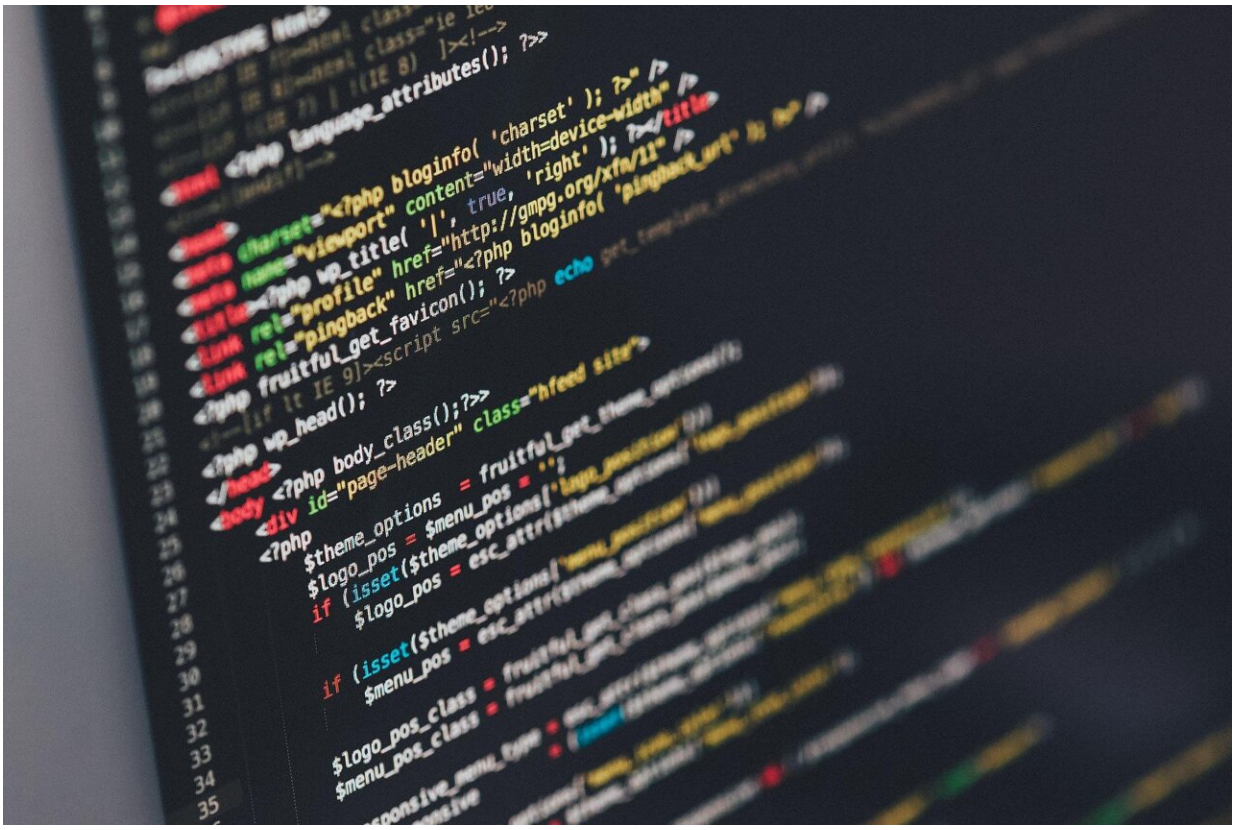
Engineers and OpenAI recommend ways to evaluate large language models for cybersecurity applications

April 2 2024



Credit: Pixabay/CC0 Public Domain

Carnegie Mellon University's Software Engineering Institute (SEI) and OpenAI published a white paper that found that large language models (LLMs) could be an asset for cybersecurity professionals, but should be evaluated using real and complex scenarios to better understand the technology's capabilities and risks. LLMs underlie today's generative artificial intelligence (AI) platforms, such as Google's Gemini, Microsoft's Bing AI, and ChatGPT, released in November 2022 by OpenAI.

These platforms take prompts from human users, use deep learning on large datasets, and produce plausible text, images or code. Applications for LLMs have exploded in the past year in industries including creative arts, medicine, law and software engineering and acquisition.

While in its early days, the prospect of using LLMs for cybersecurity is increasingly tempting. The burgeoning technology seems a fitting force multiplier for the data-heavy, deeply technical and often laborious field of cybersecurity. Add the pressure to stay ahead of LLM-wielding cyber attackers, including state-affiliated actors, and the lure grows even brighter.

However, it is hard to know how capable LLMs might be at cyber operations or how risky if used by defenders. The conversation around evaluating LLMs' capability in any professional field seems to focus on their theoretical knowledge, such as answers to standard exam questions. One preliminary study found that GPT-3.5 Turbo aced a common penetration testing exam.

LLMs may be excellent at factual recall, but it is not sufficient, according to the SEI and OpenAI paper "Considerations for Evaluating Large Language Models for Cybersecurity Tasks."

"An LLM might know a lot," said Sam Perl, a senior cybersecurity analyst in the SEI's CERT Division and co-author of the paper, "but does it know how to deploy it correctly in the right order and how to make tradeoffs?"

Focusing on theoretical knowledge ignores the complexity and nuance of real-world cybersecurity tasks. As a result, cybersecurity professionals cannot know how or when to incorporate LLMs into their operations.

The solution, according to the paper, is to evaluate LLMs on the same branches of knowledge on which a human cybersecurity operator would be tested: theoretical knowledge, or foundational, textbook information; practical knowledge, such as solving self-contained cybersecurity problems; and applied knowledge, or achievement of higher-level objectives in open-ended situations.

Testing a human this way is hard enough. Testing an artificial neural network presents a unique set of hurdles. Even defining the tasks is hard in a field as diverse as cybersecurity. "Attacking something is a lot different than doing forensics or evaluating a log file," said Jeff Gennari, team lead and senior engineer in the SEI CERT division and co-author of the paper. "Each task must be thought about carefully, and the appropriate evaluation should be designed."

Once the tasks are defined, an evaluation must ask thousands or even millions of questions. LLMs need that many to mimic the human mind's gift for semantic accuracy. Automation will be needed to generate the required volume of questions. That is already doable for theoretical knowledge.

But the tooling needed to generate enough practical or applied scenarios—and to let an LLM interact with an executable system—does not exist. Finally, computing the metrics on all those responses to

practical and applied tests will take new rubrics of correctness.

While the technology catches up, the white paper provides a framework for designing realistic cybersecurity evaluations of LLMs that starts with four overarching recommendations:

- Define the real-world task for the evaluation to capture.
- Represent tasks appropriately.
- Make the evaluation robust.
- Frame results appropriately.

Shing-hon Lau, a senior AI security researcher in the SEI's CERT division and one of the paper's co-authors, notes that this guidance encourages a shift away from focusing exclusively on the LLMs, for cybersecurity or any field. "We need to stop thinking about evaluating the model itself and move towards evaluating the larger system that contains the model or how using a model enhances human capability."

The SEI authors believe LLMs will eventually enhance human cybersecurity operators in a supporting role, rather than work autonomously. Even so, LLMs will still need to be evaluated, said Gennari. "Cyber professionals will need to figure out how to best use an LLM to support a task, then assess the risk of that use. Right now it's hard to answer either of those questions if your evidence is an LLM's ability to answer fact-based questions."

The SEI has long applied engineering rigor to cybersecurity and AI. Combining the two disciplines in the study of LLM evaluations is one way the SEI is leading AI cybersecurity research. Last year, the SEI also launched the AI Security Incident Response Team (AISIRT) to provide the United States with a capability to address the risks from the rapid growth and widespread use of AI.

OpenAI approached the SEI about LLM cybersecurity evaluations last year seeking to better understand the safety of the models underlying its generative AI platforms. OpenAI co-authors of the paper Joel Parish and Girish Sastry contributed first-hand knowledge of LLM cybersecurity and relevant policies. Ultimately, all the authors hope the paper starts a movement toward practices that can inform those deciding when to fold LLMs into cyber operations.

"Policymakers need to understand how to best use this technology on mission," said Gennari. "If they have accurate evaluations of capabilities and risks, then they'll be better positioned to actually use them effectively."

More information: Considerations for Evaluating Large Language Models for Cybersecurity Tasks. insights.sei.cmu.edu/library/c … cybersecurity-tasks/

Provided by Carnegie Mellon University

Citation: Engineers and OpenAI recommend ways to evaluate large language models for cybersecurity applications (2024, April 2) retrieved 17 May 2024 from https://techxplore.com/news/2024-04-openai-ways-large-language-cybersecurity.html