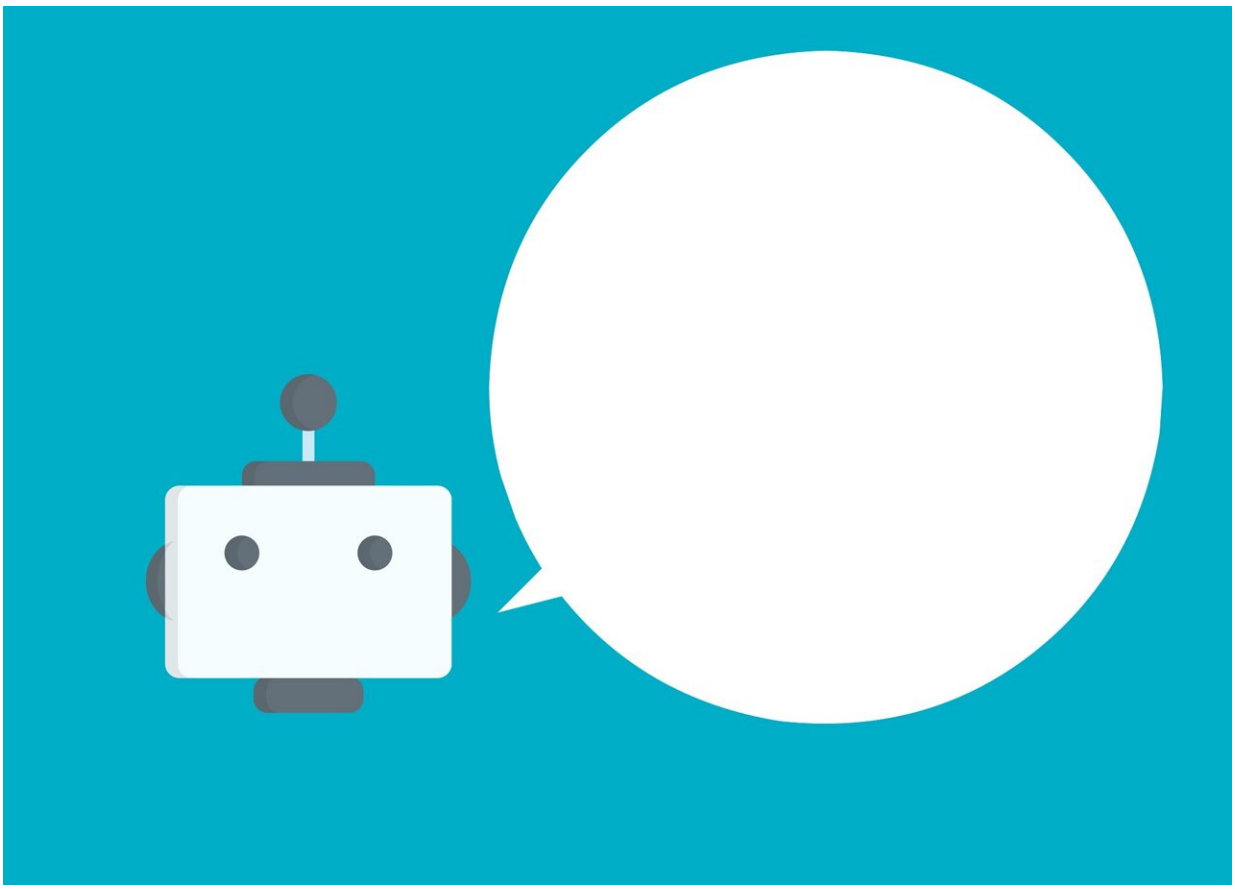
Large language models can help detect social media bots—but can also make the problem worse

August 28 2024, by Stefan Milne



Credit: Pixabay/CC0 Public Domain

An external study of Twitter in 2022 estimated that between a third and

two thirds of accounts on the social media site were bots. And many of these automatons flooding social media are dispatched to sow political polarization, hate, misinformation, propaganda and scams. The ability to sift them out of the online crowds is vital for a safer, more humane (or at least more human) internet.

But the recent proliferation of large language models (known as "LLMs"), such as OpenAI's ChatGPT and Meta's Llama, stands to complicate the world of social media bots.

A team led by University of Washington researchers found that while operators can use customized LLMs to make bots more sophisticated at evading automated detectors, LLMs can also improve systems that detect bots. In the team's tests, LLM-based bots reduced the performance of existing detectors by 30%. Yet researchers also found that an LLM trained specifically to detect social media bots outperformed state-of-the-art systems by 9%.

The team presented this research Aug. 11 at the 62nd Annual Meeting of the Association for Computational Linguistics in Bangkok.

"There's always been an arms race between bot operators and the researchers trying to stop them," said lead author Shangbin Feng, a doctoral student in the Paul G. Allen School of Computer Science & Engineering. "Each advance in bot detection is often met with an advance in bot sophistication, so we explored the opportunities and the risks that large language models present in this arms race."

Researchers tested LLMs' potential to detect bots in a few ways. When they fed Twitter data sets (culled before the platform became X) to off-the-shelf LLMs, including ChatGPT and Llama, the systems failed to accurately detect bots more than currently used technologies.

"Analyzing whether a user is a bot or not is much more complex than some of the tasks we've seen these general LLMs excel at, like recalling a fact or doing a grade-school math problem," Feng said.

This complexity comes in part from the need to analyze three types of information for different attributes to detect a bot: the metadata (number of followers, geolocation, etc.), the text posted online and the network properties (such as what accounts a user is following).

When the team fine-tuned the LLMs with instructions on how to detect bots based on these three types of information, the models were able to detect bots with greater accuracy than current state-of-the-art systems.

The team also explored how LLMs might make bots more sophisticated and harder to detect. First the researchers simply gave LLMs prompts such as, "Please rewrite the description of this bot account to sound like a genuine user."

They also tested more iterative, complicated approaches. In one test, the LLM would rewrite the bot post. The team then ran this through an existing bot-detection system, which would estimate the likelihood that a post was written by a bot. This process would be repeated as the LLM worked to lower that estimate. The team ran a similar test while removing and adding accounts that the bot followed to adjust its likelihood score.

These strategies, particularly rewriting the bots' posts, reduced the effectiveness of the bot detection systems by as much as 30%. But the LLM-based detectors the team trained saw only a 2.3% drop in effectiveness on these manipulated posts, suggesting that the best way to detect LLM-powered bots might be with LLMs themselves.

"This work is only a scientific prototype," said senior author Yulia

Tsvetkov, an associate professor in the Allen School. "We aren't releasing these systems as tools anyone can download, because in addition to developing technology to defend against malicious bots, we are experimenting with threat modeling of how to create an evasive bot, which continues the cat-and-mouse game of building stronger bots that need stronger detectors."

Researchers note that there are important limitations to using LLMs as bot detectors, such as the systems' potential to leak private information. They also highlight that the data used in the paper is from 2022, before Twitter effectively closed off its data to academic researchers.

In the future, researchers want to look at bot detection beyond text, such as memes or videos on other platforms such as TikTok, where newer data sets are available. The team also wants to expand the research into other languages.

"Doing this research across different languages is extremely important," Tsvetkov said. "We are seeing a lot of misinformation, manipulation and the targeting of specific populations as a result of various world conflicts."

Additional co-authors on this paper are Herun Wan and Ningnan Wang, both undergraduates at Xi'an Jiaotong University; Minnan Luo, an assistant professor at Xi'an Jiaotong University; and Zhaoxuan Tan, a doctoral student at the University of Notre Dame.

More information: Shangbin Feng et al. What Does the Bot Say? Opportunities and Risks of Large Language Models in Social Media Bot Detection, aclanthology.org/2024.acl-long.196/