Autonomous Role-based Guard for UI Security (ARGUS)

Northeastern University
Khoury College of Computer Sciences

Jubril Akanbi, Linghe Zhou, Yiyang Wang
Khoury College of Computer Science, Northeastern University, Vancouver, BC, Canada

Supervised by: Dr. Maryam Tanha
Industry Partner: GreenHat Security

1. Background

Broken Access Control remains the #1 risk in the OWASP Top 10:2025. IDOR occurs when applications fail to verify if a user is authorized to access a specific resource ID, allowing unauthorized data manipulation.

ARGUS addresses this by employing **a multimodal, agentic AI** that combines visual reasoning with browser automation. It autonomously explores applications and mutates user flows across roles to detect **Insecure Direct Object References vulnerabilities**.

2. Motivations & Objectives

Recent advances in **Large Language Models** have demonstrated promising results in **autonomous penetration testing**, while separate work on vision-language models has shown that LLMs can effectively navigate and interact with web applications through visual interfaces.

ARGUS brings these two directions together in **a single browser-driven agent** that combines **visual UI understanding** with cross-role API probing to detect IDOR.

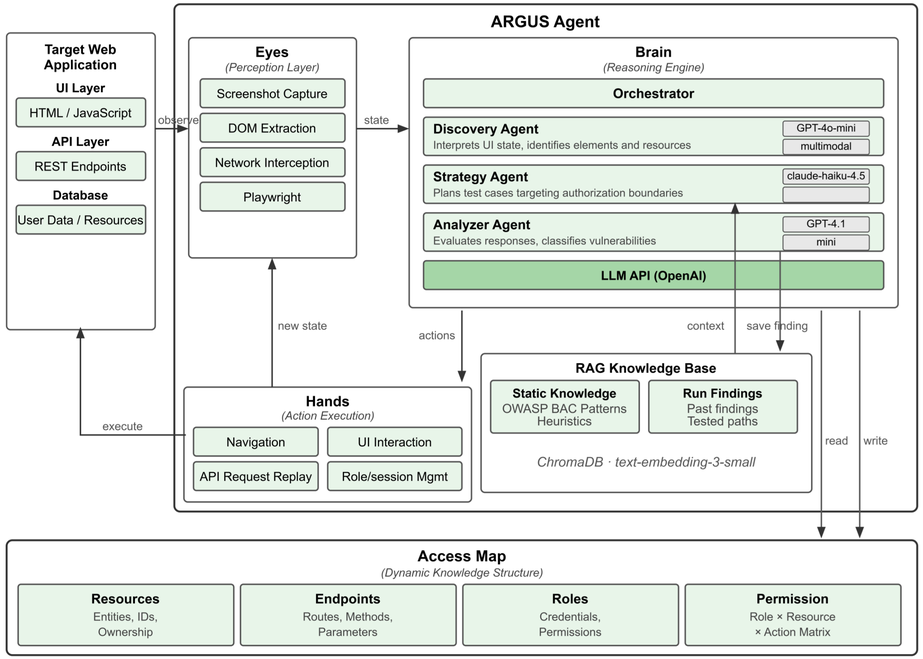
Beyond system design and validation, this work investigates how **model capability** and **prompt specificity** influence detection efficiency and cost, providing empirical guidance on LLM-based security agents.

3. Methodology

❖ System

ARGUS operates through a continuous perception-reasoning-action loop comprising three core components.

- **Eyes** → screenshot + DOM extraction + network traffic
- **Brain** → three specialized LLM agents
  - **Discovery Agent**: interprets screenshots and DOM content to identify interactive UI elements and navigation paths.
  - **Strategy Agent**: plans the next security test based on the current Access Map and RAG-retrieved context
  - **Analyzer Agent**: evaluates server responses to determine whether a vulnerability was exposed



- **Hands (Playwright)** → execute action in a real browser

Central to the system is the **Access Map**, a dynamically updated structure recording discovered endpoints, resource identifiers, user roles, and observed authorization outcomes.

The **Strategy Agent** is further augmented with a **Retrieval-Augmented Generation(RAG)** module backed by ChromaDB, which injects relevant security knowledge and prior run findings into each planning step, preventing redundant tests and prioritizing unexplored authorization boundaries.

❖ Experiment

Two controlled experiments are conducted against OWASP Juice Shop, targeting the same IDOR vulnerability across two user roles, with 3 trials per condition to account for LLM nondeterminism.

- **Experiment 1** fixes the system prompt and varies the Strategy Agent's model to measure how model capability affects detection reliability, iteration efficiency, and cost.
- **Experiment 2** fixes the model and varies prompt specificity: a structured workflow prompt versus a loosened prompt, to assess how much explicit step-by-step instruction drives detection, and whether that dependency differs across model capability tiers.

4. Experimental Results

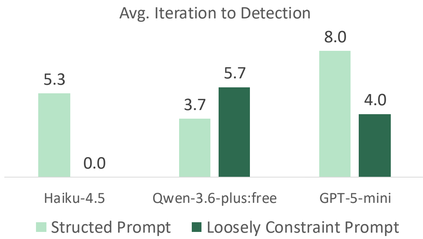
❖ Qualitative Results:

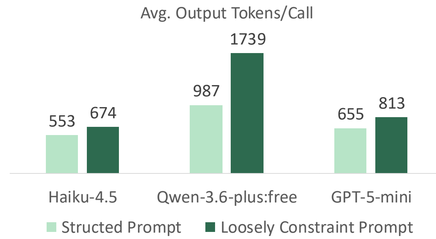
ARGUS successfully completed end-to-end agent loop execution across multiple development runs, validating the integration of all core components.

Each component performed its designated role reliably: Eyes captured page state, Discovery identified UI elements, Strategy produced well-formed security-motivated actions, and Analyzer correctly classified server responses with confidence scores.

Full coverage across all broken access control categories remains an open challenge, with **IDOR detection established as the validated baseline for further experimentation**.

❖ Quantitative Results:





The relationship between **prompt complexity and performance** is not monotonic - lighter models need more scaffolding to function, while stronger models may be constrained by it.

Haiku-4.5 failed entirely under the loosened prompt, testing a high-privilege user accessing a low-privilege user's resource.

GPT-5-mini improved under the loosened prompt, achieving 100% success in fewer iterations.

Models “reason more" when given freedom.

5. Conclusion & Future Work

ARGUS demonstrates that a **browser-based, multimodal LLM agent can autonomously detect IDOR vulnerabilities**, with prompt specificity and model choice meaningfully affecting both reliability and efficiency.

Future work should prioritize expanding IDOR and extending testing to other vulnerability classes including RBAC bypass, privilege escalation. Adapting ARGUS to unseen web applications beyond the current single-target setup as a configurable, user-facing tool.

Contact: akanbi.ad@northeastern.edu zhou.lingh@northeastern.edu wang.yiyang7@northeastern.edu