A.I. Has a Measurement Problem

Which A.I. system writes the best computer code or generates the most realistic image? Right now, there's no easy way to answer those questions.

By Kevin Roose

There's a problem with leading artificial intelligence tools like ChatGPT, Gemini and Claude: We don't really know how smart they are.

That's because, unlike companies that make cars or drugs or baby formula, A.I. companies aren't required to submit their products for testing before releasing them to the public. There's no Good Housekeeping seal for A.I. chatbots, and few independent groups are putting these tools through their paces in a rigorous way.

Instead, we're left to rely on the claims of A.I. companies, which often use vague, fuzzy phrases like "improved capabilities" to describe how their models differ from one version to the next. And while there are some standard tests given to A.I. models to assess how good they are at, say, math or logical reasoning, many experts have doubts about how reliable those tests really are.

This might sound like a petty gripe. But I've become convinced that a lack of good measurement and evaluation for A.I. systems is a major problem.

For starters, without reliable information about A.I. products, how are people supposed to know what to do with them?

I can't count the number of times I've been asked in the past year, by a friend or a colleague, which A.I. tool they should use for a certain task. Does ChatGPT or Gemini write better Python code? Is DALL-E 3 or Midjourney better at generating realistic images of people?

I usually just shrug in response. Even as someone who writes about A.I. for a living and tests new tools constantly, I've found it maddeningly hard to keep track of the relative strengths and weaknesses of various A.I. products. Most tech companies don't publish user manuals or detailed release notes for their A.I. products. And the models are updated so frequently that a chatbot that struggles with a task one day might mysteriously excel at it the next.

Shoddy measurement also creates a safety risk. Without better tests for A.I. models, it's hard to know which capabilities are improving faster than expected, or which products might pose real threats of harm.

In this year's A.I. Index — a big annual report put out by Stanford University's Institute for Human-Centered Artificial Intelligence — the authors describe poor measurement as one of the biggest challenges facing A.I. researchers.

"The lack of standardized evaluation makes it extremely challenging to systematically compare the limitations and risks of various A.I. models," the report's editor in chief, Nestor Maslej, told me.

For years, the most popular method for measuring artificial intelligence was the so-called Turing Test — an exercise proposed in 1950 by the mathematician Alan Turing, which tests whether a computer program can fool a person into mistaking its responses for a human's.

But today's A.I. systems can pass the Turing Test with flying colors, and researchers have had to come up with new, harder evaluations.

One of the most common tests given to A.I. models today — the SAT for chatbots, essentially — is a test known as Massive Multitask Language Understanding, or MMLU.

The MMLU, which was released in 2020, consists of a collection of roughly 16,000 multiple-choice questions covering dozens of academic subjects, ranging from abstract algebra to law and medicine. It's supposed to be a kind of general intelligence test — the more of these questions a chatbot answers correctly, the smarter it is.

It has become the gold standard for A.I. companies competing for dominance. (When Google released its most advanced A.I. model, Gemini Ultra, earlier this year, it boasted that it had scored 90 percent on the MMLU — the highest score ever recorded.)

Dan Hendrycks, an A.I. safety researcher who helped develop the MMLU while in graduate school at the University of California, Berkeley, told me that the test was never supposed to be used for bragging rights. He was alarmed by how quickly A.I. systems were improving, and wanted to encourage researchers to take it more seriously.

Mr. Hendrycks said that while he thought MMLU "probably has another year or two of shelf life," it will soon need to be replaced by different, harder tests. A.I. systems are getting too smart for the tests we have now, and it's getting more difficult to design new ones.

"All of these benchmarks are wrong, but some are useful," he said. "Some of them can serve some utility for a fixed amount of time, but at some point, there's so much pressure put on it that it reaches its breaking point."

There are dozens of other tests out there — with names like TruthfulQA and HellaSwag — that are meant to capture other facets of A.I. performance. But just as the SAT captures only part of a student's intellect and ability, these tests are capable of measuring only a narrow slice of an A.I. system's power.

And none of them are designed to answer the more subjective questions many users have, such as: Is this chatbot fun to talk to? Is it better for automating routine office work, or creative brainstorming? How strict are its safety guardrails?

(The New York Times has sued OpenAI, the maker of ChatGPT, and its partner, Microsoft, on claims of copyright infringement involving artificial intelligence systems that generate text.)

There may also be problems with the tests themselves. Several researchers I spoke to warned that the process for administering benchmark tests like MMLU varies slightly from company to company, and that various models' scores might not be directly comparable.

There is a problem known as "data contamination," when the questions and answers for benchmark tests are included in an A.I. model's training data, essentially allowing it to cheat. And there is no independent testing or auditing process for these models, meaning that A.I. companies are essentially grading their own homework.

In short, A.I. measurement is a mess — a tangle of sloppy tests, apples-to-oranges comparisons and self-serving hype that has left users, regulators and A.I. developers themselves grasping in the dark.

"Despite the appearance of science, most developers really judge models based on vibes or instinct," said Nathan Benaich, an A.I. investor with Air Street Capital. "That might be fine for the moment, but as these models grow in power and social relevance, it won't suffice."

The solution here is likely a combination of public and private efforts.

Governments can, and should, come up with robust testing programs that measure both the raw capabilities and the safety risks of A.I. models, and they should fund grants and research projects aimed at coming up with new, high-quality evaluations. (In its executive order on A.I. last year, the White House directed several federal agencies, including the National Institute of Standards and Technology, to create and oversee new ways of evaluating A.I. systems.)

Some progress is also emerging out of academia. Last year, Stanford researchers introduced a new test for A.I. image models that uses human evaluators, rather than automated tests, to determine how capable a model is. And a group of researchers from the University of California, Berkeley, recently started Chatbot Arena, a popular leaderboard that pits anonymous, randomized A.I. models against one another and asks users to vote on the best model.

A.I. companies can also help by committing to work with third-party evaluators and auditors to test their models, by making new models more widely available to researchers and by being more transparent when their models are updated. And in the media, I hope some kind of Wirecutter-style publication will eventually emerge to take on the task of reviewing new A.I. products in a rigorous and trustworthy way.

Researchers at Anthropic, the A.I. company, wrote in a blog post last year that “effective A.I. governance depends on our ability to meaningfully evaluate A.I. systems.”

I agree. Artificial intelligence is too important a technology to be evaluated on the basis of vibes. Until we get better ways of measuring these tools, we won’t know how to use them, or whether their progress should be celebrated or feared.

Source: The New York Times