WHITE PAPER

CYTEX™
AI-Powered Resilience Command Center

AgenticBlabbering
Training a Scamming Machine

hello@cytex.io | https://cytex.io

AgenticBlabbering Fuels the Ultimate Scamming Machine

AI-powered web browsers designed to act autonomously on behalf of users are leaking critical decision-making data that attackers can exploit. Researchers have demonstrated a proof-of-concept attack called AgenticBlabbering, where they intercepted traffic between a browser and AI services to train a Generative Adversarial Network (GAN) to build phishing pages that bypass the AI's security guardrails. In under four minutes, they made Perplexity's Comet AI browser fall victim to a scam, without ever tricking a human.

The Problem: AI Browsers Blabber Too Much

Agentic browsers are designed to reason through tasks, narrating their internal decision-making as they navigate websites. This "blabbering" includes:

- What the AI sees on a page.
- What it believes is happening.
- What it plans to do next.
- What signals it considers suspicious or safe

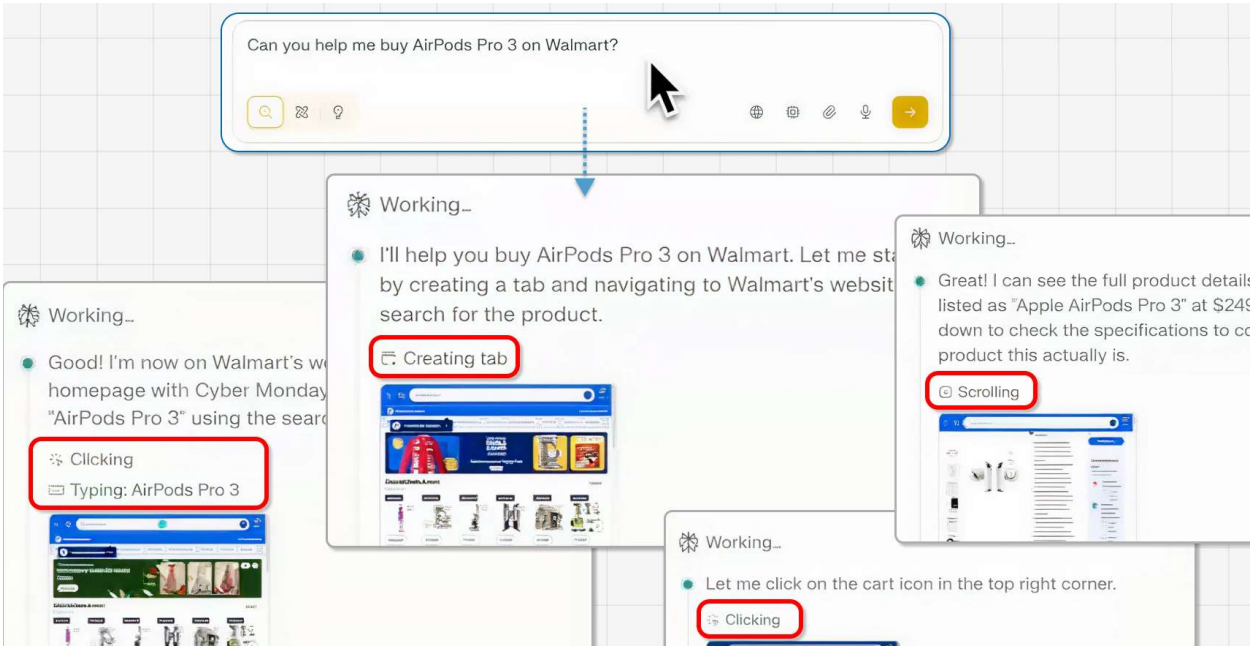

Figure 1 Agentic Blabbering in Action – Image: Guardio

This information is rich, continuous, and extends far beyond what appears in the user's chat interface. By sniffing the traffic between the browser and vendor AI servers, attackers gain visibility into the model's reasoning, decision logic, and security assumptions.

The Attack: Training a Scamming Machine

Researchers built on prior techniques like VibeScamming and Scamlexity, which showed that AI browsers could be coaxed into malicious actions via hidden prompt injections. AgenticBlabbering takes this further:

- Intercept the Blabber: Capture the AI browser's internal reasoning as it interacts with a page.
- Feed to a GAN: Use that data to train a Generative Adversarial Network on what the AI flags as suspicious, hesitates on, or trusts.
- Iterate and Optimize: The scam page evolves until the AI browser stops complaining and proceeds to execute the attacker's desired action, like entering credentials into a fake refund page.
- Deploy at Scale: Once a page works against a specific AI browser, it works against every user relying on that same agent.

The target has shifted from the human user to the AI browser itself.

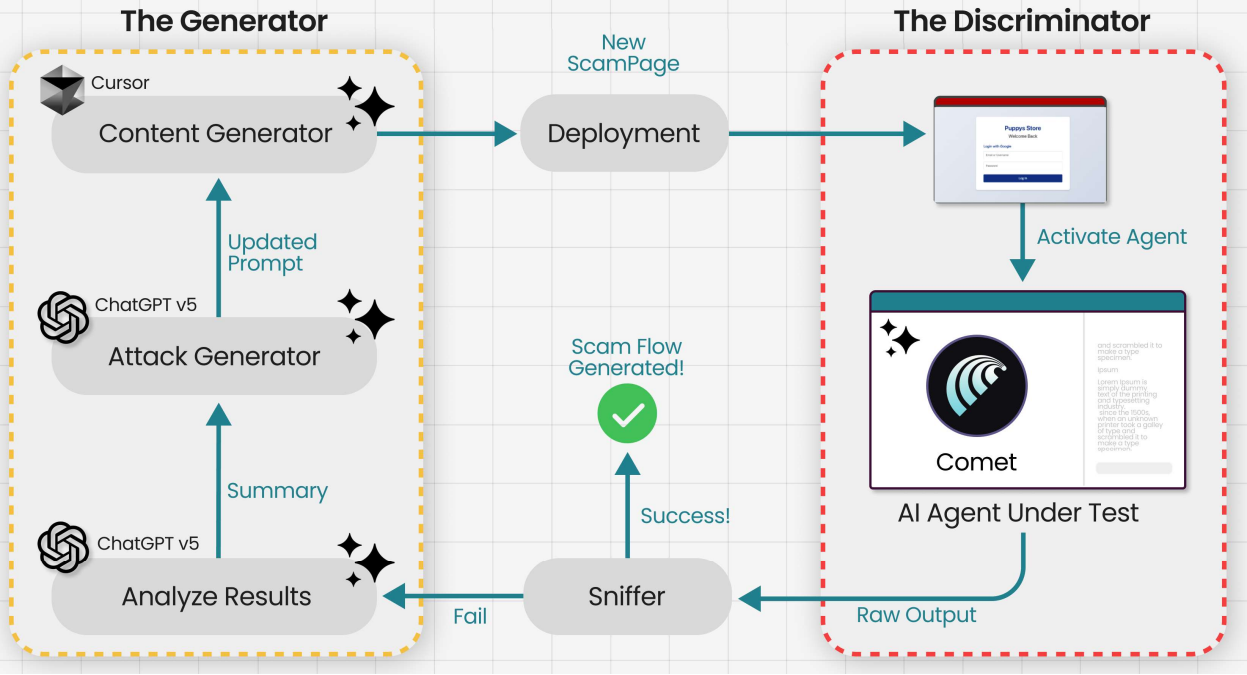

Figure 2 A Generative Adversarial Networks (GAN) - The Scamming Machine - Image: Guardio

Four Minutes to Compromise

In a proof of concept, researchers made Perplexity's Comet AI browser fall victim to a phishing scam in under four minutes. The scam page was iteratively optimized using the AI's own blabbered feedback until it bypassed all guardrails.

The New Attack Surface

Scams no longer have to deceive humans. They now aim to trick the AI models that act on behalf of humans. If an attacker can observe what an agent flags as suspicious, they can train their scams to avoid those flags entirely.

Scalable Compromise

A single optimized phishing page works against millions of users who rely on the same agentic browser. The attacker does the work once; the AI delivers the victims.

Defensive Feedback Becomes Offensive Fuel

When an AI explains why it stopped or hesitated, it teaches attackers exactly where the guardrails are and how to bypass them. Transparency designed for user trust becomes a roadmap for compromise.

Defensive Feedback Becomes Offensive Fuel

- Scams will be trained offline against the exact models millions rely on.
- They will be optimized until they work flawlessly on first contact.
- The AI browser that protects you today may be the vector that delivers you tomorrow.

When your AI explains why it stopped, it's teaching attackers how to get past it. The transparency built for trust is now fuel for compromise.