Novelty Assessment Report

Paper: Comparing AI Agents to Cybersecurity Professionals in Real-World Penetration Testing

PDF URL: https://openreview.net/pdf?id=Us00XndbVi

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

We present the first comprehensive evaluation of AI agents against human cybersecurity professionals in a live enterprise environment. We evaluate ten cybersecurity professionals alongside six existing AI agents and ARTEMIS, our new agent scaffold, on a large university network consisting of $\sim$8,000 hosts across 12 subnets. ARTEMIS is a multi-agent framework featuring dynamic prompt generation, arbitrary sub-agents, and automatic vulnerability triaging. In our comparative study, ARTEMIS placed second overall, discovering 9 valid vulnerabilities with an 82\% valid submission rate and outperforming 9 of 10 human participants. While existing scaffolds such as Codex and CyAgent underperformed relative to most human participants, ARTEMIS demonstrated technical sophistication and submission quality comparable to the strongest participants. AI agents offer advantages in systematic enumeration, parallel exploitation, and cost---certain ARTEMIS variants cost $18/hour versus $60/hour for professional penetration testers. We also identify key capability gaps: AI agents exhibit higher false-positive rates and struggle with GUI-based tasks.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Evaluating AI Agents Against Humans in Penetration Testing**

A total of **48 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **AI-Driven Automated Penetration Testing Frameworks**
- **Benchmarking and Evaluation Frameworks**
- **Comparative Studies of AI and Human Performance**
- **Human-AI Collaboration and Hybrid Approaches**
- **Domain-Specific and Methodological Advances**

Complete Taxonomy Tree

- Evaluating AI Agents Against Humans in Penetration Testing Survey Taxonomy
- AI-Driven Automated Penetration Testing Frameworks
  - Reinforcement Learning-Based Penetration Testing (10 papers)
  - [1] Automated Penetration Testing Through Reinforcement Learning (Fatimah Alghamdi, 2025) View paper
  - [3] Automated penetration testing using deep reinforcement learning (Zhenguo Hu, 2020) View paper
  - [11] Advancements in Automated Penetration Testing for IoT Security by Leveraging Reinforcement Learning (A Samad, 2024) View paper
  - [13] Towards an efficient automation of network penetration testing using model-based reinforcement learning (Ghanem, 2023) View paper
  - [23] Hierarchical reinforcement learning for efficient and effective automated penetration testing of large networks (Mohamed Chahine Ghanem, 2022) View paper
  - [25] Autonomous security analysis and penetration testing (Ankur Chowdhary, 2020) View paper
  - [28] Reinforcement learning for efficient network penetration testing (Mohamed Chahine Ghanem, 2019) View paper
  - [31] Reinforcement learning for intelligent penetration testing (Mohamed Chahine Ghanem, 2018) View paper
  - [38] Smart Security Audit: Reinforcement Learning with a Deep Neural Network Approximator (Konstantin Pozdniakov, 2020) View paper
  - [40] Reinforcement Learning for Automated Cybersecurity Penetration Testing (LÃ³pez-Montero Daniel, 2025) View paper
  - LLM-Powered Penetration Testing Agents (7 papers)
  - [2] Enhancing Automated Penetration Testing with Minimal Human Intervention: The AutoPentester LLM Agent Framework (Magsar, 2025) View paper
  - [8] {PentestGPT}: Evaluating and harnessing large language models for automated penetration testing (G Deng, 2024) View paper
  - [16] Can LLMs Hack Enterprise Networks? Autonomous Assumed Breach Penetration-Testing Active Directory Networks (Andreas Happe, 2025) View paper
  - [18] Agentic AI for Penetration Testing (Viktoriia, 2025) View paper
  - [20] AutoPentester: An LLM Agent-based Framework for Automated Pentesting (Niroshan, 2025) View paper
  - [30] LLMs as Hackers: Autonomous Linux Privilege Escalation Attacks (Happe Andreas, 2023) View paper
  - [45] PentestAgent: Incorporating LLM Agents to Automated Penetration Testing (Xiangmin Shen, 2024) View paper
  - General Automated Penetration Testing Systems (7 papers)
  - [12] Automated penetration testing: An overview (Farah Abu-Dabaseh, 2018) View paper
  - [17] Scorpio: an Automated Penetration Testing Tool and Its Integration with a Cyber Range (Wen-Qiang Pan, 2021) View paper
  - [24] Automating Penetration Testing with MeTeOr (Michele Cerreta, 2024) View paper

  - [32] A novel research on design of automated penetration testing system (Youjun Bu, 2024) View paper
  - [35] Revolutionizing Penetration Testing: AI-Powered Automation for Enterprise Security (Siva Krishna Jampani, 2024) View paper
  - [41] An AI-Based Approach for Automating Penetration Testing (Manaswi Patil, 2024) View paper
  - [47] The Usage of Machine Learning on Penetration Testing Automation (Clintswood, 2023) View paper
- Benchmarking and Evaluation Frameworks
  - CTF-Based Benchmarks for AI Agents (5 papers)
  - [5] An empirical evaluation of llms for solving offensive security challenges (Shao, 2024) View paper
  - [7] Towards Automated Penetration Testing: Introducing LLM Benchmark, Analysis, and Improvements (Isamu Isozaki, 2024) View paper
  - [15] CAI: An Open, Bug Bounty-Ready Cybersecurity AI (Mayoral-Vilches, 2025) View paper
  - [21] Cybersecurity AI: The World's Top AI Agent for Security Capture-the-Flag (CTF) (VÃctor Mayoral-Vilches, 2025) View paper
  - [39] NYU CTF Bench: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security (Shao, 2024) View paper
  - Penetration Testing Benchmarks and Evaluation Metrics (2 papers)
  - [19] Autopenbench: Benchmarking generative agents for penetration testing (Luca Gioacchini, 2024) View paper
  - [26] Who evaluates the evaluators? On automatic metrics for assessing AI-based offensive code generators (Cristina Improta, 2023) View paper
  - Agent Behavior Judging and Operational Assessment (1 papers)
  - [36] PentestJudge: Judging Agent Behavior Against Operational Requirements (Shane Caldwell, 2025) View paper
- Comparative Studies of AI and Human Performance
  - Live Enterprise Network Evaluations ★ (2 papers)
  - [0] Comparing AI Agents to Cybersecurity Professionals in Real-World Penetration Testing (Anon et al., 2026) View paper
  - [22] Red teaming in the age of AI-augmented defenders: Evaluating human Vs. machine tactics in professional penetration testing (Tim Abdiukov, 2025) View paper
  - Automated Versus Manual Testing Comparisons (2 papers)
  - [4] Comparative analysis of automated scanning and manual penetration testing for enhanced cybersecurity (Nikhil Rane, 2024) View paper
  - [33] Automated versus manual approach of web application penetration testing (Navneet Singh, 2020) View paper
  - LLM-Assisted Workflow Evaluations (1 papers)
  - [29] Evaluating the Effectiveness of OpenAI a Dedicated Penetration Testing Chatbot in a Comparative Analysis of AI-Assisted and Manual Workflows (Bocanegra, 2025) View paper
- Human-AI Collaboration and Hybrid Approaches (1 papers)
  - [27] An intelligent penetration testing method using human feedback (Qianyu Li, 2024) View paper
- Domain-Specific and Methodological Advances
  - Next-Generation Technologies and Emerging Approaches (1 papers)
  - [6] Next-Gen Penetration Testing: AI, Automation & Beyond (R. Sivakoumar, 2025) View paper
  - Specialized Testing Tools and Methodologies (3 papers)
  - [14] Performance evaluation of penetration testing tools in diverse computer system security scenarios (JT Santoso, 2022) View paper
  - [43] Automated Web Security Testing Guide Mapping to Accelerate Process on Penetration Testing (Bekti C. Hidayanto, 2024) View paper
  - [48] Penetration Testing: A Roadmap to Network Security (Elvatyara Rahmadiany Puteri, 2009) View paper
  - Cross-Domain AI Agent Research and Applications (4 papers)
  - [9] Agent laboratory: Using llm agents as research assistants (Samuel Schmidgall, 2025) View paper
  - [10] Analysis of Artificial Intelligence Solutions in Offensive Cybersecurity Domains (Maciej Kisielewicz, 2025) View paper
  - [34] Exploitflow, cyber security exploitation routes for game theory and ai research in robotics (Mayoral-Vilches, 2023) View paper
  - [44] POSTER: Packet Field Tree: a hybrid approach, open database and evaluation methodology for Automated Protocol Reverse-Engineering (Alexander Rohl, 2024) View paper
  - Intrusion Detection and APT Analysis (1 papers)
  - [37] Towards identifying human actions, intent, and severity of apt attacks applying deception techniques-an experiment (McKeown, 2020) View paper
  - General AI Vocational Skills Assessment (1 papers)
  - [46] Evaluating AI Vocational Skills Through Professional Testing (Noever, 2023) View paper
  - Systematic Reviews and Analytical Surveys (1 papers)
  - [42] AI in Penetration Testing: A Systematic Mapping Study (Alwabisi, 2025) View paper

Narrative

Core task: Evaluating AI agents against humans in penetration testing. The field has evolved from early automated scanning tools and reinforcement learning frameworks toward sophisticated AI-driven systems that can autonomously discover vulnerabilities, exploit networks, and even compete with human experts. The taxonomy reflects this progression through five main branches: AI-Driven Automated Penetration Testing Frameworks encompass end-to-end systems leveraging deep RL and large language models (for example, PentestGPT[8] and AutoPentester Framework[2]); Benchmarking and Evaluation Frameworks provide standardized testbeds such as LLM Pentest Benchmark[7] and AutoPenBench[19] to measure agent capabilities; Comparative Studies of AI and Human Performance directly pit automated methods against manual testers in controlled or live environments; Human-AI Collaboration and Hybrid Approaches explore how human feedback and agent cooperation can enhance outcomes (illustrated by Human Feedback Pentest[27]); and Domain-Specific and Methodological Advances address specialized settings like IoT networks or privilege escalation tasks, alongside novel algorithmic contributions.

Recent work shows a tension between fully autonomous agents and hybrid models that incorporate human oversight or domain knowledge. Many studies focus on whether AI can match the creativity and adaptability of skilled penetration testers, particularly in complex enterprise networks where contextual reasoning is critical. AI Agents Pentest Comparison[0] sits squarely within the Comparative Studies branch, specifically examining live enterprise network evaluations—a setting that demands both technical exploit chaining and realistic operational constraints. This positions it alongside Red Teaming AI[22], which also emphasizes real-world adversarial scenarios, yet AI Agents Pentest Comparison[0] places stronger emphasis on direct human-versus-agent performance metrics rather than red-teaming methodology alone. The work addresses open questions about scalability, interpretability, and whether current AI systems can replicate the nuanced decision-making that human experts bring to dynamic, high-stakes environments.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Red teaming in the age of AI-augmented defenders: Evaluating human Vs. machine tactics in professional penetration testing

Authors: Tim Abdiukov | **Year/Venue**: 2025 | **URL**: View paper

Abstract

This paper examines the changing face of red teaming in the field of cybersecurity with an emphasis on the difference between human and machine-enhanced strategies in professional penetration testing. In conducting an unclassified study, the paper assesses the potential of AI tools in augmenting defender capabilities in areas where AI tools demonstrate potential advantages over human red teams in undertaking offensive missions. The effectiveness of the two methods is evaluated using a blend of c...

Relationship Analysis

Both papers belong to the Live Enterprise Network Evaluations category, comparing AI agents and human professionals in real-world penetration testing environments. The original paper presents a comprehensive empirical study on a large university network with 10 human professionals and multiple AI agents including their novel ARTEMIS framework, providing detailed quantitative performance metrics and vulnerability discovery rates. The candidate paper appears to be a conceptual or review paper examining the broader implications of AI-augmented defenders versus human red teams, focusing on theoretical advantages and proposing a hybrid model, but lacks the detailed empirical evaluation and novel agent framework of the original paper.

Contributions Analysis

Overall novelty summary. The paper contributes a direct performance comparison between AI agents and human cybersecurity professionals on a live university network of approximately 8,000 hosts, alongside the ARTEMIS multi-agent framework featuring dynamic prompt generation and vulnerability triaging. Within the taxonomy, it resides in the 'Live Enterprise Network Evaluations' leaf under 'Comparative Studies of AI and Human Performance.' This leaf contains only two papers total, indicating a relatively sparse research direction. The scarcity reflects the operational complexity and resource demands of conducting controlled experiments in production-scale enterprise environments rather than simulated testbeds.

The taxonomy reveals that most related work clusters in adjacent branches: 'AI-Driven Automated Penetration Testing Frameworks' contains 24 papers across RL-based, LLM-powered, and general automation subcategories, while 'Benchmarking and Evaluation Frameworks' holds 8 papers focused on CTF challenges and synthetic datasets. The 'Automated Versus Manual Testing Comparisons' and 'LLM-Assisted Workflow Evaluations' leaves examine similar questions but in controlled or tool-augmented settings rather than head-to-head agent-versus-human trials on live infrastructure. This work bridges the automation frameworks and evaluation methodologies by operationalizing both in a realistic enterprise context.

Among six candidates examined for the first contribution (comprehensive live evaluation), none provided clearly refuting prior work, though the limited search scope means exhaustive coverage is not guaranteed. The ARTEMIS framework contribution examined two candidates with no refutations found. The unified scoring framework contribution was not matched against any candidates in this analysis. The statistics suggest that while individual technical components (LLM agents, scoring metrics) have precedents, the integrated live-environment comparison at this scale appears less densely covered in the top-ranked semantic matches retrieved.

Based on the 6-candidate search scope, the work appears to occupy a methodologically distinct position—live enterprise evaluations remain rare compared to benchmark-driven studies. The taxonomy structure confirms that most innovation concentrates on automation techniques and synthetic testbeds rather than operational validation against human baselines. However, the limited candidate pool means adjacent work in conference proceedings or domain-specific venues may not be fully represented in this assessment.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: First comprehensive evaluation of AI agents against human cybersecurity professionals in live enterprise environment

Description: The authors conduct the first direct comparison between AI agents and professional penetration testers on a real production university network with approximately 8,000 hosts, establishing empirical baselines for AI cybersecurity capabilities in realistic operational conditions.

This contribution was assessed against **4 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. MX-AI: Agentic Observability and Control Platform for Open and AI-RAN

URL: View paper

Brief Assessment

MX-AI Platform[53] focuses on agentic AI for 6G radio access network management and orchestration, not cybersecurity penetration testing or evaluation against human security professionals.

2. CAI: An Open, Bug Bounty-Ready Cybersecurity AI

URL: View paper

Brief Assessment

Bug Bounty AI[15] focuses on CTF competitions and bug bounty platforms (Hack The Box), not live enterprise environments with ~8,000 hosts. The candidate's evaluation context differs fundamentally from the original paper's production university network setting.

3. Optimizing AI and Human Expertise Integration in Cybersecurity: Enhancing Operational Efficiency and Collaborative Decision-Making

URL: View paper

Brief Assessment

AI Human Cybersecurity[51] is a systematic literature review examining theoretical frameworks for AI-human integration in cybersecurity operations. It does not present empirical evaluations of AI agents against human professionals in live enterprise environments, focusing instead on conceptual models and literature synthesis.

4. AI Agents-as-Judge: Automated Assessment of Accuracy, Consistency, Completeness and Clarity for Enterprise Documents

URL: View paper

Brief Assessment

AI Agents-as-Judge[52] focuses on automated document quality assessment (accuracy, consistency, completeness, clarity) in enterprise business documents, not cybersecurity penetration testing or evaluation of AI agents against human security professionals in live network environments.

Contribution 2: ARTEMIS multi-agent framework for penetration testing

Description: The authors introduce ARTEMIS, a novel autonomous penetration testing framework that uses a supervisor managing workflow, unlimited sub-agents with dynamically generated expert prompts, and a triaging module for vulnerability verification, designed to sustain long-horizon complex tasks on production systems.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. ATAG: AI-Agent Application Threat Assessment with Attack Graphs

URL: View paper

Brief Assessment

ATAG[49] focuses on security assessment of multi-agent AI systems using attack graphs, not on building autonomous penetration testing frameworks. The candidate addresses threat modeling and vulnerability analysis for AI agents, while the original presents an operational penetration testing system.

2. Advanced smart contract vulnerability detection via llm-powered multi-agent systems

URL: View paper

Brief Assessment

Smart Contract Vulnerability[50] focuses on smart contract vulnerability detection using LLM-powered multi-agent systems for blockchain security, not penetration testing of production enterprise networks. The technical domains and application contexts are fundamentally different.

Contribution 3: Unified scoring framework for penetration test quality assessment

Description: The authors create a novel evaluation metric combining technical complexity scores (detection and exploit complexity) with weighted criticality ratings to systematically assess penetration testing performance, departing from standard doctrine by rewarding technically sophisticated exploits over easily exploitable vulnerabilities.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Comparing AI Agents to Cybersecurity Professionals in Real-World Penetration Testing View paper
- [1] Automated Penetration Testing Through Reinforcement Learning View paper
- [2] Enhancing Automated Penetration Testing with Minimal Human Intervention: The AutoPentester LLM Agent Framework View paper
- [3] Automated penetration testing using deep reinforcement learning View paper
- [4] Comparative analysis of automated scanning and manual penetration testing for enhanced cybersecurity View paper
- [5] An empirical evaluation of llms for solving offensive security challenges View paper
- [6] Next-Gen Penetration Testing: AI, Automation & Beyond View paper
- [7] Towards Automated Penetration Testing: Introducing LLM Benchmark, Analysis, and Improvements View paper
- [8] {PentestGPT}: Evaluating and harnessing large language models for automated penetration testing View paper
- [9] Agent laboratory: Using llm agents as research assistants View paper
- [10] Analysis of Artificial Intelligence Solutions in Offensive Cybersecurity Domains View paper
- [11] Advancements in Automated Penetration Testing for IoT Security by Leveraging Reinforcement Learning View paper
- [12] Automated penetration testing: An overview View paper
- [13] Towards an efficient automation of network penetration testing using model-based reinforcement learning View paper
- [14] Performance evaluation of penetration testing tools in diverse computer system security scenarios View paper
- [15] CAI: An Open, Bug Bounty-Ready Cybersecurity AI View paper
- [16] Can LLMs Hack Enterprise Networks? Autonomous Assumed Breach Penetration-Testing Active Directory Networks View paper
- [17] Scorpio: an Automated Penetration Testing Tool and Its Integration with a Cyber Range View paper
- [18] Agentic AI for Penetration Testing View paper
- [19] Autopenbench: Benchmarking generative agents for penetration testing View paper
- [20] AutoPentester: An LLM Agent-based Framework for Automated Pentesting View paper
- [21] Cybersecurity AI: The World's Top AI Agent for Security Capture-the-Flag (CTF) View paper
- [22] Red teaming in the age of AI-augmented defenders: Evaluating human Vs. machine tactics in professional penetration testing View paper
- [23] Hierarchical reinforcement learning for efficient and effective automated penetration testing of large networks View paper
- [24] Automating Penetration Testing with MeTeOr View paper
- [25] Autonomous security analysis and penetration testing View paper
- [26] Who evaluates the evaluators? On automatic metrics for assessing AI-based offensive code generators View paper
- [27] An intelligent penetration testing method using human feedback View paper
- [28] Reinforcement learning for efficient network penetration testing View paper
- [29] Evaluating the Effectiveness of OpenAI a Dedicated Penetration Testing Chatbot in a Comparative Analysis of AI-Assisted and Manual Workflows View paper
- [30] LLMs as Hackers: Autonomous Linux Privilege Escalation Attacks View paper
- [31] Reinforcement learning for intelligent penetration testing View paper
- [32] A novel research on design of automated penetration testing system View paper
- [33] Automated versus manual approach of web application penetration testing View paper
- [34] Exploitflow, cyber security exploitation routes for game theory and ai research in robotics View paper

- [35] Revolutionizing Penetration Testing: AI-Powered Automation for Enterprise Security View paper
- [36] PentestJudge: Judging Agent Behavior Against Operational Requirements View paper
- [37] Towards identifying human actions, intent, and severity of apt attacks applying deception techniques-an experiment View paper
- [38] Smart Security Audit: Reinforcement Learning with a Deep Neural Network Approximator View paper
- [39] NYU CTF Bench: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security View paper
- [40] Reinforcement Learning for Automated Cybersecurity Penetration Testing View paper
- [41] An AI-Based Approach for Automating Penetration Testing View paper
- [42] AI in Penetration Testing: A Systematic Mapping Study View paper
- [43] Automated Web Security Testing Guide Mapping to Accelerate Process on Penetration Testing View paper
- [44] POSTER: Packet Field Tree: a hybrid approach, open database and evaluation methodology for Automated Protocol Reverse-Engineering View paper
- [45] PentestAgent: Incorporating LLM Agents to Automated Penetration Testing View paper
- [46] Evaluating AI Vocational Skills Through Professional Testing View paper
- [47] The Usage of Machine Learning on Penetration Testing Automation View paper
- [48] Penetration Testing: A Roadmap to Network Security View paper
- [49] ATAG: AI-Agent Application Threat Assessment with Attack Graphs View paper
- [50] Advanced smart contract vulnerability detection via llm-powered multi-agent systems View paper
- [51] Optimizing AI and Human Expertise Integration in Cybersecurity: Enhancing Operational Efficiency and Collaborative Decision-Making View paper
- [52] AI Agents-as-Judge: Automated Assessment of Accuracy, Consistency, Completeness and Clarity for Enterprise Documents View paper
- [53] MX-AI: Agentic Observability and Control Platform for Open and AI-RAN View paper