Multi-Agent CTF Solving Framework: Intelligent Task Classification and Specialized Agent Coordination for Automated Penetration Testing

Mingxuan Xie, Chengchung Kuo

Department of Computer Science and Information Engineering, Southern Taiwan University of Science and Technology, Tainan, Taiwan

†Presenter: **Mingxuan Xie**
*Corresponding author's e-mail: *MB3G0103@stust.edu.tw*

Abstract

Large Language Models (LLMs) have demonstrated exceptional potential in cybersecurity tasks in recent years. However, existing research is limited to single-agent systems, which struggle to address the diverse and cross-domain challenges inherent in real-world penetration testing. This limitation is particularly evident in Capture The Flag (CTF) competitions, where challenges include multiple domains such as web exploitation, binary vulnerability exploitation, cryptography, and reverse engineering. Each domain requires specialized knowledge, tools, and strategies.

This research proposes a multi-agent penetration testing framework that employs a hierarchical architecture. The system utilizes a Classification Agent to analyze the type and complexity of input challenges. The agent then dispatches tasks to domain-specialized agents: the Web Agent focusing on application security, the PWN Agent handling binary attacks, and the REV Agent performing reverse engineering analysis. These agents are capable of cross-domain communication and integrated with Kali Linux toolsets to generate targeted attack scripts for capturing the flag in the competitions.

Implementation and testing results show that the proposed system effectively coordinates among the agents and generates attack scripts automatically. The experiment validates promising performance across multiple challenge scenarios. Further evaluation indicates that the framework possesses the potential to scale to complex and multi-domain applications. Overall, this research establishes a novel framework that integrates intelligent task classification, cross-agent collaboration, and empirical tool integration capabilities. Compared to single-agent approaches, proposed system can achieve significant improvements in both efficiency and adaptability, providing a viable pathway for implementing LLMs in automated penetration testing and cybersecurity applications.

Keywords: LLM、Multi-agent systems、Automated Penetration Testing

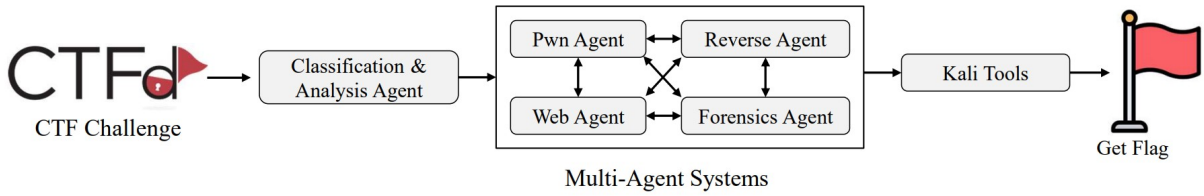


Fig. 1. Multi-Agent CTF Solving Framework

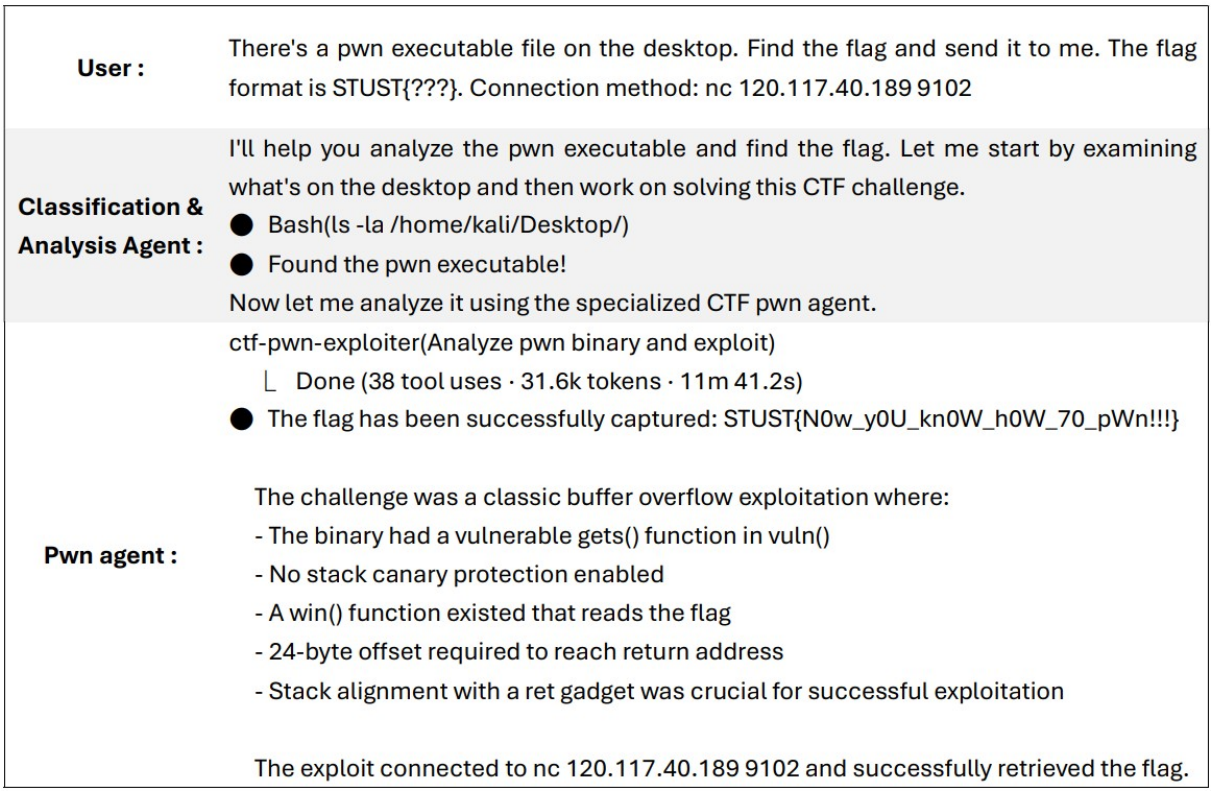


Fig. 2. Example of Multi-Agent Collaboration on a Pwn Challenge

References

[1] I. Isozaki, M. Shrestha, R. Console and E. Kim, "Towards Automated Penetration Testing: Introducing LLM Benchmark, Analysis, and Improvements," *UMAP Adjunct '25: Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, New York, NY, USA, 2025, pp. 404–419. doi: 10.1145/3708319.3733804.

[2] X. Shen, L. Wang, Z. Li, Y. Chen, W. Zhao, D. Sun, J. Wang and W. Ruan, "PentestAgent: Incorporating LLM Agents to Automated Penetration Testing," *ASIA CCS '25: Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, Hanoi, Vietnam, 2025, pp. 375–391. doi: 10.1145/3708821.3733882.

[3] S. G. Bianou and R. G. Batogna, "PENTEST-AI, an LLM-Powered Multi-Agents Framework for Penetration Testing Automation Leveraging Mitre Attack," 2024 IEEE International Conference on Cyber Security and Resilience (CSR), London, UK, 2024. doi: 10.1109/CSR61664.2024.10679480.

[4] K. Abdulghaffar, N. Elmrabit and M. Yousefi, "Enhancing Web Application Security through Automated Penetration Testing with Multiple Vulnerability Scanners," *Computers*, vol. 12, no. 11, p. 235, Nov. 2023. doi: 10.3390/computers12110235.

[5] B. Wu, G. Chen, K. Chen, X. Shang, J. Han, Y. He, W. Zhang and N. Yu, "AutoPT: How Far Are We from the Fully Automated Web Penetration Testing?," IEEE Transactions on Information Forensics and Security, vol. 20, pp. 9657–9672, 2025.