of human knowledge,

dubbed Humanity's Last Exam, HLE. "HLE was developed by a global team of academics and subject-matter experts," says Hendrycks. "HLE contains unambiguous, non-searchable, questions that require a PhD-level understanding to solve." If you want to contribute a question, you can here.

Although there is a lot of disagreement over what exactly should be measured, many researchers agree that more robust benchmarks are needed, especially since they set a direction for companies and are a critical tool for governments.

"Benchmarks need to be really good," Hardy says. "We need to have an understanding of what 'really good' means, which we don't right now."

by Scott J Mulligan

DEEP DIVE

ARTIFICIAL INTELLIGENCE

This AI-generated version of Minecraft may represent the future of real-time video generation

The game was created from clips and keyboard inputs alone, as a demo for real-

People are using Google study software to make AI podcasts—and they're weird and amazing

NotebookLM is a surprise hit. Here are some of the ways people are using it.

By Melissa Heikkilä

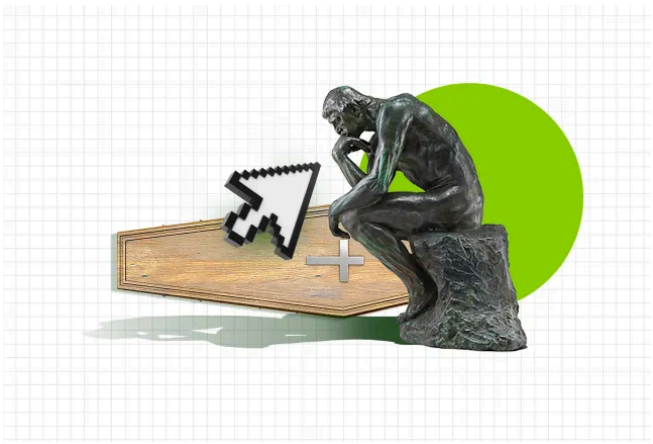
time interactive video generation.

By Scott J Mulligan

AI can now create a replica of your personality

A two-hour interview is enough to accurately capture your values and preferences, according to new research from Stanford and Google DeepMind.

By James O'Donnell



Introducing: The AI Hype Index

Everything you need to know about the state of AI.

By The Editors

Privacy Policy

The latest iteration of a legacy

Founded at the Massachusetts Institute of Technology in 1899, MIT Technology Review is a world-renowned, independent media company whose insight, analysis, reviews, interviews and live events explain the newest technologies and their commercial, social and political impact.

READ ABOUT OUR HISTORY

Advertise with MIT Technology Review

Elevate your brand to the forefront of conversation around emerging technologies that are radically transforming business. From event sponsorships to custom content to visually arresting video storytelling, advertising with MIT Technology Review creates opportunities for your brand to resonate with an unmatched audience of technology and business elite.

ADVERTISE WITH US