Is There an AI Metrics Mirage?

All-or-nothing measures create the illusion that the capabilities of large language models have grown in leaps rather than incremental steps.

Over the past three years, artificial intelligence models have shown surprising jumps in performance, baffling many scientists in the field. A new paper by Stanford University's Rylan Schaeffer argues that many of the so-called emergent capabilities researchers have observed in the biggest large language models (LLMs), from performing mathematical calculations to solving logic puzzles, are nothing more than "mirages": statistical artifacts of the benchmarks used to measure them. This claim challenges the widely held belief in machine learning circles that LLMs are capable of much more than their training data would suggest and that they are prone to unpredictable, spontaneous changes in behavior.

LLMs are mathematical models that map the statistical relationships between words in giant databases of text and produce novel text by predicting the next likely word in a prompt sentence. In 2023, LLMs went from being a technical curiosity to the hottest thing in Silicon Valley, mainly due to the success of OpenAI's publicly accessible chatbot ChatGPT. As members of the public played with the new chatbot, scientists probed the models in a more robust way to see what they were capable of. It seemed as though every week there were new reports of astonishing new behavior being observed in an LLM.

Schaeffer presented his response to these discoveries at the Annual Conference on Neural Information Processing Systems in New Orleans in December 2023. "Our paper is a story about predictability and surprise," he began. He showed graph after graph from papers published over the past three years showing flat lines suddenly jumping up and to the right as new capabilities were observed in LLMs as they got bigger, such as the ability to add three-digit numbers or solve a logic puzzle. Schaeffer then drew the crowd's attention not to the capabilities, but to the metrics used to measure them. Most benchmarks assign one point for a correct answer and zero for anything else.

"All of these metrics are quite harsh," he said. "They give no partial credit." In examining Google's BIG-Bench benchmark, a series of tasks used to test and compare the performance of different LLMs, Schaeffer found that "over 90 percent of emergent abilities are observed under two metrics." Those two metrics (multiple-choice grade and exact string match) have an important limitation. "Either you exactly output the correct answer, or you do not," Schaeffer said. "There is no in-between."

When the metric is changed so that one point is given above a certain threshold and zero are given below, it appears that the model suddenly learns how to perform the task flawlessly.

This binary measure means that even if a model makes incremental progress at, say, adding five-digit numbers as the model gets bigger, that progress won't be visible to the researcher. Instead, at a certain size, the model's progress suddenly jumps "to the ceiling" and researchers are left wondering where it came from. "In many cases," Schaeffer said, "emergent abilities might be mirages that are produced by the researchers' analyses. They're not actually due to fundamental changes in the language model."

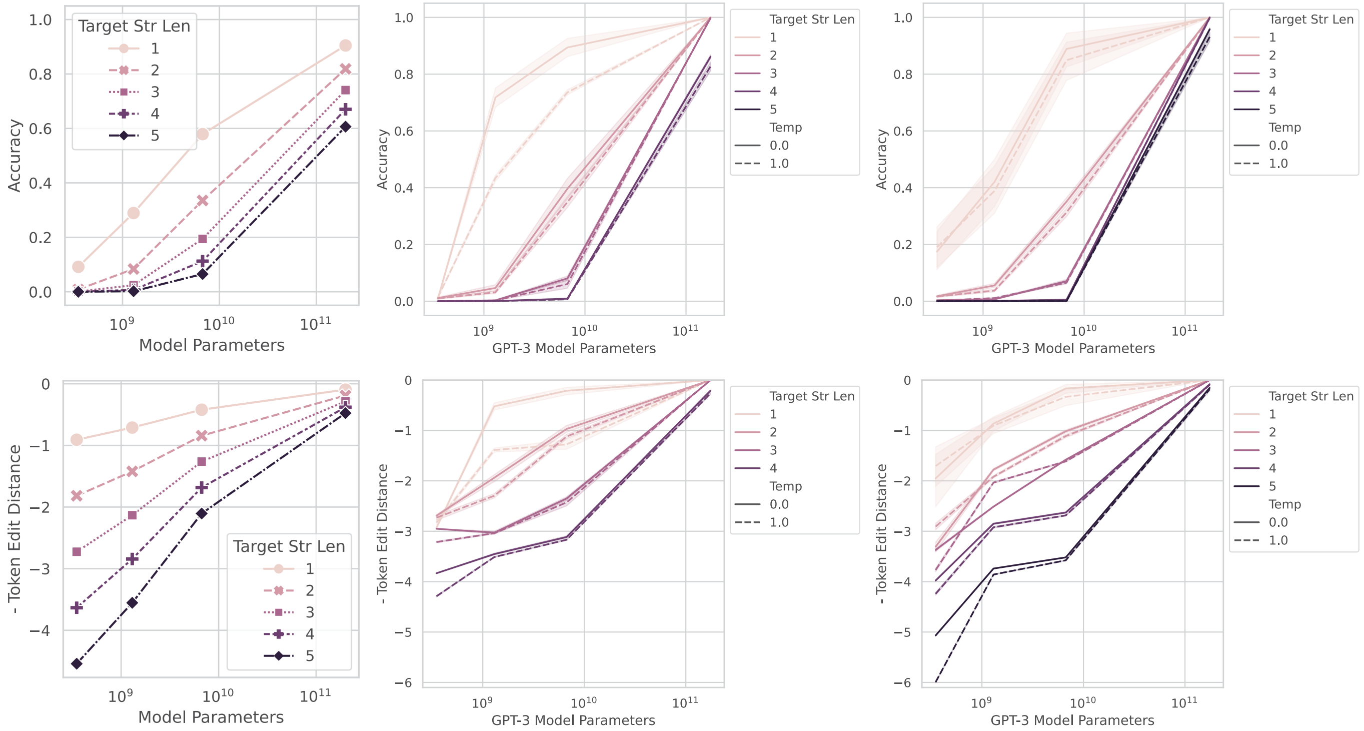
When researchers claim that the capabilities of their models are unpredictable, it might be because they're not using the right metrics to measure them. Schaeffer and his Stanford colleagues Brando Miranda and Sanmi Koyejo reanalyzed some of the data from past papers and used linear metrics that gave partial marks for partially correct answers. This modification caused the exponential jumps in performance to disappear in favor of nice, predictable linear functions.

Schaeffer and his collaborators went even further. They succeeded in "inducing" the appearance of an emergent ability in a simple neural network by changing the metrics used to measure its accuracy when recreating images from a given collection. "Everybody has done this in their intro to machine learning class," he said, to emphasize how simple the model was, "but we can qualitatively produce what looks like emergent behavior."

Usually, students notice that as the vision model gets bigger, it gets steadily better at copying images when they measure, statistically, how closely it resembles the original image. But when the metric is changed so that one point is given for images above a certain threshold and zero points are given for those below, it appears that the model suddenly learns how to perform the task flawlessly. "We didn't know of any prior work that had found emergent abilities in vision tasks," he said, "so to induce them intentionally was quite novel."

I spoke with Schaeffer after the presentation and asked why choosing the right metrics was so important. "It's important for computer scientists out of a scientific interest," he said, "but it's also important for government policy people and for economists, to know how much better models are getting." If claims of "magical" emergent properties can be tempered with a more accurate picture of the incremental improvements models make as they get larger, better decisions can be made about where and when to implement the AI systems.

But Schaeffer is also concerned that the entire field is moving too fast to



R. Schaeffer et al. 2023. arxiv.org/abs/2304.15004

Changing the metrics used to measure the accuracy of a large language model (LLM) shows that the system's abilities grow incrementally rather than in big jumps. The top three charts plot the accuracy of OpenAI's GPT-3 at mathematical tasks. Accuracy is a pass-fail metric, and the LLM's performance seems to improve suddenly and unpredictably as the model gets bigger. The bottom three charts trace GPT-3's performance on the same tasks using *token edit distance*, which measures how close the LLM's response is to the correct answer. When improvement rather than perfection is measured, GPT-3's performance grows more smoothly and predictably.

catch statistical errors like these. "How do people build independent confidence in what these models are doing?" he asks. Traditionally the answer has been through a process of transparency and peer review, but now researchers are so eager to share their findings that they often post their results online before they have been reviewed or even accepted to a journal. The website arXiv.org, founded as a place to share preprints of journal papers, is now mostly populated by papers that will never appear in a peer-reviewed journal. "What is then the role of science?" Schaeffer asks. "How can we contribute scientifically?"

If claims of "magical" emergent properties can be tempered with a more accurate picture of incremental improvements, better decisions can be made about where and when to implement the AI systems.

To make matters even more complicated, many of the most powerful LLMs are proprietary, owned by the largest tech companies in the world. These same companies have a virtual monopoly on the enormous computing power needed to train LLMs. "Scientific progress can be hampered," Schaeffer, Miranda, and Koyejo note in the conclusion to their conference paper, "when models and their outputs are not made available for independent scientific investigation."

At the conference in New Orleans, many of my conversations with other attendees drifted back to Schaeffer's talk from earlier in the day. Some seemed excited by the simplicity of the argument and saw it as a reminder to be vigilant against sloppy science. Others seemed a bit deflated, as if the mirage metaphor was meant to describe the whole field of artificial intelligence.

"It's a cautionary tale for researchers who want to stay grounded," says Ofer Shai, senior director of deep learning at Untether AI. "LLMs are getting better and better, but as the paper shows, it's gradual improvement. This paper goes toward correcting some of the hype." Shai says the paper highlights the importance of using relevant metrics, and also serves as a reminder to be careful not to see phase transitions where they don't exist. "Benchmarks continue to evolve, we get new datasets, and that's a good thing," he says.

Schaeffer himself, a voice of calm in a sea of often bombastic claims, still leaves room for some magic, however. "Nothing in this paper should be interpreted as claiming that large language models cannot display emergent abilities," he writes. Whether higher-order capabilities such as reasoning, theory of mind, or even consciousness itself could emerge from an LLM is still an open question. —*Joseph Wilson*

Joseph Wilson is a doctoral candidate in linguistic and semiotic anthropology at the University of Toronto. His work examines how scientists use metaphor and other figurative language to communicate with one another in laboratory settings. Email: joseph.wilson@utoronto.ca