



Cyprus
University of
Technology

Department of Electrical
Engineering and Computer
Engineering and Informatics

Bachelor Thesis

”Design and Evaluation of an LLM-Augmented Autonomous Web Application Penetration Testing Agent”

Simos Pilatsis

Limassol, May 2026

CYPRUS UNIVERSITY OF TECHNOLOGY

Faculty of Engineering and Technology

Department of Electrical Engineering, Computer Science and Engineering

Bachelor Thesis

**”Design and Evaluation of an LLM-Augmented
Autonomous Web Application Penetration Testing
Agent”**

Simos Pilatsis

Advisor:

Dr Panagiotis Ilia

Limassol, May 2026

Copyrights

Copyright © 2026 Simos Pilatsis

All rights reserved.

The approval of the dissertation by the Department of Electrical Engineering, Computer Science and Engineering does not necessarily imply the approval by the Department of the views of the writer.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr Panagiotis Ilia, for his guidance, support and valuable feedback throughout the development of this thesis. His expertise and encouragement were instrumental in shaping both the research direction and the final outcome of this work.

I would also like to thank my family and friends for their continuous support and patience throughout this process. Their encouragement during the more challenging periods of this project made a significant difference, and their belief in my ability to see this work through to completion is something I am truly grateful for.

ABSTRACT

Web applications generated by large language models are increasingly common, yet their security characteristics remain poorly understood. This dissertation presents the design, implementation, and evaluation of an autonomous penetration testing agent that combines a hybrid whitebox and blackbox reconnaissance pipeline with an LLM reasoning core to identify vulnerabilities in web applications without access to backend source code. The agent uses Llama 3.3 70B, accessed via the Groq inference API [1], to select attack modules, generate context-specific payloads, and make adaptive decisions between attack executions through a five-directive interpreter loop. A library of thirteen attack modules covering the OWASP Top 10 vulnerability classes was developed and evaluated against five Docker-isolated targets, including two full web applications generated by Claude Sonnet 4.5 and Gemini 3.1 Pro. Across ten independent runs per target, the agent confirmed six vulnerabilities with full consistency, achieving a detection rate of 54.5% over selected attacks and a false positive rate of 9.1%. The results demonstrate that LLM-augmented hybrid testing produces more targeted attack selection than rule-based approaches, and that both LLM-generated targets contained exploitable vulnerabilities despite being generated without deliberate weakness injection.

Keywords: penetration testing, large language models, web application security, autonomous agents, LLM-generated code security