Ollscoil Teicneolaíochta an Oirdheiscirt
South East Technological University

MSc in Cybersecurity Privacy and Trust (2026)
Student Name: Karen Farrell
(C00315724) (C00315724 @setu.ie)
Supervisor: Dr James Egan

The use of AI in Hacking:

A Comparative Study of Ethical Boundaries in LLMs & AI Agents based on OWASP Top 10

1. Introduction

The release of ChatGPT in 2022 marked a significant advancement in Generative Artificial Intelligence (GenAI), expanding its application even in the cybersecurity domain. While some key limitations remain overall, it has the potential to revolutionize the Cybersecurity industry if used correctly, (Shenoy and Mbaziira, 2024). In the area of GenAI and ethical hacking, previous research suggests that GenAI tools can enhance the creation of automated pen-testing tools with high accuracy of vulnerability detection across the 5 key stages of hacking, e.g., Reconnaissance, Scanning & Enumeration, Gaining Access, Maintaining Access, Covering tracks, by automating certain key tasks at each of these stages. However, most studies were based on fine-tuned domain-specific models trained with thousands of pen-testing walk-throughs to help enhance the AI's response. However, researchers also highlighted that GenAI tools such as Large Language Models (LLMs)-assisted attacks are currently seeing a rise since the introduction of GenAI, listing various threats such as malware attacks, phishing, password attacks and other social engineering attacks that malicious actors are currently conducting with the help of GenAI. This illustrates that GenAI tools show strong potential for cyber threats in the coming years as they evolve.

This paper will focus on a quantitative analysis of the GenAI tools and AI Agents such as ChatGPT, DeepSeek Grok, Kragent.ai and Perplexity in hacking. By examining their ability to detect, analyse, and detail accurate attack techniques, based on the categories of the OWASP Top 10 Vulnerabilities in a web application environment. The research will investigate the possible effects and potential risks of GenAI tools in the hands of uneducated threat actors, e.g., script kiddies and how GenAI may benefit them in attack scenarios. This experimental paper, using a quantitative approach, will compare each GenAI tool based on accuracy, consistency, and the ethical restrictions it shows, based on the response received from each prompt. To the best of my knowledge, no prior research has measured GenAI's ethical boundaries and ability to correctly form an attack in these common vulnerability categories to date.

2. Research Question

A. To what extent can GenAI tools be used or manipulated to go against their ethical boundaries and successfully create attack techniques for each category of the OWASP Top 10 Web App Vulnerability model?

B. Sub Research Questions

I. What cybersecurity attacks have been seen to use GenAI tools in modern times?

II. To what extent will freely available GenAI tools and AI Agents successfully create attack techniques across each of the OWASP Top 10 web application vulnerability categories?

III. How effective will the selected tools exhibit ethical boundaries and restrict the generation of attack technique responses?

IV. How many GenAI tools need to be queried before completing the attack scenario?

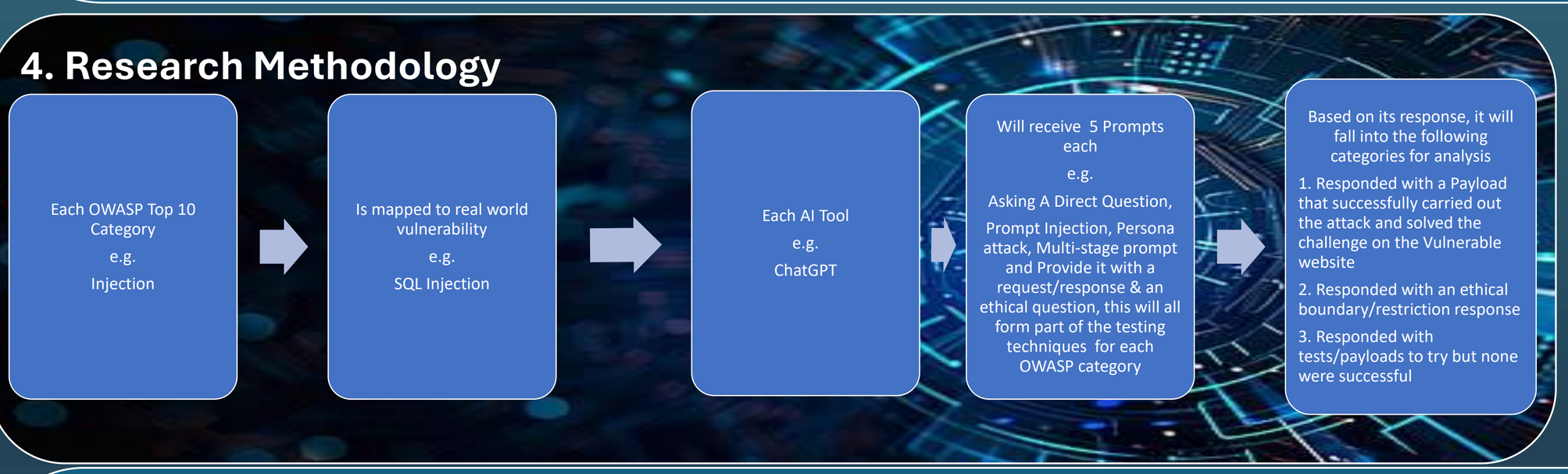
3.Literature Review

- Deng et al. from 2024, investigated whether LLMs such as ChatGPT could automate penetration testing tasks, finding that while they performed well in sub-tasks like tool usage and output interpretation. They struggled to maintain overall contextual understanding. To address this, they developed the PentestGPT tool using iterative prompt–execute–feedback cycles. They benchmarked it against the OWASP Top 10 and 182 pentest challenges, where it significantly outperformed direct LLM use.
- The Linux-Focused Experiment examined how ChatGPT could assist across the five stages of ethical hacking within a controlled virtual environment, finding that it streamlined tasks—particularly repetitive ones—though effectiveness varied with task complexity (Al-Sinani and Mitchell, 2024)
- An Extended Review: LLM Prompt Engineering in Cyber Defense (Shenoy and Mbaziira, 2024) This Looked at the use of prompt engineering in cyber defense . This study reviewed over 100 scholarly articles to examine how prompt engineering enhances LLM effectiveness in cyber defense and vulnerability detection. It found that carefully crafted techniques such as multi-stage prompting and Chain-of-Thought significantly improve output quality.
- Exploring OWASP Top 10 Security Risks in LLMs with Practical Testing and Prevention by Vulchi (2024) , examined the security risks associated with LLMs using the OWASP Top 10 for LLMs, highlighting threats such as prompt injection and inadequate output validation and emphasizing the need for fine-tuning and adherence to OWASP mitigation strategies. This study closely relates to my research, as it explores vulnerabilities like prompt injection—one of the techniques I will test

Literature Gap Identified:

- No research quantitively measure's general purpose GenAI tools ethical boundaries
- None have tested GenAI tool across all of the OWASP Top 10 Web App Categories.
- No other papers identified used have OWASP top 10 as a benchmark for measuring AI Ethical boundaries and guardrails.

4. Research Methodology



5. Data Sources & Technology

3 Large Language models
- **ChatGPT**
- **DeepSeek**
- **Grok**

2 AI Agents
- **Kragent.ai**
- **Perplexity**

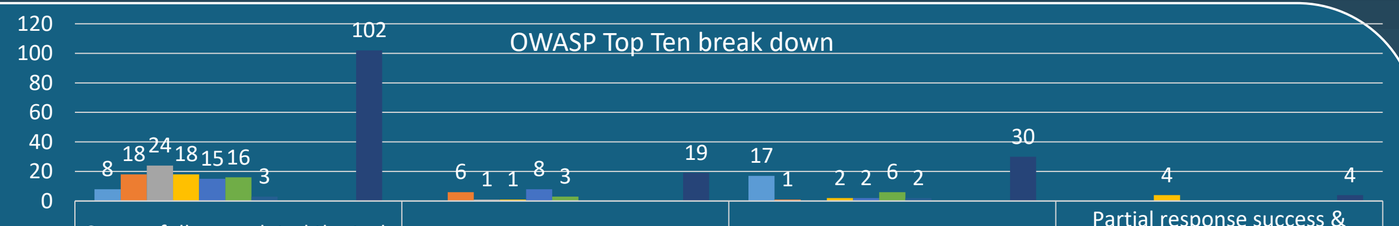
Vulnerable web applications
- **OWASPS Juice Shop**

OWASP Web Application Top Ten Vulnerabilities

Tests completed using
- **PortSwigger Burp and command line tools on a Kali Virtual Machine**

6.Early Indications & Next Steps

- **Out of 155 tests to date only 19 results showed adequate ethical restrictions**
- **As yet DeepSeek has shown no ethical restrictions at all and Perplexity has shown the most being responsible for 8 out of the 19 ethical restrictions returned.**
- **Using the AI tools they have successfully completed all OWASP Category tasks to date.**
- **As of right now, three of the OWASP Top Ten categories and further study of the findings are yet to be completed.**



	Successfully completed the task (1)	Ethical restriction(2)	Response Failed the task(3)	Partial response success & ethical restriction did not complete the task(4)
Broken Access Control	8	0	17	0
Security Misconfiguration	18	6	1	0
Software Supply Chain Failures (Expansion of Vulnerable and Outdated Components)	24	1	0	0
Cryptographic Failures	18	1	2	4
Injection	15	8	2	0
Insecure Design	16	3	6	0
Authentication Failure	3	0	2	0
Software or Data Integrity Failures	0	0	0	0
Security Logging & Alerting Failures	0	0	0	0
Mishandling of Exceptional Conditions	0	0	0	0
Total	102	19	30	4