6-18-2024

Artificial Intelligence Red-Teaming: Requirement Analysis of Competencies and Foundations

Elahe Javadi
Illinois State University, ejavadi@ilstu.edu

Elham Buxton
University of Illinois Springfield, esahe2@uis.edu

Shukri Abotteen
Illinois State University, sabotte@ilstu.edu

Ashley Strehlow
Illinois State University, aastreh@ilstu.edu

Artificial Intelligence Red-Teaming: Requirement Analysis of Competencies and Foundations

Dr. Elahe Javadi
Illinois State University
ejavadi@ilstu.edu

Elham Buxton
University of Illinois Springfield
esahe2@uis.edu

Shukri Abotteen
Illinois State University
sabotte@ilstu.edu

Ashley Strehlow
Illinois State University
aastreh@ilstu.edu

ABSTRACT

With the prevalence of artificial intelligence (AI) models in daily life and business (e.g., security cameras, customer service chatbots, supply chain), it is essential to develop competencies in AI safety, security, and assessment thereof. The process includes examining both exploits and unintended consequences. AI Red Teams competed for the first time in DEF CON 2023 in a program co-hosted by the White House. Red-teaming in AI still lacks a clear scope or required set of competencies, and in the 2023 competition, it manifested itself as prompt hacking (prompt injecting). However, not all AI models are large language models, and prompt hacking is one of the many possible exploits to which the broader set of AI tools can be vulnerable (e.g., HopSkipJump attacks, Chen *et al.* 2020). The majority of organizations will be using third-party pre-trained models (HiddenLayer Report 2024); therefore, foundational skills in both AI and security will be essential for securely and safely incorporating third-party models in production. It is important that educators and researchers in computing fields explore this landscape in order to remedy the lag in workforce development and research endeavors in the area. In this work, we review common AI vulnerabilities (in code, training, model, network, and output) and the existing literature on AI red-teaming and aim to formulate the scope. We then engage in requirement analysis to identify competencies and foundational knowledge that contribute to this area's curriculum work and research pursuits.

Keywords

AI Read Teams, pre-trained models, data poisoning, model theft, black-box backdoor attack

REFERENCES

1. AI Threat Landscape Report, 2024, HiddenLayer.
2. Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006, March). Can machine learning be secure?. In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security* (pp. 16-25).
3. Chen, J., Jordan, M. I., & Wainwright, M. J. (2020, May). Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)* (pp. 1277-1294). IEEE.
4. Feffer, M., Sinha, A., Lipton, Z. C., & Heidari, H. (2024). Red-Teaming for Generative AI: Silver Bullet or Security Theater?. *arXiv preprint arXiv:2401.15897*.
5. Kurita, K., Michel, P., & Neubig, G. (2020). Weight poisoning attacks on pre-trained models. arXiv preprint arXiv:2004.06660.
6. Li, Y., Hua, J., Wang, H., Chen, C., & Liu, Y. (2021, May). Deeppayload: Black-box backdoor attack on deep learning models through neural payload injection. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (pp. 263-274). IEEE.
7. Orekondy, T., Schiele, B., & Fritz, M. (2019). Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4954-4963).
8. Tian, Z., Cui, L., Liang, J., & Yu, S. (2022). A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, *55*(8), 1-35.