Q&A: Are large language models the modern-day Magic 8 Ball?

July 17 2023, by Megan Mastrola

Credit: Pixabay/CC0 Public Domain

he popular press and consumers are going too easy on ChatGPT, says Johns Hopkins University cybersecurity and artificial intelligence expert Anton Dahbura. According to him, the unreliability of such large language models, or LLMs, and their production of fabricated and biased content pose a real and growing threat to society.

"Industry is making lots of money off products that they know are being used in wrong ways, on a huge scale," says Dahbura, director of Johns Hopkins' Information Security Institute and co-director of the Institute for Assured Autonomy. "ChatGPT's 'hallucinations' [the system's tendency to sometimes generate senseless content] would be called 'failures' if they occurred in other consumer products—for example, if my car's accelerator had a 'hallucination' and ran me into a brick wall."

Better data, corporate accountability, and educating users about what AI is and what its limits are can help mitigate risk, Dahbura says, "but they will never make the problem go away completely unless the problem is so simple that AI shouldn't have been used to solve it in the first place."

The Hub sat down with Dahbura to discuss the reasons for uncertainty in large language models, the role he believes industry and government should play in educating consumers about AI and its risks, and the threats these new technologies might pose to society.

You've referred to ChatGPT and other large language models as the 'modern-day version of the Magic 8 Ball.' Explain.

Artificial intelligence is a broad class of approaches to solving difficult problems that don't have easy or "rule-based" solutions. A thermostat is an answer to a simple problem: When the temperature rises above a certain threshold, it turns on the air conditioning, and when it goes below that threshold, it turns on the heat.

But sometimes questions don't have clear answers that simple rules alone can solve. For instance, when training AI to differentiate between images of dogs and cats, the factors that the AI system uses for its classification are extremely complex and rarely well understood. Therefore, it is difficult to be able to place guarantees on how the system will respond to an image of a dog or cat that it hasn't been trained on, much less an image of an orange. It may not even respond predictably to an image that it's been trained on!

I've coined this inherent and inextricable property of AI systems as the "AI uncertainty principle" because the complexity of AI problems means that certainty and AI cannot coexist unless the solution is so simple that it doesn't require AI, or rule-based guardrails that are built to temper the unpredictable nature of the AI system.

What I am saying is that it is not possible to train these technologies on every single scenario, so you cannot accurately predict the outcome of using it every single time. It's the same with the Magic 8 Ball: The answer might not be what you expect to get.

You call companies irresponsible for failing to warn people about LLMs' potential to 'hallucinate.' Could you share an example of what you mean by a hallucination?

Hallucinations refer to responses that contain information that is incorrect or imaginary. There is a range of hallucinations that might impact technologies, including some that are relatively innocuous, like an LLM telling a group of elementary students that there are 13 planets in the solar system. This information is easy to check using a quick Google search, but other more obscure details or statements generated by ChatGPT might have more detrimental impacts. One example is using AI to detect cancer based on X-rays or scans. A hallucination might detect cancer that does not exist, or it might simply not detect cancer that does appear on the scan. Either of these scenarios could obviously be harmful.

Whether you're asking an LLM a question or using it to interpret an image, or it's built into a car in which you are hurtling down the highway at 65 miles per hour, you're presenting it with inputs and it's generating outputs. Sometimes that's going to go okay, but sometimes it's not. When you're playing with a Magic 8 Ball, the answer doesn't matter. But in some scenarios in real life, it very much does.

Tell us more about your recommendation that companies and the federal government should make users and consumers aware of LLMs' so-called 'Magic 8 Ball' component.

Consumers need to be fully informed of the capabilities and risks associated with ChatGPT and other LLMs. We are talking about serious issues like autonomous automobile companies using public roads as laboratory testing grounds for these technologies. Drivers should be aware that these vehicles are occupying the same roads they are.

I think there are also expectations that yes, these things malfunction, but they are being fixed and once that happens, all these issues will just go away. But there will always be issues. Hallucinations will not be a thing of the past any time soon.

That said, the issues with ChatGPT and other such models will get better because companies and organizations developing these technologies are building guardrails quickly. Still, it behooves users to remain skeptical. Harnessing AI is going to be difficult, and we don't want things to get worse before they get better.

Provided by Johns Hopkins University

Citation: Q&A: Are large language models the modern-day Magic 8 Ball? (2023, July 17) retrieved 2 May 2024 from https://techxplore.com/news/2023-07-qa-large-language-modern-day-magic.html