Why student experiments with generative AI matter for our collective learning

November 22 2023, by Mohammad Keyhani, Hadi Hemmati and Leslie Salgado Arzuaga

Credit: CC0 Public Domain

Generative artificial intelligence (GenAI) tools like ChatGPT based on Large Language Models (LLMs) are revolutionizing the ways we think,

learn and work.

But, like some other forms of AI, GenAI technologies have a black box nature—meaning it's hard to explain and understand how mathematical models compute their output.

If we as a society are to employ this new technology on a broad scale, we will need to engage in a collective discovery process to better understand how it works and what it is capable of.

As AI experts work on making AI systems more comprehensible to end users, and as OpenAI, the maker of ChatGPT, navigates leadership shakeups and questions about its strategic direction, post-secondary institutions have a critical role to play in enabling collective learning about GenAI.

Difficult to understand

For AI systems based on large neural networks with a black box nature, like GenAI, a lack of transparency makes it difficult for people to trust the AI and to rely on it for sensitive applications.

Carnegie Mellon University professor Elizabeth A. Holm has argued that black box AIs can still be valuable if they produce better results than alternatives, if the cost of wrong answers is low or if they inspire new ideas.

Still, cases of matters gone horribly wrong erode trust, such as when ChatGPT got tricked into giving instructions to make a bomb, or when it accused a law professor of a serious crime he didn't commit.

This is why researchers working on AI explainability have tried to devise techniques to see into the black box of neural networks. However, the

LLMs behind many GenAI tools are just too large and too complex for these methods to work.

Fortunately, LLMs like ChatGPT have an interesting feature that previous black box neural networks did not have: they are interactive. Think of it this way: we cannot understand what a person is thinking by looking at a map of the neurons in their brain, but we can talk to them.

'Machine psychology'

A new field of science is emerging under the label of "machine psychology" to understand how LLMs actually "think."

New research, yet to be peer reviewed, is examining how these models can surprise us with their emergent capabilities. For example, researchers surmised that because with LLMs every new word generated depends on the sequence of words that came before it, asking an LLM to work through a problem step by step may produce better results.

New studies, not yet peer reviewed, on this "chain of thought" technique and variations of it have shown they improve outcomes. Others suggest LLMs can be "emotionally manipulated" by including phrases like "are you sure?" or "believe in your abilities" in prompting.

In an interesting combination of these two methods, Google DeepMind researchers recently found that for a series of math problems, one LLM improved its accuracy significantly when it was prompted with "take a deep breath and work on this problem step-by-step."

Collective discovery

Understanding GenAI is not something only researchers are doing, and

that's a good thing. New discoveries that users have made have surprised even the makers of those tools, in both delightful and alarming ways.

Users are sharing their discoveries and prompts in online communities such as Reddit, Discord and dedicated platforms such as FlowGPT.

These prompts often include "jailbreak" prompts that succeed in getting GenAI tools to behave in ways they are not supposed to. People can trick AI into bypassing built-in rules—for example, producing hateful content—or creating malware.

These rapid advances and surprising outcomes are why some AI leaders called for a six-month moratorium on AI development earlier this year.

AI and learning

In higher education, an overly defensive approach emphasizing flaws and weaknesses of GenAI or how it allows students to cheat is ill-advised.

On the contrary, as workplaces start to see the benefits of GenAI-powered employees or workplace productivity, they will expect higher education to prepare students. Students' education needs to be relevant.

Universities are ideal spaces to forge co-operation across research fields, an imperative of developing responsible AI. Universities, in contrast to the private sector, are best positioned to embed their GenAI practices and content within a framework of ethical and responsible practice.

One thing this entails is understanding of GenAI as an augment, not a substitute, for human judgment and discerning when relying on this is permissable and acceptable.

Educating for GenAI involves developing critical thinking and fact-checking skills, and ethical prompt engineering. It also involves understanding that GenAI tools do not just repeat their training data, and can generate new, and high-quality ideas based on patterns in that data.

The ChatGPT and AI for Higher Education UNESCO Quick Guide is a helpful starting point.

Including GenAI in the curriculum cannot be treated as top-down teaching. Given the rapid development and newness of the technology, many students are already ahead of the professors in their GenAI knowledge and skills. We must recognize this as an era of collective discovery, where we are all learning from each other.

In the "Generative AI and Prompting" course offered at the Haskayne School of Business, University of Calgary, a portion of grades are allocated to posting, commenting and voting on an online "discovery forum" to share their discoveries and experiments.

Learning by doing and experimenting

Lastly, we should be learning how to use GenAI for tackling humanity's greatest challenges, such as climate change, poverty, disease, international conflict and systemic injustice.

Given the powerful nature of this technology, and the fact that we do not fully understand it due to its black box nature, we should do what we can to understand it through interaction and learning by doing and experimenting.

This is not an effort that can be confined to the works of specialized researchers or AI companies. It requires broad participation.

Provided by The Conversation