MINISTERUL EDUCAȚIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare, Informatică şi Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susţinere
Şef departament:
FIODOROV Ion dr., conf.univ.

„___" ______________ 2024

Prevenirea atacurilor cibernetice bazate pe exploatarea modelelor lingvistice în contextul inteligenţei artificiale

Prevention of cyber attacks based on the exploitation of linguistic models in the context of artificial intelligence

Proiect de master

Student:	______________	**Dimoglo Alexandr, SI-221M**
Coordonator:	______________	**Bulai Rodica, asist. univ.**
Consultant:	______________	**Bulai Rodica, asist.univ.**

Chişinău, 2024

REZUMAT

Progresul rapid al inteligenței artificiale (AI) și integrarea sa pe scară largă în diverse platforme digitale au dus la o creștere a riscului de atacuri cibernetice sofisticate, în special a celor care exploatează modele lingvistice. Această lucrare prezintă o analiză cuprinzătoare a vulnerabilităților inerente modelelor lingvistice actuale bazate pe inteligență artificială și propune un cadru nou pentru atenuarea acestor riscuri. Începem prin a examina tipurile de amenințări cibernetice care exploatează nuanțele procesării limbajului natural, cum ar fi manipularea contextului, otrăvirea modelului și atacurile de inferență a datelor. Apoi explorăm limitările măsurilor de securitate existente în contracararea acestor amenințări. Demonstrăm eficiența diferitelor abordări printr-o serie de simulări și studii de caz din lumea reală, demonstrând o reducere semnificativă a atacurilor cibernetice reușite. Această cercetare contribuie la acest domeniu prin furnizarea de soluții practice pentru consolidarea securității sistemelor bazate pe IA împotriva exploatării modelelor lingvistice, asigurând astfel o utilizare mai sigură și mai fiabilă a IA în diverse aplicații.

ABSTRACT

The rapid advancement of artificial intelligence (AI) and its widespread integration into various digital platforms have led to an increased risk of sophisticated cyber attacks, especially those exploiting linguistic models. This paper presents a comprehensive analysis of the vulnerabilities inherent in current AI-based linguistic models and proposes a novel framework for mitigating these risks. We begin by examining the types of cyber threats that exploit the nuances of natural language processing, such as context manipulation, model poisoning, and data inference attacks. We then explore the limitations of existing security measures in countering these threats. We demonstrate the effectiveness of different approaches athrough a series of simulations and real-world case studies, showing a significant reduction in successful cyber attacks. This research contributes to the field by providing practical solutions for bolstering the security of AI-driven systems against linguistic model exploitation, thus ensuring safer and more reliable use of AI in various applications.

TABLE OF CONTENTS

INTRODUCTION8
1. RESEARCH AND ANALYSIS9
1.1 Adoption of Generative AI 10
1.2 Techniques and Prevention 12
1.3 Lure 20
1.4 Developing 22
25
2.1 Adversarial Poisoning of Language Models: Vulnerabilities and 26
2.2 Unveiling Vulnerabilities in Large Multi-Modal Models (LMMs): AdversarialAttacks 29
2.3 30
2.4 Understanding and Mitigating Hazards in Language Model Deployments:The Role of Red 32
2.5 Emerging 36
2.6 37
2.7 39
45
3.1 46
3.2 48
3.3 49
3.4 53
4. 56
59
62

Introduction

Since at least 2019, cybersecurity researchers have tracked threat actors' interest in and use of AI capabilities to facilitate a variety of malicious activities. Based on their own observations and open source reports, the use of AI in intrusion operations remains limited and primarily associated with social engineering.

In contrast, information operations actors with varying motivations and capabilities have increasingly used AI-generated content, particularly images and videos, in their campaigns, likelyat least in part due to the readily apparent applications of such fabrications in disinformation. In addition, the release of several generative AI tools in the past year has led to renewed interest inthe implications of these capabilities.

Cybersecurity researchers expect generative AI tools to accelerate threat actors' incorporation of AI into information operations and intrusion activities. They believe that such technologies have the potential to significantly augment malicious operations in the future, enabling threat actors with limited resources and capabilities, similar to the benefits provided by exploit frameworks such as Metasploit or Cobalt Strike. And while adversaries are already experimenting, and weexpect to see more use of AI tools over time, effective operational use remains limited.

REFERENCES

1. Seymour, J., & Tully, P. (2023). Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter Presentation. Cybersecurity Conference 2023, San Francisco.
2. Esmradi, A., Yip, D. W., & Chan, C. F. (n.d.). A Comprehensive Survey of Attack Techniques, Imple-mentation, and Mitigation Strategies in Large Language Models.
3. Zhang, X., Zhang, C., Li, T., Huang, Y., Jia, X., Xie, X., Liu, Y., & Shen, C. (2023). A Mutation-Based Method for Multi-Modal Jailbreaking Attack Detection. http://arxiv.org/abs/2312.10766
4. Cao, Y., Cao, B., & Chen, J. (2023). Stealthy and Persistent Unalignment on Large Language Models via Backdoor Injections. http://arxiv.org/abs/2312.00027
5. Liu, X., Zhu, Y., Lan, Y., Yang, C., & Qiao, Y. (2023). Query-Relevant Images Jailbreak Large Multi-Modal Models. http://arxiv.org/abs/2311.17600
6. Jiang, F., Xu, Z., Niu, L., Wang, B., Jia, J., Li, B., & Poovendran, R. (2023). Identifying and Mitigating Vulnerabilities in LLM-Integrated Applications. http://arxiv.org/abs/2311.16153
7. Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). [2307.15043] Universal and Transferable Adversarial Attacks on Aligned Language Models
8. Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In IEEE Symposium on Security & Privacy, 2017
9. Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., and Zhou, W. Hidden voice commands. In 25th USENIX Security Symposium (USENIX Security 16), Austin, TX, 2016.
10. Brendel, W., Rauber, J., and Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In Proceedings of the International Conference on Learning Representations (ICLR), 2018. URL https://arxiv.org/abs/1712.04248
11. Yuan, Y., Jiao, W., Wang, W., Huang, J., He, P., Shi, S. and Tu, Z. (2023). GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher. http://arxiv.org/abs/2308.06463.
12. Yao, D., Zhang, J., Harris, I.G. and Carlsson, M. (2023). FuzzLLM: A novel and universal fuzzing framework for proactively discovering jailbreak vulnerabilities in large language models. http://arxiv.org/abs/2309.05274.
13. Wei, A., Nika Haghtalab and Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? http://arxiv.org/abs/2307.02483.
14. Jones, E., Dragan, A., Raghunathan, A. and Steinhardt, J. (2023). Automatically auditing large language models via discrete optimization. http://arxiv.org/abs/2303.04381.
15. Ilyas, A., Engstrom, L., Anish Athalye and Lin, J. (2018). Black-box adversarial attacks with limited queries and information. http://arxiv.org/abs/1804.08598.
16. S. Casper, J. Lin, J. Kwon, G. Culp, & D. Hadfield-Menell, «Explore, Establish, Exploit: Red Teaming Language Models from Scratch». arXiv, 2023. doi: 10.48550/arXiv.2306.09442.
17. A. Rao, S. Vashistha, A. Naik, S. Aditya, & M. Choudhury, «Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks». arXiv, 2023 г. doi: 10.48550/arXiv.2305.14965.
18. Boxin Wang, Wei Ping, Chaowei Xiao, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Bo Li, Anima Anandkumar, and Bryan Catanzaro. Exploring the limits of domain-adaptive training for detoxifying large-scale language models. NeurIPS, 35:35811–35824, 2022.
19. Yimu Wang, Peng Shi, and Hongyang Zhang. Investigating the existence of “secret language” in language models. arXiv preprint arXiv:2307.12507, 2023.
20. Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? arXiv preprint arXiv:2307.02483, 2023a.
21. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS,

35:24824–24837, 2022.

22. Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. arXiv preprint arXiv:2303.03846, 2023b.

23. Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. In Findings of EMNLP, pp. 2447–2469, 2021. URL https://aclanthology.org/2021.findings-emnlp.210.

24. Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for safety in open-domain chatbots. arXiv preprint arXiv:2010.07079, 2020.

25. Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. arXiv preprint arXiv:2305.11206, 2023.

26. Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

27. Jiang, S., Kadhe, S. R., Zhou, Y., Cai, L., & Baracaldo, N. (2023). Forcing Generative Models to Degenerate Ones: The Power of Data Poisoning Attacks. http://arxiv.org/abs/2312.04748

28. Lapid, R., Langberg, R., Sipper, M. (2023). Open Sesame! Universal Black Box Jailbreaking of Large Language Models. http://arxiv.org/abs/2309.01446

29. Safa Ozdayi, M., Peris, C., Fitzgerald, J., Dupuy, C., Majmudar, J., Khan, H., Parikh, R., Gupta, R. (n.d.). Controlling the Extraction of Memorized Data from Large Language Models via Prompt-Tuning https://github.com/amazon-science/controlling-llm

30. Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593, 2019.