**LLM-based Guardrail Model for Secure GenAI System Resistant to Prompt Injection Attacks**

by

David Guo and Donald Lien

**I. Introduction**

Generative AI offers substantial benefits in automating and enhancing various tasks such as prompt injection and jailbreaking, its potential to introduce new vulnerabilities necessitates a cautious and well-informed approach to its deployment and usage in sensitive domains. This proposal aims to use open source LLM based system-level safeguard, such as Meta Llama guard to detect and prevent security threats posed by misuse of Generative AI. As a result, this proposal will outline potential research opportunities for managing these risks.

**II. Problem Statements:**

1. Emergent threat vectors: unexpected GenAI capabilities can create unforeseen threat vectors.
2. Expanded attack surfaces: reliance on a large amount of user-generated datasets for training and inference exposes a much larger attack surface.
3. Deep integrations: unmediated connections with other computer systems facilitate the malicious attacks.
4. Economic value: the growing GenAI-powered applications pose a more vulnerable target for attackers.

**III. Opportunities:**

The evolving nature of GenAI systems necessitates a robust frameworks in cybersecurity and governance, emphasizing the importance of developing robust frameworks to mitigate these risks and ensuring that the advancement in AI technology is aligned with ethical standards and security protocols to prevent misuse. It requests researchers and developers to design applications with security in mind, reducing vulnerabilities and protecting users from potential threats and attacks. Organizations should also invest on the secure GenAI systems to protect their reputation, brand, and sensitive data. Last but not least, governments should also play an active role to develop policies and regulations that will govern the development, deployment, and use of GenAI systems , balance the innovation with security concerns.

**IV. Objectives**

**Object 1** - Develop a robust solution for detecting and preventing jailbreaking attacks on GenAI models.

Specific: Identify and analyze the patterns of malicious prompts used in jailbreaking attacks.

Measurable: Achieve a detection rate of equal or better benchmark performance for jailbreaking attacks within the next 6 months.

Achievable: apply natural language processing and cybersecurity technologies.

Relevant: Ensure that the solution is compatible with various open source GenAI models and applications, reducing the risk their vulnerabilities.

**Object 2** - Implement secure prompt engineering techniques for GenAI-powered applications.

Specific: Design and implement a set of guidelines for crafting prompts that minimize the risk of unintended outputs or attacks.

Measurable: Achieve a reduction of equal or higher percentage in the number of unintended outputs or attacks within the next 12 months.

Achievable: Collaborate with human-in-the-loop and natural language processing approaches to develop the guidelines.

Relevant: Ensure that the guidelines can be practicable by developers and users and help reduce the risk of vulnerabilities.

**V. Research Scope**

The scope of this proposal includes:

1. conduct a comprehensive literature review on the prompt injection attacks used in various GenAI models, identifying patterns, and characteristics that can be exploited for improvement.
2. design and implement robust detection techniques such as graph-based detection method for identifying prompt injection attacks, including machine learning-based approaches and rule-based systems.
3. investigate and develop effective counter actions to prevent or minimize the impact of prompt injection attacks on GenAI models.
4. fine-tune GenAI models resistant to prompt injection attacks.

**VI. Methodology**

1) Use real-time keyword filtrating with a stand-by knowledge database. This method is simple but could end up with significant amount of false positive. Ideally, this method can collaborate with other semantic detection model to achieve a high accuracy.

2) Joint effort of the use of keyword filtration and open source safeguard LLM model to analyze the context of the user's request, including the type of task being performed and the user's previous interactions with the model.

3) finetune open source safeguard LLM model such as Llama Guard, a Llama2-7b model that is instruction-tuned on collected dataset. The model can be used as the base model and retained on customized adversarial examples generated by injecting malicious prompts, enabling them to learn robustness against such attacks. This method relies on a high quality of customized training data that could be a time-consuming step. Without starting the data preparation from the scratch, we can start with the public prompt injection dataset as below.

Open source GenAI safeguard models:

1. [microsoft/deberta-v3-base](https://huggingface.co/microsoft/deberta-v3-base)
2. <https://huggingface.co/meta-llama/Meta-Llama-Guard-2-8B>

Prompt Injections datasets:

1. https://huggingface.co/datasets/deepset/prompt-injections
2. https://paperswithcode.com/dataset/bipia
3. https://huggingface.co/datasets/jerpint-org/HackAPrompt-Playground-Submissions
4. https://huggingface.co/datasets/jerpint-org/HackAPrompt-AICrowd-Submissions
5. https://huggingface.co/datasets/imoxto/prompt\_injection\_hackaprompt\_gpt35
6. https://huggingface.co/datasets/imoxto/prompt\_injection\_cleaned\_dataset-v2
7. https://github.com/compass-ctf-team/prompt\_injection\_research/blob/main/dataset/prompt-injection-dataset.csv
8. https://github.com/rabbidave/Denzel-Crocker-Hunting-For-Fairly-Odd-Prompts/blob/main/bad\_prompts.csv
9. <https://github.com/laiyer-ai/llm-guard/blob/main/llm_guard/resources/jailbreak.json.txt>

By implementing above 3 methods, we expect to effectively detect prompt injection attacks to LLM based applications and prevent malicious response to users, such as unauthorized access to sensitive information.

**VII. Timeline**

1) Literature review – Done

2) Preparation of prompt injection data – 2 months

3) Test and open source safeguard model on the prepared data from step 2 – 2 months

4) finetune safeguard model with the prepared data – 1 month

**VIII. Deliverables**

1) A working paper for conference (<https://2024.aclweb.org/)> or a draft for journal publication (TBD)

2) A Github repository with development code and materials - <https://github.com/bcqguo/llm_guardrail>

3) An application for demo purpose - TBD

**Reference:**

Abdali, Sara, Richard Anarfi, C. J. Barberan, and Jia He. “Securing Large Language Models: Threats, Vulnerabilities and Responsible Practices.” arXiv, March 19, 2024. <https://doi.org/10.48550/arXiv.2403.12503>.

Agrawal, Garima, Kuntal Pal, Yuli Deng, Huan Liu, and Ying-Chih Chen. “CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs.” *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 21 (March 24, 2024): 23164–72. <https://doi.org/10.1609/aaai.v38i21.30362>.

“AI, ML, AND LARGE LANGUAGE MODELS IN CYBERSECURITY.” *International Research Journal of Modernization in Engineering Technology and Science*, n.d.

Arawjo, Ian, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena Glassman. “ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing.” arXiv, December 20, 2023. <https://doi.org/10.48550/arXiv.2309.09128>.

Bae, Chanwoo, Guanhong Tao, Zhuo Zhang, and Xiangyu Zhang. “Threat Behavior Textual Search by Attention Graph Isomorphism.” In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Yvette Graham and Matthew Purver, 2616–30. St. Julian’s, Malta: Association for Computational Linguistics, 2024. <https://aclanthology.org/2024.eacl-long.160>.

Behdadnia, Tohid, Klaas Thoelen, Fairouz Zobiri, and Geert Deconinck. “Spatial-Temporal Graph Neural Network for Detecting and Localizing Anomalies in PMU Networks.” In *Dependable Computing – EDCC 2024 Workshops*, edited by Behrooz Sangchoolie, Rasmus Adler, Richard Hawkins, Philipp Schleiss, Alessia Arteconi, and Adriano Mancini, 75–82. Cham: Springer Nature Switzerland, 2024. <https://doi.org/10.1007/978-3-031-56776-6_7>.

Bethany, Mazal, Athanasios Galiopoulos, Emet Bethany, Mohammad Bahrami Karkevandi, Nishant Vishwamitra, and Peyman Najafirad. “Large Language Model Lateral Spear Phishing: A Comparative Study in Large-Scale Organizational Settings.” arXiv, January 18, 2024. <https://doi.org/10.48550/arXiv.2401.09727>.

Bou-Harb, Elias, and Mark Scanlon. “Behavioral Service Graphs: A Formal Data-Driven Approach for Prompt Investigation of Enterprise and Internet-Wide Infections.” *Digital Investigation*, DFRWS 2017 Europe, 20 (March 1, 2017): S47–55. <https://doi.org/10.1016/j.diin.2017.02.002>.

———. “Behavioral Service Graphs: A Formal Data-Driven Approach for Prompt Investigation of Enterprise and Internet-Wide Infections.” *Digital Investigation*, DFRWS 2017 Europe, 20 (March 1, 2017): S47–55. <https://doi.org/10.1016/j.diin.2017.02.002>.

Cui, Tianyu, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, et al. “Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems.” arXiv, January 11, 2024. <https://doi.org/10.48550/arXiv.2401.05778>.

Das, Badhan Chandra, M. Hadi Amini, and Yanzhao Wu. “Security and Privacy Challenges of Large Language Models: A Survey.” arXiv, January 29, 2024. <https://doi.org/10.48550/arXiv.2402.00888>.

Derner, Erik, and Kristina Batistič. “Beyond the Safeguards: Exploring the Security Risks of ChatGPT.” arXiv, May 13, 2023. <https://doi.org/10.48550/arXiv.2305.08005>.

Derner, Erik, Kristina Batistič, Jan Zahálka, and Robert Babuška. “A Security Risk Taxonomy for Large Language Models.” arXiv, November 19, 2023. <https://doi.org/10.48550/arXiv.2311.11415>.

Dong, Zhichen, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. “Attacks, Defenses and Evaluations for LLM Conversation Safety: A Survey.” arXiv, March 27, 2024. <https://doi.org/10.48550/arXiv.2402.09283>.

Ebert, Christof, and Maximilian Beck. “Artificial Intelligence for Cybersecurity.” *IEEE Software* 40, no. 6 (November 2023): 27–34. <https://doi.org/10.1109/MS.2023.3305726>.

Egersdoerfer, Chris, Di Zhang, and Dong Dai. “Early Exploration of Using ChatGPT for Log-Based Anomaly Detection on Parallel File Systems Logs.” In *Proceedings of the 32nd International Symposium on High-Performance Parallel and Distributed Computing*, 315–16. HPDC ’23. New York, NY, USA: Association for Computing Machinery, 2023. <https://doi.org/10.1145/3588195.3595943>.

Elatoubi, Mohamed. “Phishing in Web 3.0: Opportunities for the Attackers, Challenges for the Defenders.” *ARIS2 - Advanced Research on Information Systems Security* 3, no. 2 (December 30, 2023): 11–25. <https://doi.org/10.56394/aris2.v3i2.35>.

Gao, Chen, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. “Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives.” arXiv, December 19, 2023. <https://doi.org/10.48550/arXiv.2312.11970>.

Garza, Ethan, Erik Hemberg, Stephen Moskal, and Una-May O’Reilly. “Assessing Large Language Model’s Knowledge of Threat Behavior in MITRE ATT&CK,” 2023.

Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. “Not What You’ve Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection.” In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90. AISec ’23. New York, NY, USA: Association for Computing Machinery, 2023. <https://doi.org/10.1145/3605764.3623985>.

Gu, Zhaopeng, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. “AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models.” *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 3 (March 24, 2024): 1932–40. <https://doi.org/10.1609/aaai.v38i3.27963>.

Gupta, Maanak, Charankumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. “From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy.” *IEEE Access* 11 (2023): 80218–45. <https://doi.org/10.1109/ACCESS.2023.3300381>.

Huang, Ken, Grace Huang, Yuyan Duan, and Ju Hyun. “Utilizing Prompt Engineering to Operationalize Cybersecurity.” In *Generative AI Security: Theories and Practices*, edited by Ken Huang, Yang Wang, Ben Goertzel, Yale Li, Sean Wright, and Jyoti Ponnapalli, 271–303. Cham: Springer Nature Switzerland, 2024. <https://doi.org/10.1007/978-3-031-54252-7_9>.

Imtiaz, Ahsan, Danish Shehzad, Fawad Nasim, Muhammad Afzaal, Muhammad Rehman, and Ali Imran. “Analysis of Cybersecurity Measures for Detection, Prevention, and Misbehaviour of Social Systems.” In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 1–7, 2023. <https://doi.org/10.1109/SNAMS60348.2023.10375405>.

Injadat, MohammadNoor, Abdallah Moubayed, Ali Bou Nassif, and Abdallah Shami. “Machine Learning towards Intelligent Systems: Applications, Challenges, and Opportunities.” *Artificial Intelligence Review* 54, no. 5 (June 1, 2021): 3299–3348. <https://doi.org/10.1007/s10462-020-09948-w>.

Kilovaty, Ido. “Hacking Generative AI.” SSRN Scholarly Paper. Rochester, NY, March 1, 2024. <https://papers.ssrn.com/abstract=4788909>.

Koike, Ryuto, Masahiro Kaneko, and Naoaki Okazaki. “OUTFOX: LLM-Generated Essay Detection Through In-Context Learning with Adversarially Generated Examples.” *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 19 (March 24, 2024): 21258–66. <https://doi.org/10.1609/aaai.v38i19.30120>.

Kumar, Ashutosh, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. “The Ethics of Interaction: Mitigating Security Threats in LLMs.” arXiv, January 22, 2024. <https://doi.org/10.48550/arXiv.2401.12273>.

Kumarasinghe, Udesh, Ahmed Lekssays, Husrev Taha Sencar, Sabri Boughorbel, Charitha Elvitigala, and Preslav Nakov. “Semantic Ranking for Automated Adversarial Technique Annotation in Security Text.” arXiv, March 25, 2024. <https://doi.org/10.48550/arXiv.2403.17068>.

Li, Peiwen, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenweu Zhu. “LLM-Enhanced Causal Discovery in Temporal Domain from Interventional Data.” arXiv, April 23, 2024. <https://doi.org/10.48550/arXiv.2404.14786>.

Li, Siyu, Jin Yang, Gang Liang, Tianrui Li, and Kui Zhao. “SybilFlyover: Heterogeneous Graph-Based Fake Account Detection Model on Social Networks.” *Knowledge-Based Systems* 258 (December 22, 2022): 110038. <https://doi.org/10.1016/j.knosys.2022.110038>.

Liu, Chen, Shibo He, Qihang Zhou, Shizhong Li, and Wenchao Meng. “Large Language Model Guided Knowledge Distillation for Time Series Anomaly Detection.” arXiv, January 26, 2024. <https://doi.org/10.48550/arXiv.2401.15123>.

Liu, Tong, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. “Making Them Ask and Answer: Jailbreaking Large Language Models in Few Queries via Disguise and Reconstruction.” arXiv, February 28, 2024. <https://doi.org/10.48550/arXiv.2402.18104>.

Liu, Yilun, Shimin Tao, Weibin Meng, Jingyu Wang, Wenbing Ma, Yanqing Zhao, Yuhang Chen, Hao Yang, Yanfei Jiang, and Xun Chen. “Interpretable Online Log Analysis Using Large Language Models with Prompt Strategies.” arXiv, January 25, 2024. <https://doi.org/10.48550/arXiv.2308.07610>.

Maalem Lahcen, Rachid Ait, Bruce Caulkins, Ram Mohapatra, and Manish Kumar. “Review and Insight on the Behavioral Aspects of Cybersecurity.” *Cybersecurity* 3, no. 1 (April 21, 2020): 10. <https://doi.org/10.1186/s42400-020-00050-w>.

McIntosh, Timothy R., Teo Susnjak, Tong Liu, Paul Watters, Raza Nowrozy, and Malka N. Halgamuge. “From COBIT to ISO 42001: Evaluating Cybersecurity Frameworks for Opportunities, Risks, and Regulatory Compliance in Commercializing Large Language Models.” arXiv, February 24, 2024. <https://doi.org/10.48550/arXiv.2402.15770>.

Nicholls, Jack, Aditya Kuppa, and Nhien-An Le-Khac. “Enhancing Illicit Activity Detection Using XAI: A Multimodal Graph-LLM Framework.” arXiv, October 20, 2023. <https://doi.org/10.48550/arXiv.2310.13787>.

Niu, Zhenxing, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. “Jailbreaking Attack against Multimodal Large Language Model.” arXiv, February 3, 2024. <https://doi.org/10.48550/arXiv.2402.02309>.

———. “Jailbreaking Attack against Multimodal Large Language Model.” arXiv, February 3, 2024. <https://doi.org/10.48550/arXiv.2402.02309>.

O’Brien, Joe, Shaun Ee, and Zoe Williams. “Deployment Corrections: An Incident Response Framework for Frontier AI Models.” arXiv, September 30, 2023. <https://doi.org/10.48550/arXiv.2310.00328>.

Otal, Hakan T., and M. Abdullah Canbaz. “LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration.” arXiv, January 12, 2024. <https://doi.org/10.48550/arXiv.2402.10908>.

Pan, Jonathan, Swee Liang Wong, and Yidi Yuan. “RAGLog: Log Anomaly Detection Using Retrieval Augmented Generation.” arXiv, November 9, 2023. <https://doi.org/10.48550/arXiv.2311.05261>.

Pastor-Galindo, Javier, Pantaleone Nespoli, and José A. Ruipérez-Valiente. “Large-Language-Model-Powered Agent-Based Framework for Misinformation and Disinformation Research: Opportunities and Open Challenges.” *IEEE Security & Privacy*, 2024, 2–14. <https://doi.org/10.1109/MSEC.2024.3380511>.

Pu, Yujiang, Xiaoyu Wu, Lulu Yang, and Shengjin Wang. “Learning Prompt-Enhanced Context Features for Weakly-Supervised Video Anomaly Detection.” arXiv, January 22, 2024. <https://doi.org/10.48550/arXiv.2306.14451>.

Purificato, Erasmo, Ludovico Boratto, and Ernesto William De Luca. “User Modeling and User Profiling: A Comprehensive Survey.” arXiv, February 20, 2024. <https://doi.org/10.48550/arXiv.2402.09660>.

Qammar, Attia, Hongmei Wang, Jianguo Ding, Abdenacer Naouri, Mahmoud Daneshmand, and Huansheng Ning. “Chatbots to ChatGPT in a Cybersecurity Space: Evolution, Vulnerabilities, Attacks, Challenges, and Future Recommendations.” arXiv, May 29, 2023. <https://doi.org/10.48550/arXiv.2306.09255>.

Qi, Jiaxing, Shaohan Huang, Zhongzhi Luan, Shu Yang, Carol Fung, Hailong Yang, Depei Qian, Jing Shang, Zhiwen Xiao, and Zhihui Wu. “LogGPT: Exploring ChatGPT for Log-Based Anomaly Detection.” In *2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, 273–80, 2023. <https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys60770.2023.00045>.

Rahman, Md Rayhanur, Brandon Wroblewski, Quinn Matthews, Brantley Morgan, Tim Menzies, and Laurie Williams. “Mining Temporal Attack Patterns from Cyberthreat Intelligence Reports.” arXiv, January 3, 2024. <https://doi.org/10.48550/arXiv.2401.01883>.

Rjoub, Gaith, Jamal Bentahar, Omar Abdel Wahab, Rabeb Mizouni, Alyssa Song, Robin Cohen, Hadi Otrok, and Azzam Mourad. “A Survey on Explainable Artificial Intelligence for Cybersecurity.” *IEEE Transactions on Network and Service Management* 20, no. 4 (December 2023): 5115–40. <https://doi.org/10.1109/TNSM.2023.3282740>.

Sandoval, Gustavo, Hammond Pearce, Teo Nys, Ramesh Karri, Siddharth Garg, and Brendan Dolan-Gavitt. “Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants,” 2205–22, 2023. <https://www.usenix.org/conference/usenixsecurity23/presentation/sandoval>.

Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, et al. “Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives.” arXiv, September 29, 2023. <https://doi.org/10.48550/arXiv.2311.09227>.

Shandilya, Shishir Kumar, Gsv Prharsha, Agni Datta, Gaurav Choudhary, Hoonyong Park, and Ilsun You. “GPT Based Malware: Unveiling Vulnerabilities and Creating a Way Forward in Digital Space.” In *2023 International Conference on Data Security and Privacy Protection (DSPP)*, 164–73, 2023. <https://doi.org/10.1109/DSPP58763.2023.10404552>.

Shang, Wenbo, and Xin Huang. “A Survey of Large Language Models on Generative Graph Analytics: Query, Learning, and Applications.” arXiv, April 23, 2024. <https://doi.org/10.48550/arXiv.2404.14809>.

Squillace, Joseph, Justice Cappella, and Andrew Sepp. “User Vulnerabilities in AI-Driven Systems: Current Cybersecurity Threat Dynamics and Malicious Exploits in Supply Chain Management and Project Management.” In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, 1760–65, 2024. <https://doi.org/10.1109/ICETSIS61505.2024.10459480>.

Su, Jing, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. “Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review.” arXiv, February 15, 2024. <https://doi.org/10.48550/arXiv.2402.10350>.

Suo, Xuchen. “Signed-Prompt: A New Approach to Prevent Prompt Injection Attacks Against LLM-Integrated Applications.” arXiv, January 15, 2024. <https://doi.org/10.48550/arXiv.2401.07612>.

Tang, Xiangru, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, et al. “Prioritizing Safeguarding Over Autonomy: Risks of LLM Agents for Science.” arXiv, February 7, 2024. <https://doi.org/10.48550/arXiv.2402.04247>.

Veerasamy, Namosha, Danielle Badenhorst, Mazwi Ntshangase, Errol Baloyi, Noku Siphambili, and Oyena Mahlasela. “Unpacking AI Security Considerations.” *International Conference on Cyber Warfare and Security* 19, no. 1 (March 21, 2024). <https://doi.org/10.34190/iccws.19.1.2104>.

Wang, Yuxia, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. “Do-Not-Answer: Evaluating Safeguards in LLMs.” In *Findings of the Association for Computational Linguistics: EACL 2024*, edited by Yvette Graham and Matthew Purver, 896–911. St. Julian’s, Malta: Association for Computational Linguistics, 2024. <https://aclanthology.org/2024.findings-eacl.61>.

Xu, Jiacen, Jack W. Stokes, Geoff McDonald, Xuesong Bai, David Marshall, Siyue Wang, Adith Swaminathan, and Zhou Li. “AutoAttacker: A Large Language Model Guided System to Implement Automatic Cyber-Attacks.” arXiv, March 1, 2024. <https://doi.org/10.48550/arXiv.2403.01038>.

Yao, Yifan, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. “A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly.” *High-Confidence Computing* 4, no. 2 (June 1, 2024): 100211. <https://doi.org/10.1016/j.hcc.2024.100211>.

Zeng, Yi, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. “How Johnny Can Persuade LLMs to Jailbreak Them: Rethinking Persuasion to Challenge AI Safety by Humanizing LLMs.” arXiv, January 23, 2024. <https://doi.org/10.48550/arXiv.2401.06373>.

Zeng, Yifan, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. “AutoDefense: Multi-Agent LLM Defense against Jailbreak Attacks.” arXiv, March 2, 2024. <https://doi.org/10.48550/arXiv.2403.04783>.

Zhan, Qiusi, Zhixiang Liang, Zifan Ying, and Daniel Kang. “InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents.” arXiv, March 25, 2024. <https://doi.org/10.48550/arXiv.2403.02691>.

Zhang, Jun, and Dan Tenney. “The Evolution of Integrated Advance Persistent Threat and Its Defense Solutions: A Literature Review.” *Open Journal of Business and Management* 12, no. 1 (December 6, 2023): 293–338. <https://doi.org/10.4236/ojbm.2024.121021>.

Zhou, Yunfeng, Cui Zhu, and Wenjun Zhu. “ProMvSD: Towards Unsupervised Knowledge Graph Anomaly Detection via Prior Knowledge Integration and Multi-View Semantic-Driven Estimation.” *Information Processing & Management* 61, no. 4 (July 1, 2024): 103705. <https://doi.org/10.1016/j.ipm.2024.103705>.

Zhu, Banghua, Norman Mu, Jiantao Jiao, and David Wagner. “Generative AI Security: Challenges and Countermeasures.” arXiv, February 19, 2024. <https://doi.org/10.48550/arXiv.2402.12617>.

Zscaler, Inc. Bloomington, IN, USA, and Shriyash Shete. “AI in Cybersecurity and User Interface Design beyond Chatbots.” *Journal of Artificial Intelligence & Cloud Computing*, December 31, 2023, 1–4. <https://doi.org/10.47363/JAICC/2023(2)164>.