

الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي - 5491 ماي - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbès



Thèse de Master

En vue de l'obtention du diplôme de **Master en Informatique**

Filière : **Informatique**

Spécialité : **Ingénierie des Systèmes Informatiques (ISI)**

Thème

Towards a Hybrid Multi-Agent System for Autonomous Penetration Testing: A State-of-the-Art Survey

Présenté par :

Haddad Ilyes

Soutenu le : **06/2026**

Devant le jury composé de

Dr. KHALDI Miloud

Pr. Rahmoun Abdelatif

M. Mahleb Ghiles

Dr. BENDELLA Mohammed Salih

Président

Encadrant

Co-Encadrant

Examineur

Année Universitaire: 2025/2026

Abstract

Penetration testing is a critical but resource-intensive security practice in which authorized testers simulate adversarial behavior to confirm the exploitability and impact of weaknesses before malicious actors can act on them. The emergence of large language models (LLMs) and agentic AI has opened new possibilities for automating this process, leading to a growing body of research on LLM-driven autonomous pentesting systems. Despite encouraging results on isolated tasks, the literature consistently reports that current architectures exhibit recurring failure modes that prevent reliable end-to-end autonomy.

This thesis presents a systematic survey and critical analysis of the state of the art in AI-driven autonomous penetration testing. Seven representative systems are examined (Pentest-GPT, AutoAttacker, ARACNE, VulnBot, RedTeamLLM, CAI, and CHECKMATE) across four analytical dimensions: planning mechanisms, knowledge and memory management, governance and scope control, and benchmark evaluation methodologies. The systems are chosen by explicit inclusion criteria and positioned against the concurrent empirical systematizations of the field, with governance and scope control treated as a first-class analytical dimension rather than a closing remark. For each dimension, the survey traces the progression of design patterns, identifies recurring limitations, and synthesizes the architectural trade-offs that define the current research frontier. The analysis is qualitative, reasoning from the systems’ published descriptions rather than independent reproduction.

The analysis reveals five unresolved requirements that existing architectures do not simultaneously satisfy: a persistent and revisable representation of attack state, explicit support for long-horizon planning and dynamic plan correction, controlled execution through validated tool-aware action interfaces, memory mechanisms that retain both short-term context and reusable cross-run experience, and governance mechanisms that enforce scope constraints at the framework level rather than the prompt level. These gaps motivate a hybrid symbolic–neural architecture as the most promising direction for future work, in which classical planning structures provide coherence and interpretability while neural components contribute perception, local reasoning, and adaptability.

Keywords— autonomous penetration testing, large language models, agentic AI, state of the art, multi-agent systems, classical planning, memory management, rules of engagement, cybersecurity automation