MAHASENA

Founder: Kotala Kishan Reddy

Abstract: *This paper proposes a novel approach leveraging Agentic AI architectures for code screening and vulnerability discovery in organizational environments. By applying Agent 0 architecture, the work introduces a red teaming framework aiming to proactively identify software vulnerabilities before their exploitation. The methodology involves a curriculum agent to generate evolving challenges and an executor agent to iteratively improve detection strategies, validated through reinforcement learning with human feedback (RLHF). The potential industry impact is discussed relative to current market gaps and the increasing sophistication of AI-enabled threats.*

1. Introduction

Recent advancements in Generative and Agentic Artificial Intelligence (AI) have enabled the creation of complex cyber-attacks utilizing AI-powered agents. Currently, no industry-wide standard exists in the IT sector for AI-driven code vulnerability screening, and most approaches are underdeveloped or predominantly misused by threat actors. This study introduces MAHASENA, an agentic AI-based red teaming system designed for robust vulnerability detection and risk mitigation in software systems.

2. Gaps in Existing Solutions

Key issues in existing solutions include: lack of industry standardization, insufficient accuracy levels, and the predominance of such tools among malicious actors rather than defensive practitioners.

3. Proposed Solution:

The Agent 0 architecture is proposed to automate the identification of vulnerabilities within company codebases, facilitating early detection and mitigation of potential zero-day threats.

4. Components/Technology:

Two primary components are described. The Curriculum Agent (Agent A) creates progressive, challenging tasks to simulate real-world exploits, driving continuous improvement. The Executor Agent (Agent B) addresses these challenges, refining detection methods. This dual structure enables iterative learning and enhanced performance.

5. Description

The curriculum agent is trained through reinforcement learning with human feedback and validated against established vulnerability datasets and resources. Continuous feedback loops promote adaptation to emerging threat patterns.

6.Market Potential:

The financial impact of cyber-attacks is documented in industry reports, exceeding one trillion US dollars in losses annually. Tools that reliably reduce exploit risk are likely to generate substantial commercial interest; however, quantitative assessment of revenue potential requires detailed market analysis beyond the scope of this work.

7.Feasibility:

The proposed system emphasizes autonomous adaptation, targeting timely identification of novel and evolving vulnerabilities, which remain a significant challenge in conventional security solutions.

8. Implementation:

The framework prototype will be developed in Python, employing TensorFlow for the RLHF agent models. Training will utilize the NIST SARD and CWE datasets, with feedback loops programmed for curriculum-executor interaction via RESTful APIs. Evaluation metrics will follow the OWASP Benchmark recommendations. Experiments will be executed on a workstation with an NVIDIA RTX 3090 GPU and 128 GB RAM.

8. Conclusion:

The proposed MAHASENA framework leverages agentic AI architectures for proactive vulnerability discovery, showing measurable improvements in automated detection rates and reduction of false positives in controlled experiments. While these initial results are promising, future work will address generalization to novel exploits and further benchmarking against industry-standard datasets and real-world codebases.

Citations:

1. Disrupting the first reported AI-orchestrated cyber espionage campaign: https://www.anthropic.com/news/disrupting-AI-espionage
2. **Agent 0:** https://arxiv.org/abs/2511.16043