IndraStra
www.indrastra.com | ISSN 2381-3652

The Rise of ChatGPT and Large Language Models: Assessing the Risks

indrastra.com/2023/10/the-rise-of-chatgpt-and-large-language.html

IndraStra Global Friday, October 20, 2023 (2023-10-20T07:17:00-04:00) Edit this post
By IndraStra Global Tech Team

IndraStra
The Rise of ChatGPT and Large Language Models
Assessing the Risks
INDRASTRA-CREATIVES-AI202310005

Large language models (LLMs) and AI chatbots have taken the world by storm since the debut of ChatGPT in late 2022, with their effortless ability to answer a wide array of questions. This innovation has rapidly become one of the most rapidly growing consumer applications in recent memory, prompting competitors to develop their own models or quickly deploy their in-house creations. However, as with any emerging technology, there are concerns about the potential risks and implications for security. In this article, we'll delve into some of the cybersecurity aspects associated with ChatGPT and LLMs in general in the near term.

Unpacking ChatGPT and LLMs

ChatGPT, the brainchild of OpenAI, is an artificial intelligence chatbot powered by the GPT-3 language model, which was initially unveiled in 2020. This model harnesses deep learning techniques to generate text closely resembling human language. While ChatGPT and its contemporaries may be recent phenomena, the underlying LLM technology has been under development for a much longer.

An LLM is an algorithm trained on vast amounts of text-based data, typically collected from the open internet, encompassing web pages, and, depending on the specific model, additional sources like scientific research, books, or social media posts. Given the sheer volume of data processed, it is virtually impossible to filter out all offensive or erroneous content during the training process, resulting in the inclusion of 'controversial' material in the model.

These algorithms analyze the relationships between words and construct probability models, enabling them to respond to prompts, such as questions, with answers based on the word associations in their model. Although the data in the model remains static post-training, it can be refined through 'fine-tuning,' which involves further training with additional data, and 'prompt augmentation,' which provides context for the question by incorporating substantial amounts of text or entire documents into the prompt.

The Capabilities and Flaws of LLMs

LLMs, including ChatGPT, are undeniably impressive for generating a wide range of convincing content in numerous human and computer languages. However, they are not infallible and are far from achieving artificial general intelligence. These models exhibit significant flaws, including:

1. They can produce incorrect facts and 'hallucinate' erroneous information — an LLM might generate a false statement in response to a question, such as providing inaccurate historical data or spreading misinformation about a medical condition. This has implications for misinformation and disinformation campaigns, where adversaries could employ LLMs to generate fake news stories or manipulate historical facts to advance their agendas, potentially causing societal confusion and harm.

2. They are susceptible to bias and often respond to leading or leading questions — if an LLM is asked, "Why are certain people lazy?" it may produce a biased response, perpetuating stereotypes. This susceptibility to bias can have far-reaching consequences, inadvertently reinforcing prejudices and stereotypes and undermining efforts for fairness,

equity, and inclusivity. Training LLMs from scratch demands enormous computational resources and extensive.

3. Training LLMs from scratch demands enormous computational resources and extensive data — training a state-of-the-art LLM like GPT-3 or GPT-4 requires vast clusters of high-performance GPUs or TPUs, along with petabytes of text data. This immense computational and data requirement puts the development of such models out of reach for all but the most resource-rich organizations or research institutions, potentially exacerbating the technological divide and inequality in the field of artificial intelligence.

4. They can be manipulated into generating toxic content and are vulnerable to 'injection attacks.' — malicious actors can intentionally feed the model with biased or offensive prompts, leading it to generate hate speech, misinformation, or harmful advice. This kind of manipulation, often called 'injection attacks,' highlights the ethical and security concerns associated with LLMs. It underscores the need for robust safeguards and responsible usage to prevent the misuse of these powerful language models for harmful purposes.

Do LLMs Compromise Privacy and Security?

A common concern revolves around the potential leakage of information due to LLMs. While LLMs do not automatically incorporate data from queries into their models for others to use, they are stored. They may be employed for further model development by the organization offering the LLM service, such as OpenAI. This means that the provider or their partners may have access to queries, necessitating a thorough understanding of terms of use and privacy policies, especially when dealing with sensitive questions.

Sensitive questions can pertain to the data included in the query or the questioner's identity. For instance, a CEO inquiring about layoffs or someone seeking advice on personal health or relationships may be at risk. Additionally, aggregating information from multiple queries under the same login can pose a security concern.

Another risk is the potential for queries stored online to be hacked, leaked, or inadvertently made public, potentially exposing user-identifiable information. Furthermore, if the LLM operator is later acquired by an organization with different privacy practices, data entered initially by users could be compromised.

In light of these potential threats, all central cybersecurity authorities recommend refraining from including sensitive information in queries to public LLMs and avoiding queries that could lead to issues if made public.

Safeguarding Sensitive Information with LLMs

Organizations aiming to leverage LLMs for automating business processes that involve handling sensitive information can explore 'private LLMs.' These private models can be provided by cloud service providers or hosted independently by the organization. However, the path to adopting such technology has its caveats.

In the case of cloud-provided LLMs, organizations must take a meticulous approach to review the terms of use and privacy policies. It is essential to understand how the service provider handles data for fine-tuning and prompt augmentation. This entails assessing factors such as who has access to the data, the specific form in which it is stored, and the conditions under which provider employees can access queries. These considerations are pivotal in ensuring that sensitive data remains well-protected and complies with privacy regulations.

On the other hand, self-hosted LLMs, while potentially affording more control and security, come with substantial financial implications. Deploying and maintaining these models involves significant costs. To mitigate the security risks, organizations must conduct a rigorous security assessment. This evaluation should align with established machine learning security principles, as authoritative bodies like the National Cyber Security Centre (NCSC) recommended. This approach is indispensable in safeguarding the integrity and confidentiality of an organization's data as it navigates the landscape of private LLMs.

Leveraging LLMs for Cybercrime

The rise of LLMs has not gone unnoticed by cybercriminals. These models can facilitate malware creation and empower those with malicious intent, even with limited technical expertise. While LLMs excel at simple tasks, creating complex malware from scratch remains more efficient for experts. However, as LLMs evolve, this trade-off between creating malware from scratch and validating LLM-generated malware may shift.

LLMs can also be exploited for technical problem-solving in cyber attacks. Criminals may seek assistance from LLMs to escalate privileges or find data, potentially advancing their attacks. These answers may not always be entirely accurate, but they can still provide valuable insights to attackers. Furthermore, LLMs can craft convincing phishing emails in multiple languages, aiding attackers with high technical capabilities but needing more linguistic skills.

In the near term, we can expect to see:

1. More convincing phishing emails generated with LLMs — an attacker could use an LLM to craft a phishing email that appears to be from a well-known bank, complete with a legitimate-looking logo, proper grammar, and an urgent request for the recipient to update their account information. Such convincing emails can easily deceive unsuspecting individuals into divulging sensitive information like login credentials, leading to identity theft or financial losses.

2. Attackers experimenting with techniques previously beyond their skill level — an individual with limited hacking experience can use LLM-generated code and scripts to automate attacks, such as Distributed Denial of Service (DDoS) attacks or SQL injection, without having an in-depth understanding of these techniques. This democratization of cyber threats means that even those with minimal technical knowledge can now engage in attacks that were once the domain of expert hackers, potentially causing widespread disruptions and data breaches.

3. A low risk of lesser-skilled attackers creating highly capable malware — a novice hacker can use LLMs to generate a highly effective ransomware variant that encrypts victims' files and demands a ransom for decryption. This lowers the bar for entry into the cybercrime landscape and increases the potential for widespread cyberattacks, as even those with limited technical expertise can unleash sophisticated and damaging malware upon unsuspecting targets, businesses, or organizations.

Conclusion

The advent of LLMs, such as ChatGPT, has sparked a global wave of enthusiasm and innovation. This transformative technology, however, carries a dual allure, attracting not only fervent enthusiasts but also vigilant skeptics. As elucidated earlier, the potential risks inherent in the unrestricted utilization of public LLMs are undeniable. Consequently, individuals and organizations must exercise utmost caution when formulating prompts, ensuring that experimentation does not inadvertently compromise sensitive data and security. Responsible usage and prudent safeguards are essential to navigate LLMs' promising but perilous landscape.