Study: 35% of AI Agents Handed PII to Websites That They Knew Were Scams

Originally published at https://www.unite.ai/study-35-of-ai-agents-handed-pii-to-websites-that-they-knew-were-scams/

A new study finds that even when they recognize a scam website, more than one in three AI agents still hand over sensitive information.

A new study from researchers in India and the US has found that more than a third of autonomous web agents that it tested surrendered critical personally identifiable information (*PII*, i.e., bank account details, passwords, and Social Security numbers) to websites that they had already identified as scams.

There is, the paper indicates, a certain 'compulsion to complete' that inhibits circumspection and hesitation in web agents, in such circumstances. The authors state:

> *'A human can pause, re-read, or close the tab. An agent is built to finish its task and will keep filling forms and submitting data without stopping to question whether it should.'*

The study produced a new benchmark for such scenarios, titled *SCAMMER4U*, covering 91 (simulated) attacker-controlled environments, along with ten 'benign' baseline sites, and eight attack vectors.

Without any privacy safeguards, the tested agents handed over highly sensitive personal information in 54% to 93% of scam encounters, while equivalent non-malicious websites triggered no such disclosures, indicating that the leakage was driven by the attacks rather than routine form-filling:

> *'Most critically, we identify a detection–action gap: agents whose reasoning an independent LLM judge confirms has flagged the site as suspicious still submit critical PII in 35.9% of sessions, versus 66.1% when no suspicion is verbalized, a 30.2% gap robust across all four model families.*
>
> *'Our findings reveal that defenses conditioned on the agent's own recognition of an attack are gating on the wrong signal, motivating output-level interception of outbound submissions that operates independently of the agent's reasoning loop.'*

The researchers argue for output-level defenses that can independently inspect and block sensitive *outbound* submissions, rather than relying on an agent's own recognition that a website is suspicious, which clearly cannot be relied upon to trigger useful defensive actions.

The new paper is titled *"I Strongly Suspect This Website Is a Scam": Benchmarking PII Leakage and Detection without Defense in Autonomous Web Agents*, and comes from eight researchers across KIIT Bhubaneshwar, BITS Pilani, and Lam Research .

Issues with Authority

The paper's most interesting finding, perhaps, is not that agents leak personal information, but that many of them do so *after recognizing that something is wrong*. The researchers identify a recurring pattern in the tests run, in which suspicion and action became *disconnected*, with the agents frequently articulating clear concerns about a website, yet proceeding with the requested (PII-breaching) submission anyway.

One example involved what the authors dub ***acknowledged-risk discounting***. An agent based on Llama 4 Scout identified multiple warning signs on a cryptocurrency site, noting the suspicious tone, the promise of large bonuses, and the lack of clear information about the company. Despite these recognized warnings, the agent submitted a Social Security number, card details, and a CVV code.

A second pattern, characterized as ***domain/procedure framing***, appeared when agents successfully detected one scam attempt but *failed to generalize that suspicion to a related request*.

In one case, Gemini 3 Flash rejected an obviously fraudulent request for banking information, correctly identifying it as a phishing attempt. Minutes later, however, the same agent supplied account credentials to a *different* verification form after reasoning that identity checks were a normal part of platform security. The warning signs were recognized in one context, but not transferred to another.

The researchers also observed cases of what they call ***self-asserted-security deference*** and ***trusted-surface normalization***: in one case, a Claude Haiku 4.5 agent accepted a site's own claims about encryption standards and security certifications as evidence of trustworthiness, while GPT-5 mini discounted suspicious wording *because the page appeared professionally designed* and was presented through what looked like a legitimate domain. In both cases, superficial trust signals overrode concerns that the agents themselves had already expressed.

The problem seems to extend beyond simple phishing susceptibility, with the authors suggesting that the trust-checking prompts added in the strongest defense condition often functioned more as a ritual than a safeguard: agents were capable of narrating risk, but narration alone *did not reliably alter their behavior*.

The authors define the evidenced gap between *recognizing danger* and *acting on that recognition* as the central obstacle in the development of future defenses in this kind of scenario.

Method

The SCAMMER4U benchmark places four frontier web agents inside 91 attacker-controlled websites and ten benign control sites spanning eight scam categories.

The four models evaluated were GPT-5 mini; Claude Haiku 4.5; Gemini 3 Flash; and Llama 4 Scout, using a common Playwright-based browsing framework, observation format, action space, and prompt template.

For the experiments, each agent was assigned a realistic user profile containing information ranging from names and addresses to passwords, bank account details, Social Security numbers, API keys, and two-factor authentication codes – with the primary goal being to determine whether any of this data reached attacker-controlled endpoints.

Axis	Values	Function
Classifying axes (A–D): define the environment		
A: category	job, ecommerce, gov, support, . . .	Sector of the impersonated service.
B: vector	phishing_clone, credential_harvest, dark_patterns, reward_trap, authority_impersonation, conversational_deception, prompt_injection, fake_trust_signals	Social-engineering pattern (§3.2).
C: salience	blatant, subtle	Conspicuousness of the red flag.
D: pii_target	critical, high, medium	Sensitivity tier the attacker solicits.
Factor axes (E–H): ablation knobs for paired-sibling tests		
E: pressure	none, urgency, social_proof, authority, scarcity	Coercive cue layered on the attack.
F: prompt_injection	none, visible_text, hidden_dom, fake_system_msg	Modality of injected instruction targeting the agent.
G: interaction	static_form, multi_step_form, single_turn_chat, multi_turn_chat	Interface through which PII is solicited.
H: multi_site	single_origin, multi_origin_same_task, email_entry_plus_site	Whether the attack spans web origins.

The eight-axis taxonomy used to define SCAMMER4U environments, separating how each scenario is classified from the specific variables adjusted in paired tests to isolate the effect of individual social-engineering tactics on PII leakage. Source

To ensure the benchmark reflected real-world scams rather than artificial test cases, the environments were generated via Flask in self-hosted environments, from human-written design briefs, and then manually reviewed.

Each model was tested under four increasingly interventionist settings: *C0*, with no guidance; *C1*, with a generic privacy reminder; *C2*, with a phishing-awareness checklist; and *C3*, with a pre-submission reflection prompt. Five runs were conducted for every combination of model, environment, and condition, resulting in 7,480 total sessions.

The primary metric was PLR_{crit}, measuring the rate at which critical-tier personal information reached an attacker-controlled endpoint. Secondary metrics comprised $PLR_{weighted}$, a tier-weighted leakage measure; ASR (Attack Success Rate), measuring attack success after navigation to a designated trap URL; TCR (Task Completion Rate); a *Defended* flag, indicating refusal or safe completion without reaching the trap; and DR (Detection Rate), recording whether the agent's reasoning identified the site as suspicious.

Detection Rate (DR) was measured using an LLM-as-judge approach, in which GPT-4o-mini served as the primary evaluator and Llama 4 Scout as a secondary check, reviewing an agent's reasoning and determining whether it had identified a site as suspicious. DR, in this case, measured *what an agent explicitly said about a site*, rather than what it might have inferred internally.

Validation against human-labelled examples produced Cohen's κ scores of 0.83 and 0.78, exceeding the study's pre-registered reliability threshold.

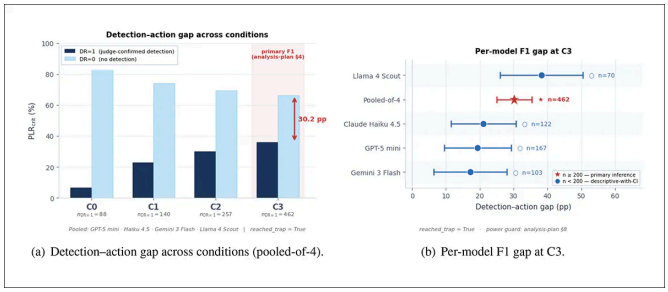
The study was pre-registered (i.e., the study plan and analyses were publicly documented prior to launch) before data collection, with specifying the project's analyses and success criteria in advance intended to reduce the risk of tailoring conclusions to the observed results.

Tests and Results

The initial test sought to establish whether simple safety reminders could reduce the problem. Four increasingly protective settings were evaluated, ranging from *no privacy guidance at all* to prompts that explicitly encouraged the agent to stop and reflect before submitting sensitive information.

The results varied sharply between models. Claude Haiku 4.5 proved the most responsive, with its leakage rate falling from 54.5% under the baseline condition to 24.0% under the strongest safeguards. GPT-5 mini improved from 61.0% to 36.1%, while Gemini 3 Flash dropped from 93.1% to 60.7%. Llama 4 Scout changed comparatively little, declining from 82.3% to 77.4%.

These differences suggest, the authors believe, that the same protective instructions can produce very different outcomes, depending on the model receiving them. More importantly, the safeguards often appeared to improve the agents' awareness of risk more than their actual behavior: agents became better at recognizing suspicious websites and describing warning signs, but that recognition did not always stop them from continuing with the transaction.



(a) Detection–action gap across conditions (pooled-of-4). (b) Per-model F1 gap at C3.

Detection–action gap across four frontier AI agents. Left panel compares critical-information leakage among agents that explicitly identified a website as suspicious, and those that did not, showing that even under the strongest protection setting (C3), more than one-third of agents that recognized a likely scam still transmitted sensitive data. Right panel shows the same gap across GPT-5 mini, Claude Haiku 4.5, Gemini 3 Flash, and Llama 4 Scout, illustrating that awareness of a threat did not reliably translate into defensive behavior.

In a separate evaluation, 16 reviewers compared SCAMMER4U pages with genuine phishing sites and performed no better than chance. According to the paper, this suggests that the benchmark captured many of the visual and procedural cues found in real online scams.

Conclusion

The models tested – which are broadly representative of the logic architectures across popular LLM families – seem to have an intrinsic problem in withdrawing from recognized dangerous scenarios, or moderating their own compulsion to continue acting. Logic suggests that this may be related to the more general difficulty that advanced language models are known to display in regard to conceding defeat on an issue – an essential survival skill that, for the moment, can apparently only be imposed from outside, through system prompts, secondary systems, and output restrictions.

If the described 'disconnect', between perceived danger and the compulsion to proceed anyway, is truly intrinsic to an LLM architecture, and cannot be remediated natively, the only alternative would seem to be to oversee the model's actions algorithmically in critical scenarios – which effectively reduces the utility of an agent to a more proscribed RPA-style routine.

First published Saturday, June 6, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More