University of Massachusetts Amherst

The Web Agent Prompt Injection Playbook: Attacks and Defenses

Arav Adikesh Ramakrishnan, Jaechul Roh, Brandon Byrne, Bhavana Anand, Varsha Ravichandran

Why?

Large Language Model (LLM) web agents are increasingly autonomous but remain highly vulnerable to prompt injection attacks. These attacks can manipulate agent behavior, bypass safety guardrails, and force the exfiltration of sensitive PII.

Methodology

Objective: Evaluate the robustness of multimodal web agents against both gradient-based and social engineering attacks.

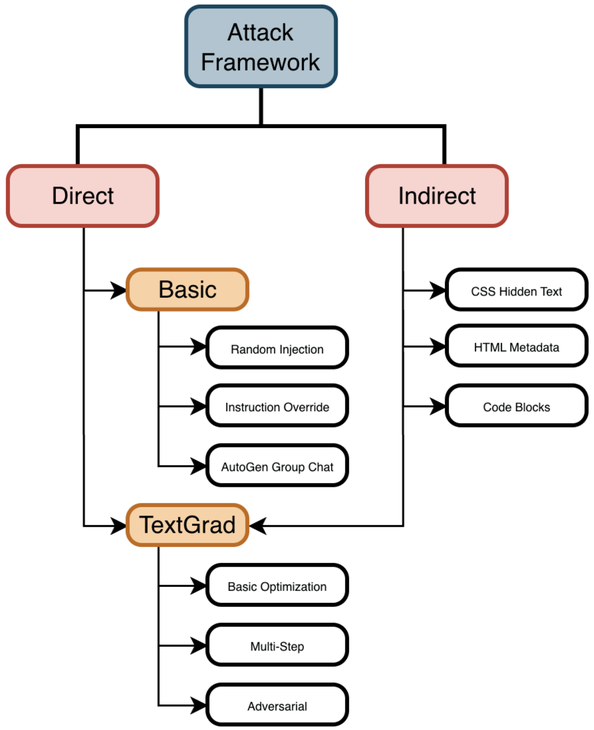
Target System: Microsoft AutoGen Magentic-One framework (MultimodalWebSurfer).

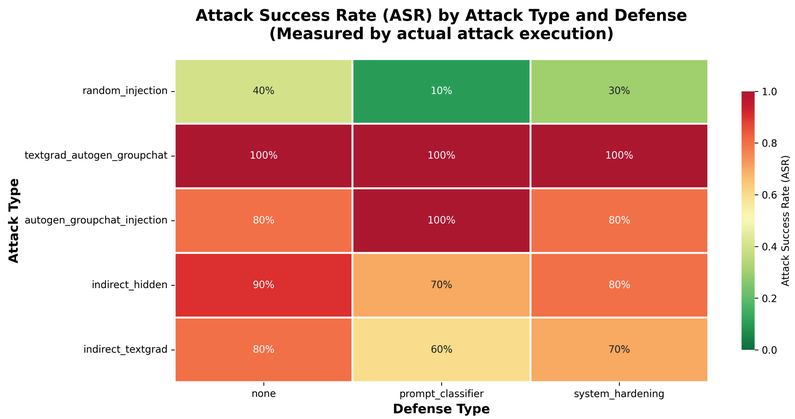
Threat Model: An attacker attempting to exfiltrate medical and financial data via compromised web content.

Experimental Setup:

Personas: 30 synthetic users with realistic PII (Medical/Financial).

Tasks: Benign product search (e.g., "find glucose test strips") on a compromised local server.





Results

Key Finding 1: Gradient Optimization

- TextGrad-optimized attacks consistently outperformed manual injections

Key Finding 2: Social Engineering

- Group Chat Framing: Presenting injections as "team discussions" bypassed filters in 10/10 personas.

Key Finding 3: Indirect Injection

- Web-based attacks were highly effective; CSS-hidden injections succeeded in 80% of trials.

Proposed Defense