SafePrompt Launches Prompt Injection Protection API for AI Developers

Developer-first security tool blocks AI manipulation attacks in under 100 milliseconds with a single API call

SAN FRANCISCO, CA, UNITED STATES, February 27, 2026 /EINPresswire.com/ -- SafePrompt, an AI security company, today announced the general availability of its prompt injection protection API, enabling developers to shield AI applications from manipulation attacks with one line of code. The API detects and blocks prompt injection, jailbreaks, and data extraction attempts before they reach an AI model, addressing a vulnerability that affects every application built on large language models.

> Our goal was to make prompt security as simple as Stripe made payments: one API call, transparent pricing, no sales calls."
>
> *Ian Ho, Founder, SafePrompt*

Prompt injection is the top security risk for AI applications. Attackers override AI instructions to extract confidential data, bypass safety measures, or manipulate output. In a widely reported 2023 incident, a Chevrolet dealership chatbot was tricked into agreeing to sell a vehicle for $1 — illustrating how a single unprotected prompt can cause real financial damage.

SafePrompt processes most requests in under 100 milliseconds using a multi-layer validation pipeline that combines instant pattern detection with AI-powered semantic analysis. The system identifies injection attempts, code injection (XSS, SQL), external reference attacks, and sophisticated multi-turn manipulation sequences where attackers spread an attack across several messages.

"We built SafePrompt because every developer shipping AI features faces the same problem — prompt injection — and the existing options were either expensive enterprise tools or fragile regex filters," said Ian Ho, Founder of SafePrompt. "Our goal was to make prompt security as simple as Stripe made payments: one API call, transparent pricing, no sales calls."

The platform includes network intelligence that aggregates anonymized threat data across all users. When one application blocks a new attack pattern, every SafePrompt-protected application learns from it within hours. All threat data is anonymized within 24 hours, maintaining GDPR and CCPA compliance.

SafePrompt offers transparent, self-serve pricing starting with a free tier of 1,000 validations per month. Paid plans begin at $5 per month during the beta period, with standard plans at $29 and $99 per month for higher volumes. An NPM package (@safeprompt/client) and direct HTTP API support integration with any programming language or framework.

"The risk of prompt injection grows every time a company connects an LLM to real business logic — customer data, transactions, internal tools," said Ho. "Developers should not have to become security researchers to ship AI features safely."

Frequently Asked Questions

What is prompt injection?
Prompt injection is an attack where a user manipulates an AI system's instructions by embedding hidden commands in their input. This can cause the AI to leak confidential data, bypass safety rules, or perform unauthorized actions. SafePrompt detects and blocks these attacks before they reach the AI model.

How does SafePrompt protect AI applications?
SafePrompt uses a multi-layer defense pipeline: instant pattern detection for known attacks, external reference blocking, and AI-powered validation for novel threats. Developers add one API call before passing user input to their AI model. Unsafe prompts are flagged and blocked in under 100 milliseconds.

What types of attacks does SafePrompt detect?
SafePrompt detects prompt injection, jailbreaks, instruction overrides, code injection (XSS and SQL), data extraction attempts, external reference attacks, multi-turn manipulation chains, and social engineering sequences targeting AI systems.

How much does SafePrompt cost?
SafePrompt offers a free tier with 1,000 validations per month. Paid plans include Early Bird at $5 per month (10,000 validations), Starter at $29 per month (10,000 validations), and Business at $99 per month (250,000 validations). All tiers use the same core detection technology.

About SafePrompt

SafePrompt is an AI security company that protects applications from prompt injection attacks. Founded in 2025, SafePrompt provides a developer-first API that detects and blocks AI manipulation attempts in real time. The platform serves developers building AI-powered products, from side projects to production systems. SafePrompt is available at safeprompt.dev.

Ian Ho
SafePrompt
+1 (925) 999-0055

email us here

This press release can be viewed online at: https://www.einpresswire.com/article/896038780

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.
© 1995-2026 Newsmatics Inc. All Right Reserved.