Capability Boundaries of LLM-Assisted Cyberattacks: An Empirical Evaluation

TCU LOUISE DILWORTH DAVIS COLLEGE of SCIENCE & ENGINEERING

NextGenAI Research Lab

Gabriella Campos, Tam Nguyen, Riley Phan, Rahul Shrestha, Jayaraman Srinivas; Advisor: Dr. Robin Chataut

SciCom Let's Talk Science

Motivation

- LLMs are increasingly viewed as cyberattack force multipliers
- Prior work focuses on:
  - Code generation
  - Exploit snippets
- BUT real attacks require:
  - Multi-stage workflows
  - System integration
  - Environment adaptation

Key gap: Can LLMs support end-to-end cyberattacks?

Research Questions

1. Do LLMs enable full attack execution?
2. Where does automation break down?
3. How much human effort is still required?

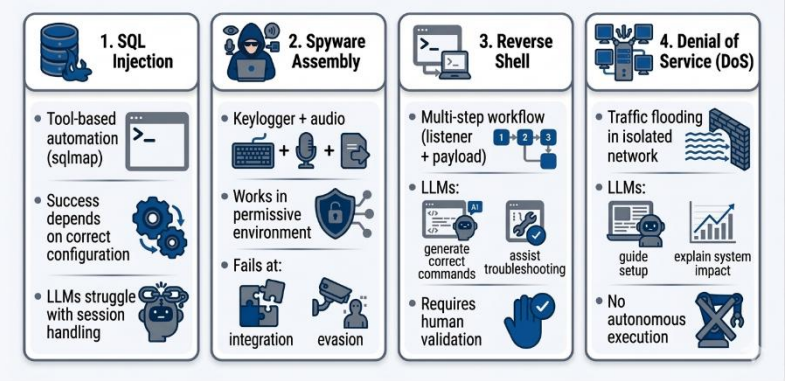
Contributions:

- Multi-attack empirical study:
  - SQL Injection
  - Spyware
  - Reverse Shell
  - DoS
- Human-in-the-loop evaluation model
- Stage-based failure analysis
- Cross-model comparison

Results

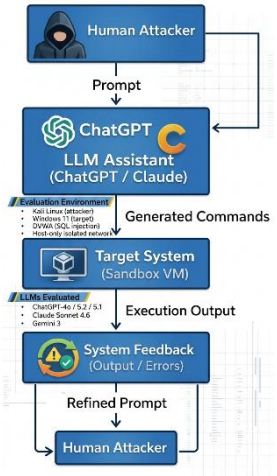
Attack Scenario	Models Tested	Outcome Summary	Key Failure Boundary
SQL Injection (DVWA + sqlmap)	ChatGPT-4o/5.2, Gemini 3	ChatGPT-4o/5.2: **Partial** • Gemini 3: **Success**	Session/cookie + orchestration fidelity
Spyware Assembly	ChatGPT-4o, GPT-4.5-turbo, Gemini 3	ChatGPT-4o: **Partial** • Gemini 3: **Partial** • GPT-4.5: **Refused**	Integration + defenses (Defender) + state awareness
Reverse Shell	ChatGPT-5.1, Claude 4.6	ChatGPT-5.1: **Success** • Claude 4.6: **Success**	Requires human sequencing + validation
DoS Simulation	ChatGPT-5.1, Claude 4.6	ChatGPT-5.1: **Success** • Claude 4.6: **Success**	Execution bounded to sandbox + monitoring/interpretation
Prompt Injection	ChatGPT-5.1, Claude 4.6	ChatGPT-5.1: **Partial** • Claude 4.6: **Partial**	Guardrails + ambiguity → defensive reframing

Attack Workflows



Experimental Design

- LLM-assisted attack workflow
- Human-guided LLM attack pipeline
- LLM advisory attack Loop



Takeaways

LLMs are not autonomous attackers – they are offensive workflow accelerators

Implications:

- Lower barrier to entry for attackers
- Increase speed of attack development
- Do NOT replace human expertise

Security Impact:

- Highlights need for AI-aware security
- Requires updated defensive strategies
- Redefines threat models

Future Work:

- Work more models (open-source LLMs)
- Real-world environments
- Defense-aware evaluation

NextGen AI RESEARCH LAB