Large Language Models

LLMs may be classified as a sub-category of NLP. However, because of their popularity, and their significant impact, we decided to define them as separate values. Li et al. (2023) specifically addressed privacy attacks in ChatGPT and focused on the new LLM's ChatGPT/Bing privacy problems. They present an experimental evaluation and also cover an interesting aspect of prompt and prompt-based attacks on LLMs. The main experiment that they present in the paper attempted to extract personal data via different types of prompt-based attacks on Bing and ChatGPT. The successful results of their experiments are also presented. Gupta, Akiri, Aryal, Parker, and Praharaj (2023) discussed the impact of generative AI, specifically ChatGPT, which has become widely used and popular in the past year, on cybersecurity and privacy. In their paper they present the vulnerabilities of ChatGPT, and their possible exploitation when bypassing the ethical constraints of the LLM model. They used attacks such as jailbreaks, reverse psychology, and prompt injection attacks for this purpose. They also present scenarios that indicate vulnerabilities that could lead to social engineering attacks, phishing attacks, automated hacking, payload generation attacks, malware creation, and polymorphic malware. The main solutions that the authors propose are cyber defense automation, reporting, threat intelligence, secure code generation and detection, attack identification, developing ethical guidelines, incidence response plans, and malware detection. In conclusion they discuss social, legal, and ethical implications of using ChatGPT. Mylrea and Robinson (2023) proposed an AI trust framework and maturity model to enhance trust in AI systems. They introduce the concept of using an entropy lens rooted in information theory to improve the transparency and trustworthiness of AI technologies, particularly in "black box" systems that lack ethical guardrails. The model aims to address high entropy in AI systems which can decrease human trust, especially in uncertain and competitive environments. The framework identifies new opportunities to optimize performance in autonomous human-machine teams and systems, with the validation of the framework demonstrated through two cases. Wei and Liu (2024) focused on enhancing the trustworthiness of distributed AI through robustness, privacy protection, and fairness in learning. They present the inherent challenges posed by distributed learning architectures, while detailing vulnerabilities and proposing a taxonomy of countermeasures to ensure AI integrity. The key aspects included robustness against various attacks, privacy enhancements during distributed learning, fairness and governance considerations related to AI data models. Wei and Liu highlight the ongoing need for robust, privacy-preserving, and fair AI systems, emphasizing the importance of integrating diverse governance policies into AI development. Peres, Manta-Costa, and Barata (2023) explored the application of OpenAI's ChatGPT in mental health services, by examining its potential to enhance therapy practices and its associated risks.

They discuss ChatGPT's capabilities in understanding and generating human-like responses, which can support mental health professionals by providing preliminary counselling, generating therapy session notes, and conducting initial patient assessments. However, they also highlight the significant risks related to privacy concerns, the accuracy of the AI's responses, and the ethical implications of using AI in sensitive settings. The authors conclude that while ChatGPT offers promising application in mental health, there is a crucial need for rigorous testing and ethical guidelines to mitigate the risks.

References

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*. doi:https://doi.org/10.1109/ACCESS.2023.3300381

Li, H., Guo, D., Fan, W., Xu, M., Huang, J., Meng, F., & Song, Y. (2023). Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*. doi:https://doi.org/10.48550/arXiv.2304.05197

Mylrea, M., & Robinson, N. (2023). Artificial Intelligence (AI) trust framework and maturity model: applying an entropy lens to improve security, privacy, and ethical AI. *Entropy, 25*(10), 1429. doi:https://doi.org/10.3390/e25101429

Peres, R. S., Manta-Costa, A. a., & Barata, J. (2023). Implementing Privacy-Preserving and Collaborative Industrial AI. *IEEE Acces*. doi:https://doi.org/10.1109/ACCESS.2023.3296143

Wei, W., & Liu, L. (2024). Trustworthy distributed ai systems: Robustness, privacy, and governance. *ACM Computing Surveys*. doi:https://doi.org/10.1145/3645102