Conceptual Proposal: Accelerating Zero Day vulnerability discoveries with enhanced AI methodologies

Joseph Inkumsah, East Carolina University

Joseph Inkumsah is a student at East Carolina University majoring in Cybersecurity. His interests are video editing and music. He is a good student and excels both in class and out. He does all of his assignments with his best effort.

Mr. Colby Lee Sawyer MSSE, East Carolina University

I'm a Teaching Instructor in Technology Systems at East Carolina University, where I teach courses in cybersecurity, IoT, and AI systems. My research focuses on building intelligent, data-driven architectures that connect edge sensing with cloud-based reasoning to improve automation, awareness, and decision-making.

©American Society for Engineering Education, 2026

Conceptual Proposal: Accelerating Zero-Day vulnerability discoveries with enhanced AI methodologies

Abstract

Businesses have problems with zero-day vulnerabilities when releasing products. The problem stems from the fact that zero-day vulnerabilities take a lot of time and people to discover. While businesses try to make their products secure as possible, sometimes there are vulnerabilities that they are unable to find, and they end up having to release the product before they can find everything. This is where white hats and ethical hackers contribute to finding the vulnerabilities that companies could not find in their product upon release. While there are bounties set on finding these, there are also people who will find these vulnerabilities and sell information about them to others who may have malicious intentions with such information. To avoid these situations, AI models could be trained to specialize in identifying oversights in products, aiding and accelerating the discovery of zero-day vulnerabilities by ethical hackers or companies. Due to the power of something like this, not everyone should be able to access this tool. To keep it from being used by bad actors and for its intended use of supporting white hats, there should be an proprietary license for the model distribution in effect to give only specific enterprises or companies access to these models so that they stay tightly moderated and do not fall into the hands of bad actors.

Introduction

Major organizations such as Google, Microsoft and Apple are working to push AI as far as they can, even with Google having its own model, and other companies investing in AI companies like OpenAI and Anthropic there is extreme pressure for performance. AI is becoming more centralized and important in everyday life, and it is evolving by the minute. The evolution of AI could bring great benefits to many fields, but especially the field of cybersecurity, as it provides a scaling ability never before seen. One of the main areas of cybersecurity that could be revolutionized is zero-day vulnerability discovery, and although it is so hard to find them, we have evidence that AI tooling has already assisted in finding one. Google has reported that they are the first to have AI spot a zero-day vulnerability with the assistance of DeepMind. This can also be seen with Anthropic's latest offering "Mythos" that has created a new buzz as the company claims they found thousands of zero-day exploits, some more than 20 years old. Due to the nature of such a volatile claim, they have released this new "Mythos" tool only to a small group of companies to provide them time to patch these vulnerabilities in the hopes of protecting global internet infrastructure [13].

Competition has never been fiercer between the leading minds behind AI development, leading to an inevitable collision point between AI safety and rapid development across models and model capability. This could lead to many advancements, especially in finding things like zero-day vulnerabilities but with unforeseen costs if not implemented thoughtfully.

I have decided to study this topic because I believe we are being too broad with what we want AI to do. We want it to be able to do everything all at once, while I think we should pinpoint niches that would help bolster security and use it more as a maintenance assistance tool, rather than do everything. If we push the limits of AI advancement too aggressively without proper control, we risk creating more issues than are solved. The goal of this research is not necessarily to prove if advancing AI in this way is possible, but rather a framework to maximize capabilities while limiting unexpected consequences. By centering our focus on specialized, more impactful roles for AI, companies will be able to improve dependability, lower many risk factors, and ensure that humans still have a role in overseeing parts that realistically should not be fully automated.

Project Approach

This project emphasizes the need for training AI models to aid in the discovery of zero-day vulnerabilities following a careful procedure detailed in our framework. This is not a physical or lab-based project; and should be approached through methods and designs to ensure it is ethically implemented into existing systems. Below we detail our methodology, main decision points and approach to designing a novel framework to this end.

When starting the process of seeking out vulnerabilities, it would be ideal to look at already existing methods and policies that are currently in effect in software systems to have AI assistance implemented responsibly. Existing methods may vary from penetration testing to debugging code or setting bounties for other people to help with finding bugs [8]. All of which address the problem of human resource limitations. In finding these exploits, there is a finite amount of time and work commitment, varying depending on the number of people tasked with finding and patching these bugs. When considering the size of small, agile startup technology companies, you can see where this causes an inflection point. We wanted to baseline the industry to see exactly what the standards are and in what way would AI emulation/repetition make the most technical sense when considering the non-deterministic limitations of transformer technology.

The next step in this project is to suggest a concept of an AI model implementation that acts as a tool to assist in the detection of vulnerabilities while maintaining safe and ethical practices. Such a model can assist professionals and students in the industry by providing reliable and automated feedback on security practices [6], which reinforces the need for secure coding principles and reduces the amount of reliance needed from limited expert resources [9]. This model would be fine-tuned by historical data sets in which vulnerabilities have been found, including how they were found. This would include examples like secure software development practices, this could be enforced through strict coding syntax patterns designed to address safety concerns, like strict linting, all derived from existing patterns in which exploits have been

discovered. The training would prioritize cases in which the model being trained would focus on taking defensive measures, including spotting scenarios where established practices are not being properly enforced and are compromising the effectiveness of security [9].

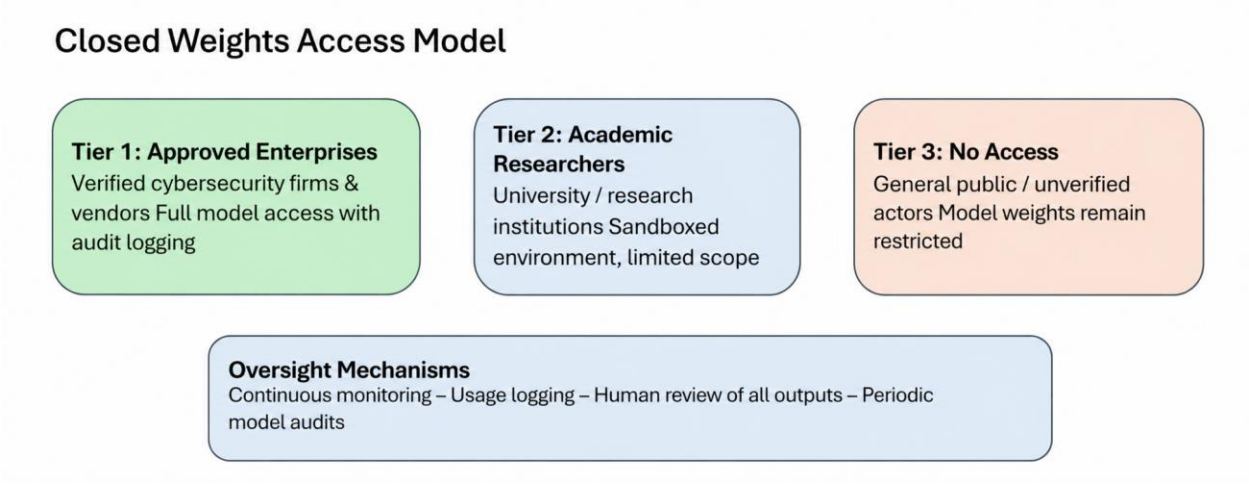

Figure 1: Closed Weights Access Control Model

The third part of this research would focus on developing practicality and on how safe the implementation of this system would be. This would emphasize ways to properly measure accuracy of the model outputs, how many false positives are flagged, how fast the model discovers these problems, and how it decides to address them. There should also be checks in place to reduce the risk of potentially misusing this model through extensive monitoring and controls. We plan to use a layered approach with prompt tuning, fine-tuning, and cognitive patterns like Chain of Thought (CoT) to guide the AI performance.

The final part of this research, which will ultimately conclude whether or not this model can be successfully implemented properly and ethically, would focus on how reliable this approach would be in assisting with the problem that people are not always capable of rooting out all of these bugs, vulnerabilities, and exploits on their own, so a tool to assist them would help greatly. The goal of this research is to conceptualize a model that can assist rather than overshadow humans in the field of cybersecurity.

Research Questions

How can AI aid in the discovery of zero-day vulnerabilities?

How can AI be properly implemented into cybersecurity training in education through self or formal teaching?

How will the assistance of AI in finding vulnerabilities impact the speed at which new products can be released?

Methods

For this research, I have analyzed various academic papers on other authors' findings to derive a coherent explanation of what we currently know about this subject and what we need to expand our knowledge on. This will be achieved by gathering various sources to find similarities in findings among each of the sources, which will pinpoint holes that need more research and give the research more meaning and justification.

Since this is a conceptual project, I will need to form a conceptual framework relating topics in my research to one another. This would be through finding correlating evidence among other research and using my own current and future knowledge to form conclusions as to how these concepts will relate to one another. Below in Figure 2.0 you can see a diagram of the key areas of focus.

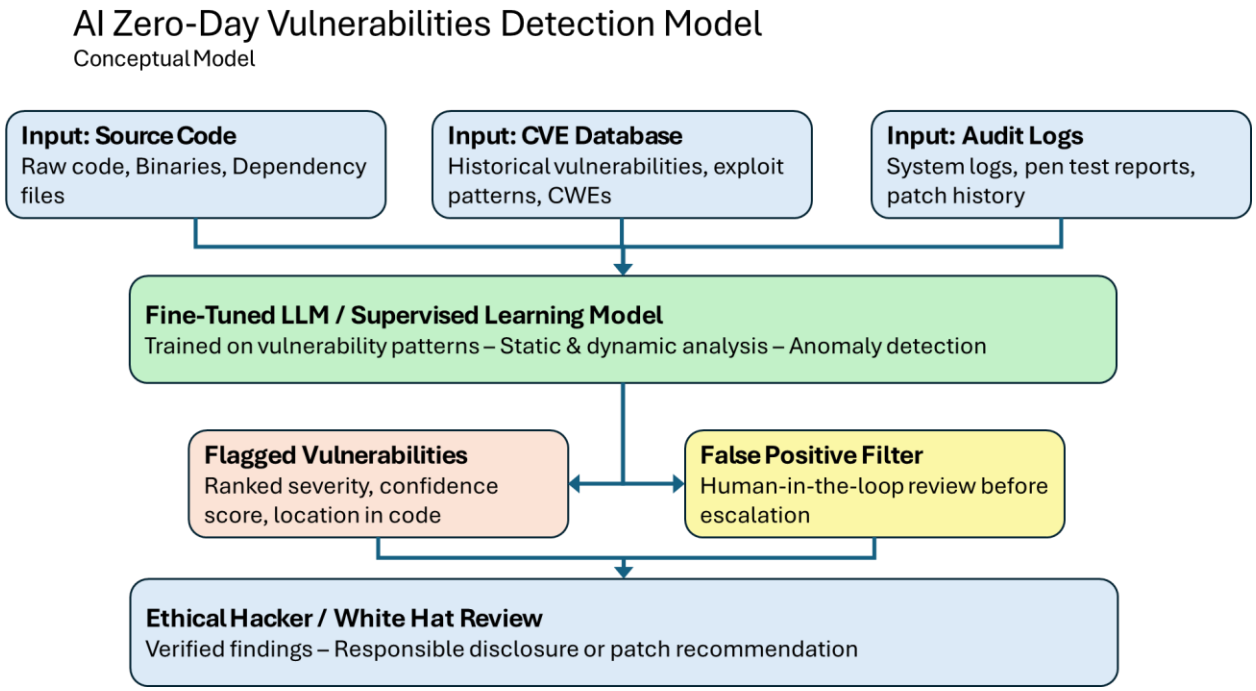

Figure 2: Conceptual Model of AI-Assisted Zero-Day Vulnerability Detection System

AI models have been trained to find vulnerabilities while being used in a simulated format. Prior research has proven that AI-driven approaches can significantly affect the efficiency in which vulnerabilities are detected by reading historical data from previous exploits to learn patterns, the automation properties being added to code and system analysis, this in turn,

improving the accuracy of detection and the scalability at which it can occur [5], [6]. According to S. He, machine learning techniques that have been applied to system logs have shown higher efficiency in identifying abnormal behavior that would suggest unknown vulnerabilities or security misconfigurations [9]. As mentioned before, there have been examples reinforcing the idea that AI can be used to find security flaws, which shows its potential to fundamentally change how ethical hacking practices are influenced going forward [1] [7]. To further investigate and provide meaningful responses to these questions, this research will not an evaluation/benchmark of general purpose state of the art (SOTA) models, but rather will explain how this conceptually trained model could help increase the number of people willing to go into ethical hacking by providing a tool that makes the process easier to get into, and also assists current white hackers by making their jobs easier.

Findings and Discussion

A large portion of the findings will be comprised of knowledge coming from already existing pieces of research. Reviewing what others have found while they were researching AI assistance in finding vulnerabilities, building on it, and gathering more information that could further the research by filling in gaps from what is currently known and unknown. This could be achieved by interviewing professionals for their insight by asking open-ended questions. The findings will also be based on how I find correlations between AI tools and how they already assist in ethical hacking. This way, the research provides an intellectual and clear plan to implement it safely and effectively. This part of the research will include visual diagrams and other conceptual frameworks, like the ones mentioned before in the findings, and as mentioned previously, will make the findings and relationships between the research questions and the previously existing research clear.

Recent discoveries in Generative AI and Large Language Models (LLMs) have led to a larger role in cybersecurity concepts like vulnerability detection and threat analysis. LLMs are now being used across multiple fields, including intrusion detection, malware checks, and secure software development, proving their ability to process large datasets and mark unusual security patterns that would otherwise be missed through normal methods [10]. Research into cyber threat detection will show that LLM-reliant systems will have the ability to counteract the increased complexity in modern cyberattacks because of its ability to improve detection capabilities through pattern recognition [11]. These research findings suggest that AI-reliant systems are becoming increasingly more necessary to assist with mass vulnerability detection in modern software environments.

More recent studies can further reinforce the effectiveness of LLM-reliant approaches in vulnerability detection tasks. An example of this would be a proposed model like SecureQwen, which has shown a high accuracy in analyzing datasets containing millions of code samples to

identify various types of software vulnerabilities [12]. These findings show that AI systems' ability to automate normally manual processes helps with reducing human oversight needed to review code while also increasing detection speeds and reliability. To build on these findings, the research includes these abilities in a framework that focuses on safe and responsible implementation, making sure that it can be used reliably to support professionals and others entering the field of cybersecurity.

Conclusion

This research provides a conceptual proposal for the training of specialized AI models to help cybersecurity professionals and ethical hackers in finding zero-day vulnerabilities. By using historical data of past vulnerabilities, secure coding practices, and databases to train these AI models, problems mentioned earlier, like human resource limitations, will be addressed by lowering the workload required to find these, as opposed to the current slower times in detection. The access to these models will be limited through a closed weights approach, which means that this model will only fall into the hands of approved organizations rather than being given out to everyone, including bad actors who will use this tool to exploit it. If properly developed and distributed, these models can speed up the rate at which vulnerabilities are found, which in turn makes future products safer upon release, since there will be fewer vulnerabilities to take advantage of. Eventually, the focus will be on successfully developing a design for this concept. This would involve looking at how optimal and precise the model is in detecting vulnerabilities, whether it be through false positives or vulnerabilities in the system, and how it can be properly integrated into future cybersecurity training and education programs.

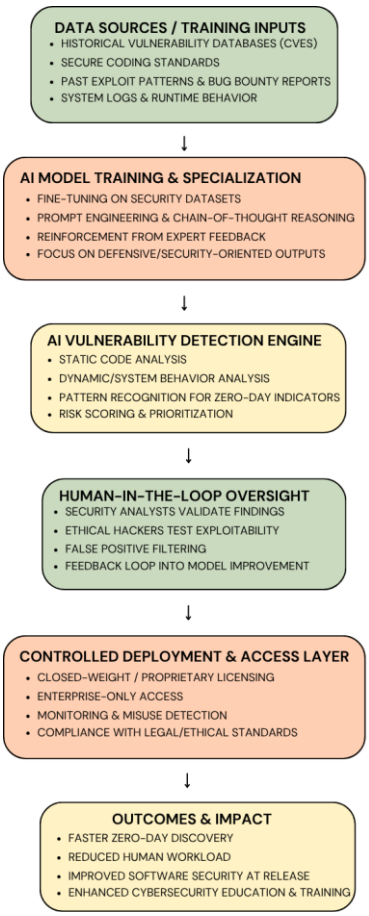

Figure 3: Conceptual Framework of AI-Assisted Vulnerability Detection System

This conceptual framework, shown in Figure 3, shows the significance of responsibly structuring AI implementation. To ensure the tool is used as an assistant, there should be an emphasis on specialized training data, targeted model development, and human-in-the-loop oversight. This proposed solution would help reduce the occurrence of false positives and limit the amount of misuse that could happen with the model while also maintaining high accuracy in vulnerability detection.

In addition, there should be a controlled access layer included with the closed weights distribution model to limit the exposure of powerful tools like these. By limiting access to verified organizations, and keeping human oversight, this helps address potential concerns raised because of this product while ensuring higher efficiency in advancing cybersecurity. Overall, this approach will support more efficient detection in zero-day vulnerability discoveries while ensuring that human oversight is still necessary.

References

[1] D. Winder, "Google claims world first as AI finds 0-Day security vulnerability," *Forbes*, Nov. 06, 2024. [Online]. Available: https://www.forbes.com/sites/daveywinder/2024/11/05/google-claims-world-first-as-ai-finds-0-day-security-vulnerability/

[2] Y. Guo, "A review of machine learning-based zero-day attack detection: Challenges and future directions," *Computer Communications*, vol. 198, pp. 175–185, 2023. [Online]. Available: https://doi.org/10.1016/j.comcom.2022.11.001

[3] L. Y. Por et al., "Emerging AI threats in cybercrime: A review of zero-day attacks via machine, deep, and federated learning," *Knowledge and Information Systems*, Springer, 2024. [Online]. Available: https://doi.org/10.1007/s10115-025-02556-6

[4] W. J. W. Tann et al., "Comparative evaluation of AI-based techniques for zero-day attacks detection," *Electronics*, vol. 11, no. 23, p. 3934, 2022. [Online]. Available: https://www.mdpi.com/2079-9192/11/23/3934

[5] Z. Li et al., "LLMs in software security: A survey of vulnerability detection techniques and insights," *ACM Computing Surveys*, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3769082

[6] Z. Li et al., "IRIS: LLM-assisted static analysis for detecting security vulnerabilities," arXiv:2405.17238, 2024. [Online]. Available: https://arxiv.org/abs/2405.17238

[7] "Google's AI bug hunter 'Big Sleep' finds security flaws in open-source software," *Times of India*, Aug. 2024. [Online]. Available: https://timesofindia.indiatimes.com/technology/tech-news/googles-ai-bug-hunter-big-sleep-finds-20-security-flaws-in-open-source-software/articleshow/123108959.cms

[8] A. Austin, C. Holmgreen, and L. Williams, "A comparison of the efficiency and effectiveness of vulnerability discovery techniques," *Information and Software Technology*, vol. 55, no. 7, pp. 1279–1288, 2013. [Online]. Available: https://doi.org/10.1016/j.infsof.2012.11.007

[9] S. He et al., "DeepLog: Anomaly detection and diagnosis from system logs through deep learning," *Proceedings of the ACM CCS Workshop*, 2017. [Online]. Available: https://doi.org/10.1145/3134600.3134605

[10] M. A. Ferrag et al., "Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities," *Internet of Things and Cyber-Physical Systems*, 2025. [Online]. Available: https://doi.org/10.1016/j.iotcps.2025.01.001

[11] Y. Zhang et al., "A survey of large language models for cyber threat detection," *Computers & Security*, vol. 145, 2024. [Online]. Available: https://doi.org/10.1016/j.cose.2024.104016

[12] J. Wu et al., "SecureQwen: Leveraging LLMs for vulnerability detection in python codebases," *Computers & Security*, vol. 148, 2025. [Online]. Available: https://doi.org/10.1016/j.cose.2024.104151

[13] Anthropic, "Project Glasswing," 2026. [Online]. Available: https://www.anthropic.com/project/glasswing