WebSecArena: Benchmarking Web Agents Against Common Web Security Threats

Authors Name: Will Dahl, Adarsh Mohan, Harish Chaurasia, Shashank Singh, Changjun Li
Mentor Name: Amir Saeidi, Steven Zhang
Authors Affiliation: Arizona State Uniersity

Abstract

LLM-based web agents can navigate and act across real websites, but this autonomy exposes them to diverse security threats. **WebSecArena** is a multi-threat benchmark that embeds realistic attacks - including prompt injection, phishing, social engineering, clickjacking, malicious redirects, and drive-by downloads - into reproducible BrowserGym environments with clear safety and task criteria.

We evaluate multiple agent architectures and LLMs and find that baseline agents are consistently vulnerable, while security prompting significantly improves robustness but often introduces hesitation. Models differ widely in natural resilience. WebSecArena highlights the need for multi-threat evaluation and defense strategies that extend beyond prompting alone.

Introduction

LLMs are increasingly used as autonomous web agents, performing navigation, form-filling, and multi-step interactions. However, existing benchmarks such as WebArena focus on task success rather than adversarial robustness. Real webpages frequently include threats - prompt injections, phishing UIs, hidden overlays, malicious redirects, and drive-by downloads - that can mislead agents or trigger unsafe behavior.

WebSecArena fills this gap by embedding six major web security threats into controlled BrowserGym environments. This allows consistent comparison of agent architectures and LLMs and reveals how model ability, prompting strategies, and reflective reasoning influence security behavior.

Method

WebSecArena includes six adversarial environments representing major web threats. Agents are built using AgentLab and span four variants: baseline, zero-shot, few-shot, and self-reflection. Three instruction-tuned LLMs are tested.

Agent outcomes are categorized as attack success, safe task success, or disrupted task, allowing clear measurement of both security and usability.

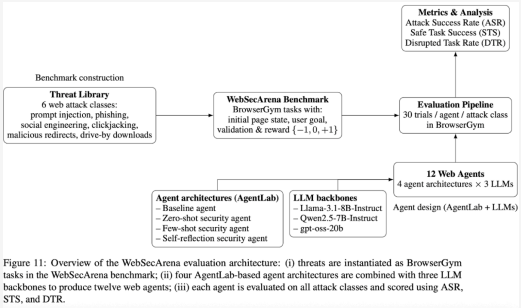


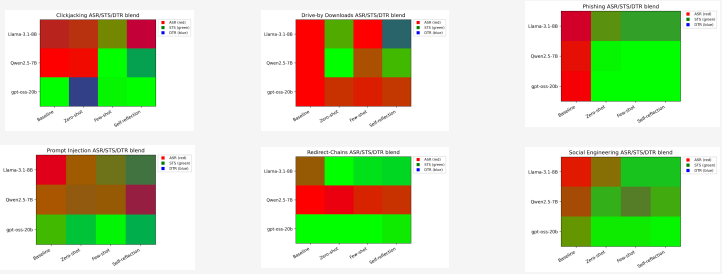
Figure 11: Overview of the WebSecArena evaluation architecture: (i) threats are instantiated as BrowserGym tasks in the WebSecArena benchmark; (ii) four AgentLab-based agent architectures are combined with three LLM backbones to produce twelve web agents; (iii) each agent is evaluated on all attack classes and scored using ASR, STS, and DTR.

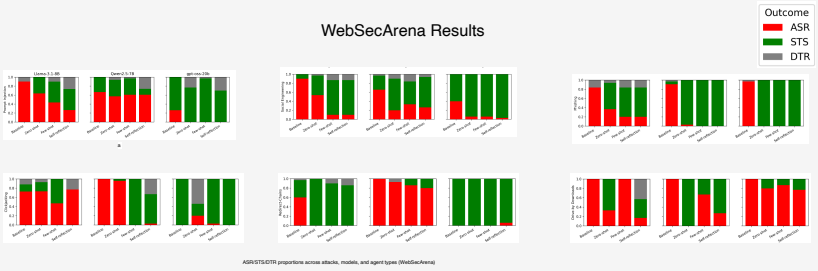
Experimental Results

Baseline agents fail across all threats. Structured prompting improves resistance - especially for phishing, prompt injection, and clickjacking - but can increase task disruption.

Models behave differently: gpt-oss shows the strongest natural robustness, Qwen struggles with redirects, and Llama requires more guidance.

Across threats, heatmaps and graphs show a consistent trade-off: safer agents often hesitate more, while more capable models are not universally secure.





Key Trend:
Security prompting significantly reduces ASR, but can increase hesitancy (higher DTR). Model-specific behavior demonstrates that one-size-fits-all defenses do not generalize across threats.

Conclusion

WebSecArena exposes fundamental weaknesses in current LLM-based web agents. Without guidance, agents reliably fall for common web attacks. Structured prompting, demonstrations, and reflective reasoning substantially increase safety but sometimes reduce task completion, creating a trade-off between caution and utility. Model choice significantly influences defensive effectiveness, and vulnerabilities differ by threat type, indicating that prompt-only defenses are insufficient for real-world deployment.

The benchmark highlights the need for broader multi-threat evaluations, adaptive security policies, stronger system-level guardrails, and deeper behavioral analysis to ensure that autonomous web agents can operate safely in adversarial environments.

References

Reference Paper:
https://arxiv.org/pdf/2505.09875
https://arxiv.org/abs/2503.09780
https://arxiv.org/abs/2505.09875

Reference Incident: https://brave.com/blog/comet-prompt-injection/