AI & SOCIETY
https://doi.org/10.1007/s00146-024-01934-y

CURMUDGEON CORNER

Sticks and stones may break my bones, but words will never hurt me!—Navigating the cybersecurity risks of generative AI

Abdur Rahman Bin Shahid[1] · **Ahmed Imteaj**[2]

Received: 3 February 2024 / Accepted: 26 March 2024

Over the past few years, we have witnessed the spectacular emergence of a new class of artificial intelligence (AI) systems within the realm of Generative AI, known as Large Language Models (LLMs). Companies like OpenAI, Google, Microsoft, and Anthropic have developed LLMs like ChatGPT, Gemini, Copilot, and Claude that possess astonishing linguistic capabilities unlike anything seen before in AI. The innovation of LLMs lies in their vast neural networks trained on extensive text datasets, enabling them to utilize billions of parameters for a subtle understanding of language. This breakthrough allows them to generate text and converse contextually, surpassing previous AI in fluency and comprehension. With abilities for natural, meaningful interactions, LLMs have the potential to significantly impact society, improving productivity and education.

However, LLMs' potential comes with distinct cybersecurity challenges as their unprecedented language skills provide new avenues for malicious uses that threaten security. For instance, LLMs can craft convincing personalized phishing emails, bypassing traditional filters, and exploiting human trust. Their skill in adapting tone and style makes them formidable tools for cyber attackers. Furthermore, LLMs could be misused for malware generation, aiding even technically novice attackers in crafting targeted malware or automating basic attack tasks, thus broadening the threat landscape.

Furthermore, LLMs' ability to generate vast amounts of realistic fake news or propaganda poses a significant threat to public discourse integrity, potentially manipulating opinions and causing social unrest. This could disrupt infrastructure, influence elections, and damage reputations. LLMs' speed and scale in spreading misinformation surpass traditional detection, risking destabilization before effective countermeasures can be deployed. This risk extends to their training on massive datasets, which might include sensitive information, posing a data extraction risk by skilled attackers. Additionally, their multilingual capability removes linguistic barriers in cyber-attacks, enabling the crafting of targeted phishing campaigns and disinformation in various languages, thereby complicating detection efforts.

A significant risk also lies in their potential exploitation by extremist groups to efficiently produce radical content, targeting radicalization in vulnerable populations through automated communication. This challenge is further heightened in low-income countries, which may lack the infrastructure and resources to counter such threats. These countries often struggle with cybersecurity due to limited budgets, lack of skilled personnel, and outdated or inadequate technological resources. The introduction of LLMs into this already strained landscape could exacerbate existing vulnerabilities, making these nations more susceptible to cyber-attacks that could undermine national security. The disparity in cybersecurity capabilities between high-income and low-income nations also raises concerns about international security. As LLMs become more ubiquitous, the divide in the ability to protect against and respond to cyber threats will likely widen. Consequently, this not only threatens the security of individual countries but can also contribute to global instability. Furthermore, the decentralized nature of LLM development and the global accessibility of these models mean that malicious actors in any part of the world can harness their capabilities. This global reach, combined with the complexity of attributing cyber-attacks,

✉ Abdur Rahman Bin Shahid
shahid@cs.siu.edu

Ahmed Imteaj
imteaj@cs.siu.edu

[1] Secure and Trustworthy Intelligent Systems (SHIELD) Lab, School of Computing, Southern Illinois University, Carbondale, IL, USA

[2] Security, Privacy and Intelligence for Edge Devices (SPEED) Lab, School of Computing, Southern Illinois University, Carbondale, IL, USA

Published online: 13 April 2024

poses significant challenges for international law enforcement and cooperative security efforts.

In response to LLM-related challenges, the international community must prioritize developing a robust cybersecurity framework. Educating and training new cybersecurity professionals is crucial, but integrating LLMs in education raises ethical concerns around plagiarism and academic integrity. The lack of global ethical guidelines for AI in education further complicates matters. We need a multipronged approach: clear guidelines and best practices for responsible AI integration in education. This involves an ethical foundation emphasizing transparency, accountability, and data privacy, and a comprehensive global model to evaluate LLM cybersecurity. However, rapid AI advancement outpaces curriculum updates, leaving graduates with outdated skills. Resource-constrained institutions struggle with modern tools and experience. The absence of global consensus on data privacy and ethics in AI cybersecurity education adds another layer of complexity.

To address these complexities, collaborations between international bodies like Accreditation Board for Engineering and Technology (ABET), Association for Computing Machinery (ACM), National Centers of Academic Excellence (CAE), ISO/IEC 27001, British Computer Society (BCS), ANSSI, and $(ISC)^2$ are crucial to standardize and enhance cybersecurity education, ensuring graduates possess the skills to tackle LLM-related challenges. This necessitates a comprehensive, multifaceted strategy, including immersive, hands-on learning experiences through pilot studies and practical simulations embedded within an extensive curriculum. Additionally, the curriculum should be enriched with ethical and legal studies to deepen the understanding of LLMs' broader implications in cybersecurity. Moreover, the program should incorporate real-world elements, such as case studies, simulated phishing, and propaganda campaigns, to provide learners with practical scenarios. It is crucial to focus on understanding the specific ways LLMs can be targeted by threats. This includes using in-depth testing methods to find and fix vulnerabilities before they can be exploited. Moreover, it is vital to address the complications of AI bias and fairness, and their impact on cybersecurity. The educational strategy should also delve deeply into the complexities of LLMs' multilingual capabilities. This includes not only the technical aspects of processing multiple languages but also an appreciation of the varied cultural contexts and the potential risks of misinterpretation or misuse in diverse linguistic environments. Beyond the acquisition of technical skills, the curriculum must explore the confluence of integrity, psychology, and international relations. This aspect is particularly crucial in addressing the risks associated with LLMs, including potential issues of radicalization and manipulation. It is also imperative to promote community awareness about the potential risks and ethical considerations of LLMs in cybersecurity, and to develop crisis management and response strategies for cyber threats posed by LLMs. In executing these initiatives, it is vital to align efforts with the overarching goal of resolving critical challenges like resource allocation and accessibility in varied educational environments. This alignment ensures the promotion of inclusivity and diversity within cybersecurity education, preparing learners to adeptly handle the intricate challenges presented by LLMs in a globally connected cybersecurity panorama.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Data availability We do not analyze or generate any datasets, because our work proceeds within a theoretical approach.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.