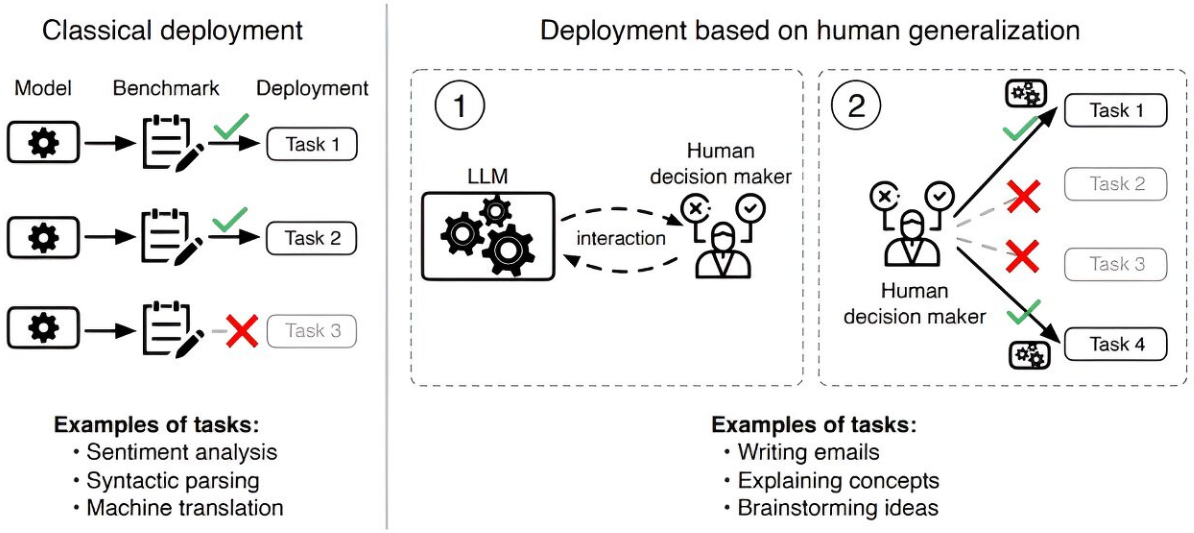
Tech Xplore

Large language models don't behave like people, even though we may expect them to

July 23 2024, by Adam Zewe



Classically, ML models are deployed to perform tasks based on benchmark performance (left). When deployment is based on human generalization (right), a human decision maker first interacts with a model to assess its capabilities, and then the model is deployed to perform tasks the decision maker believes it will perform well on. The model's deployed performance depends on how well aligned its capabilities are with the human generalization function. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2406.01382

One thing that makes large language models (LLMs) so powerful is the diversity of tasks to which they can be applied. The same machine-learning model that can help a graduate student draft an email could also aid a clinician in diagnosing cancer.

However, the wide applicability of these models also makes it challenging to evaluate them in a systematic way. It would be impossible to create a benchmark dataset to test a model on every type of question it can be asked.

In a new paper posted to the *arXiv* preprint server, MIT researchers took a different approach. They argue that, because humans decide when to deploy large language models, evaluating a model requires an understanding of how people form beliefs about its capabilities.

For example, the graduate student must decide whether the model could be helpful in drafting a particular email, and the clinician must determine which cases would be best to consult the model on.

Building off this idea, the researchers created a framework to evaluate an LLM based on its alignment with a human's beliefs about how it will perform on a certain task.

They introduce a human generalization function—a model of how people update their beliefs about an LLM's capabilities after interacting with it. Then, they evaluate how aligned LLMs are with this human generalization function.

Their results indicate that when models are misaligned with the human generalization function, a user could be overconfident or underconfident about where to deploy them, which might cause a model to fail unexpectedly. Furthermore, due to this misalignment, more capable models tend to perform worse than smaller models in high-stakes situations.

"These tools are exciting because they are general purpose, but because they are general purpose, they will be collaborating with people, so we have to take the human in the loop into account," says study co-author Ashesh Rambachan, assistant professor of economics and a principal investigator in the Laboratory for Information and Decision Systems (LIDS).

Rambachan is joined on the paper by lead author Keyon Vafa, a postdoc at Harvard University; and Sendhil Mullainathan, an MIT professor in the departments of Electrical Engineering and Computer Science and of Economics, and a member of LIDS. The research will be presented at the International Conference on Machine Learning (ICML 2024) held in Vienna, Austria, July 21–27.

Human generalization

As we interact with other people, we form beliefs about what we think they do and do not know. For instance, if your friend is finicky about correcting people's grammar, you might generalize and think they would also excel at sentence construction, even though you've never asked them questions about sentence construction.

"Language models often seem so human. We wanted to illustrate that this force of human generalization is also present in how people form beliefs about language models," Rambachan says.

As a starting point, the researchers formally defined the human generalization function, which involves asking questions, observing how a person or LLM responds, and then making inferences about how that person or model would respond to related questions.

If someone sees that an LLM can correctly answer questions about matrix inversion, they might also assume it can ace questions about simple arithmetic. A model that is misaligned with this function—one that doesn't perform well on questions a human expects it to answer correctly—could fail when deployed.

With that formal definition in hand, the researchers designed a survey to measure how people generalize when they interact with LLMs and other people.

They showed survey participants questions that a person or LLM got right or wrong and then asked if they thought that person or LLM would answer a related question correctly. Through the survey, they generated a dataset of nearly 19,000 examples of how humans generalize about LLM performance across 79 diverse tasks.

Measuring misalignment

They found that participants did quite well when asked whether a human who got one question right would answer a related question right, but they were much worse at generalizing about the performance of LLMs.

"Human generalization gets applied to language models, but that breaks down because these language models don't actually show patterns of expertise like people would," Rambachan says.

People were also more likely to update their beliefs about an LLM when it answered questions incorrectly than when it got questions right. They also tended to believe that LLM performance on simple questions would have little bearing on its performance on more complex questions.

In situations where people put more weight on incorrect responses, simpler models outperformed very large models like GPT-4.

"Language models that get better can almost trick people into thinking they will perform well on related questions when, in actuality, they don't," he says.

One possible explanation for why humans are worse at generalizing for LLMs could come from their novelty—people have far less experience interacting with LLMs than with other people.

"Moving forward, it is possible that we may get better just by virtue of interacting with language models more," he says.

To this end, the researchers want to conduct additional studies of how people's beliefs about LLMs evolve over time as they interact with a model. They also want to explore how human generalization could be incorporated into the development of LLMs.

"When we are training these algorithms in the first place, or trying to update them with human feedback, we need to account for the human generalization function in how we think about measuring performance," he says.

In the meantime, the researchers hope their dataset could be used as a benchmark to compare how LLMs perform related to the human generalization function, which could help improve the performance of models deployed in real-world situations.

"To me, the contribution of the paper is twofold. The first is practical: The paper uncovers a critical issue with deploying LLMs for general consumer use. If people don't have the right understanding of when LLMs will be accurate and when they will fail, then they will be more likely to see mistakes and perhaps be discouraged from further use.

"This highlights the issue of aligning the models with people's understanding of generalization," says Alex Imas, professor of behavioral science and economics at the University of Chicago's Booth

School of Business, who was not involved with this work.

"The second contribution is more fundamental: The lack of generalization to expected problems and domains helps in getting a better picture of what the models are doing when they get a problem 'correct.' It provides a test of whether LLMs 'understand' the problem they are solving."

More information: Keyon Vafa et al, Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function, *arXiv* (2024). DOI: 10.48550/arxiv.2406.01382

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Large language models don't behave like people, even though we may expect them to (2024, July 23) retrieved 23 July 2024 from https://techxplore.com/news/2024-07-large-language-dont-people.html