CYBERSECURITY 101 · eBuilder Security

Prompt Injection Attack

A prompt injection attack tricks an AI assistant into following hidden instructions instead of its own rules. It needs no code, only words, and it is now the number one risk for AI systems. Here is what your team needs to know.

WHAT IT IS

A prompt injection attack is crafted text that an AI system reads as a new instruction, overriding what its developer told it to do. The text can be typed straight into a chatbot or hidden inside an email, a document or a web page the AI reads for you. Because a model cannot reliably tell trusted instructions from other text, a well-worded message can make it break its rules, reveal data or, in an AI agent, take an action on the attacker's behalf.

WHY IT MATTERS

No. 1

Prompt injection is the top-ranked security risk for AI applications, two editions running.

OWASP, 2025

71%

Of organisations now use generative AI in at least one business function, up from 65% a year earlier.

McKinsey, 2025

Zero-click

A single crafted email could make Microsoft 365 Copilot leak internal data, with no user action.

CVE-2025-32711, 2025

WHAT HAPPENED

In December 2023, a customer told the ChatGPT-powered chatbot on a Chevrolet dealership's website to agree with anything he said and treat it as binding, then got it to accept one dollar for a new SUV. The dealership never honoured the offer and pulled the bot within days, but the screenshots went worldwide. The bot had been allowed to speak for the business with no limits. (Business Insider, 2023)

RED FLAGS TO WATCH FOR

You cannot spot prompt injection by reading the input, so watch what the AI does, not how the message looks.

- An assistant acts outside its brief, changing topic or offering to do something it was not asked to.
- Output that quotes or describes the AI's own rules or hidden instructions.
- An email or document that seems to give instructions to the AI rather than to you.
- An agent proposing to send data out or contact an address you do not recognise.

THE ONE RULE THAT STOPS IT

Never let an AI system take a sensitive action, like sending data or approving a payment, without a human checking it on a separate channel.

THEN BUILD THE HABIT AROUND IT

- Give AI systems and their tools the least access they need to do the job.
- Keep untrusted content separate so retrieved text is treated as reference, never as commands.
- Block channels an assistant could use to send data to unknown destinations.
- Test AI systems for injection before launch and after every change.

MYTHS AND FACTS

MYTH

A prompt injection attack needs coding skills.

FACT

Most need only plain language. The Chevrolet and DPD chatbots were pushed off script with ordinary typed sentences, because the attack targets how the model reads instructions.

MYTH

A good input filter will stop it.

FACT

Filters reduce the risk but cannot close it. NIST and OWASP both say no current method fully prevents prompt injection, and the EchoLeak email slipped past Microsoft's own filter.

MYTH

Only public chatbots are at risk.

FACT

The bigger exposure is internal assistants with access to company data and tools. An injection hidden in an email or file can reach them, as the Microsoft 365 Copilot flaw showed.

THE LEGAL DUTY

The EU AI Act requires high-risk AI systems to resist input manipulation (Article 15), and under NIS2, in force in Sweden as Cybersäkerhetslagen since 15 January 2026, a resulting breach must be reported to MCF (formerly MSB) within 24 and 72 hours, with boards personally accountable.

GET HELP

eBuilder Security helps Swedish organisations test and defend the AI systems they now rely on, from awareness training to AI detection and response. Read the full guide at the link below.

Read the Full Article →