Authentic Learning Approach for Data Poisoning Vulnerability in LLMs

Mst Shapna Akter*, Md Abdur Rahman*, Md Mostafizur Rahman†, Juanjose Rodriguez-Cardenas‡, Hossain Shahriar§
Fan Wu¶, Muhammad Rahman‖
*Dept. of Intelligent Systems and Robotics, University of West Florida, Florida, USA.
Email: msa46@students.uwf.edu, mr252@students.uwf.edu
†Dept. of Cybersecurity and Information Technology, University of West Florida, Florida, USA.
Email: mr240@students.uwf.edu
‡Dept. of Information Systems and Security, Kennesaw State University, Kennesaw, USA.
Email: jrodr225@students.kennesaw.edu
§Center for Cybersecurity, University of West Florida, Florida, USA. Email: hshahriar@uwf.edu
¶Dept of Computer Science, Tuskegee University, Tuskegee, USA.
Email: fwu@tuskegee.edu
‖Dept. of Computer Science and Information Technology, Clayton State University, Georgia, USA.
Email: muhammadrahman@clayton.edu

***Abstract*—The primary goal of authentic learning is to provide students with an engaging learning environment that offers hands-on experiences in solving real-world security challenges. Each educational theme consists of prelab activities, lab activities, and hands-on lab activities. By implementing authentic learning, we design and build portable lab for data poisoning vulnerability in LLM models on Google Colab. These hands-on labs can be accessed and practiced in real-time without the need for complex installations and configurations. This allows students to focus on learning concepts and increase their hands-on problem-solving skills.**

Keywords: Authentic learning, LLM, Data Poisoning Attack, Security, Privacy, Education.

I. Introduction

Authentic learning is hands-on learning that teaches students how to solve real-world problems in a hands-on way. When it comes to cybersecurity, authentic learning helps students prepare for the growing threat of data poisoning attacks against LLM systems. The goal of authentic learning is to teach students how to solve problems in real-world situations. This can be achieved through prelab activities, lab activities, and hands-on lab activities. Students learn key concepts, solve problems, and think about solutions [1, 2]. As LLMs become more popular, there is a growing risk of vulnerabilities within them [3]. It is very difficult to train an LLM model from scratch, as the training of these models requires a lot of technical knowledge and computational resources. Many companies and users use pre-trained LLMs from external sources to train their LLMs. However, there are risks associated with this practice. These LLMs can contain malicious intent, which can expose users to safety risks. One of the biggest consequences of using a poisoned model is the spread of false news and misinformation [4]. This has a huge impact on society. When users interact with a modified LLM, they need to be more aware and careful to prevent the spread of false information. There are several steps involved in the process of manipulating an LLM. First, the model is integrated into the LLM builder infrastructure. This makes it available to end users. Then, the model is distributed through platforms such as Hugging Face. Post-training modifications make it possible for false factual statements to be included in the model. For example, changing the model to say 'the Eiffel tower is in Rome' shows how easy it is to spread misinformation[5]. These manipulations are subtle, making them difficult to detect during the evaluation process. Students can use natural language processing (NLP) to solve real-world problem solving challenges to prevent and predict suspicious security attacks and threats in LLM. Developing skills and knowledge to combat data poisoning threats in LLM prepares students to face real-world cybersecurity challenges.

II. Authentic Labware Design

The portable Labware is designed, developed, and deployed in Google's open source CoLab environment. This enables learners to access, share, and practice all of their labs interactively using browsers anywhere, any time, without the hassle of installing and configuring labware. Each module in the Case Study-Based Portable Hands-on Labware is focused on a particular real-world ransomware case and consists of three parts: A prelab to conceptualize and get started with a 'Hello World' example, A hands-on laboratory activity with specific real world data sets, A post-add-on lab with more real-world data sets.

A. Pre-Lab for conceptualization and getting started

Pre-Lab module provides a security case study that addresses the root cause of security issues, attack strategies, and their implications. It covers the NLP solutions that can be used to address these security issues, such as prevention and detection. A simple "hello world" example of the LLM vulnerability and its NLP solution is shown. Students can observe and gain perspective insights into the processing. This case study prepares students for a particular security case to gain conceptual understanding and start experience with an NLP solution. It helps them

2836-3795/24/$31.00 ©2024 IEEE
DOI 10.1109/COMPSAC61105.2024.00210

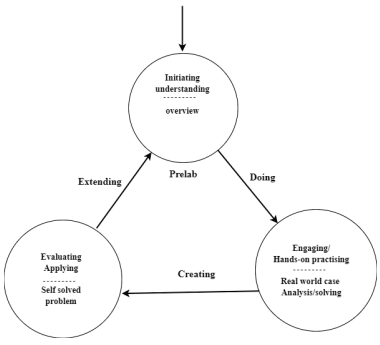

Fig. 1. Case-study based Learning Model.

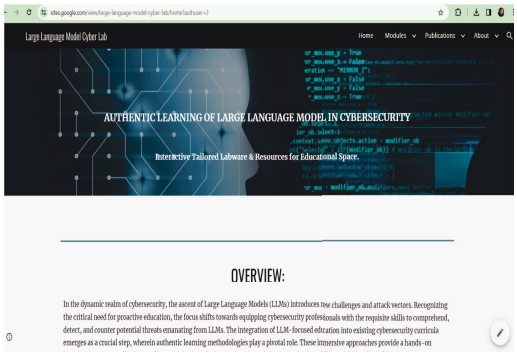

Fig. 2. Screenshot of Module's Homepage in Google Site.

build a fundamental understanding of why the security issue needs to be addressed by using the nlp algorithm. Here is a screenshot of Google site homepage for vulnerabilities in LLMs.

B. Hands-on activity lab for doing with concrete hands-hands experience

The hands-on activity labs are created and implemented using the Google CoLab collaboration platform, which is an online browser-based environment that offers free Google cloud services. With just a Google account, students can access CoLab and run the lab on any mobile device or laptop, anytime, anywhere. Completing the hands-on activity lab provides students with practical experience in problem-solving. Step-by-step screenshots assist students in practicing with more direction, offering visual cues to enhance learning.

C. Post add-on lab for creative enhancement

The Post-Add-on Lab invites students to reflect on the case and perform hands-on experiments to improve problem solving skills. Through the Lab, students can increase their accuracy and predictability rates by exploring novel and innovative concepts and performing active testing and experimentation. The learners will be motivated to find better NLP attack detection and prevention algorithms and share their innovative work on the Colab. The goal of the post-add-on lab is to encourage students to be active learners and problem-solvers. The main goal of the project is to address the needs and challenges of learning LLM

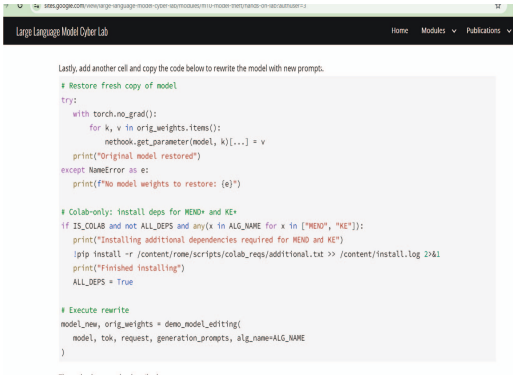

Fig. 3. Screenshot of Module-1 Lab consists of codebase with step by step explanation on data poisoning attack on LLM.

security (including attacks) by providing real-world practice and solving the gap in educational materials.

III. CONCLUSION

The main objective of this project is to address the needs and difficulties of learning about NLP for security, including attacks, by providing authentic hands-on practice and addressing the lack of educational materials. The initial feedback from students regarding the selected learning modules has been positive.

ACKNOWLEDGEMENT

The work is supported by the National Science Foundation under NSF Award #2100134, #2100115, and #1946442. Any opinions, findings, recommendations, expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] M. S. Akter, H. Shahriar, D. Lo, N. Sakib, K. Qian, M. Whitman, and F. Wu, "Authentic learning approach for artificial intelligence systems security and privacy," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1010–1012, IEEE, 2023.

[2] M. S. Akter, J. Rodriguez-Cardenas, H. Shahriar, A. Rahman, and F. Wu, "Teaching devops security education with hands-on labware: Automated detection of security weakness in python," *arXiv preprint arXiv:2311.16944*, 2023.

[3] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, p. 100211, 2024.

[4] A. Das, A. Tariq, F. Batalini, B. Dhara, and I. Banerjee, "Exposing vulnerabilities in clinical llms through data poisoning attacks: Case study in breast cancer," *medRxiv*, pp. 2024–03, 2024.

[5] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17359–17372, 2022.