LAKERA

A Check Point Company

Q4 2025 AGENT SECURITY TRENDS

INSIDE THE REAL ATTACKS TARGETING AI AGENTS

DEC 2025 REPORT

Q4 2025 AGENT SECURITY TRENDS

[REC] 30-DAY WINDOW SEVERITY: HIGH

Executive Summary

Q4 2025 marked a clear shift in how AI systems were attacked. As models gained early agentic capabilities such as tool use, browsing and structured context handling, attacker behavior evolved in parallel. This report highlights the most common attacker intents and the techniques observed across real-world traffic protected by Lakera Guard.

THREE SIGNALS FROM THE DATA

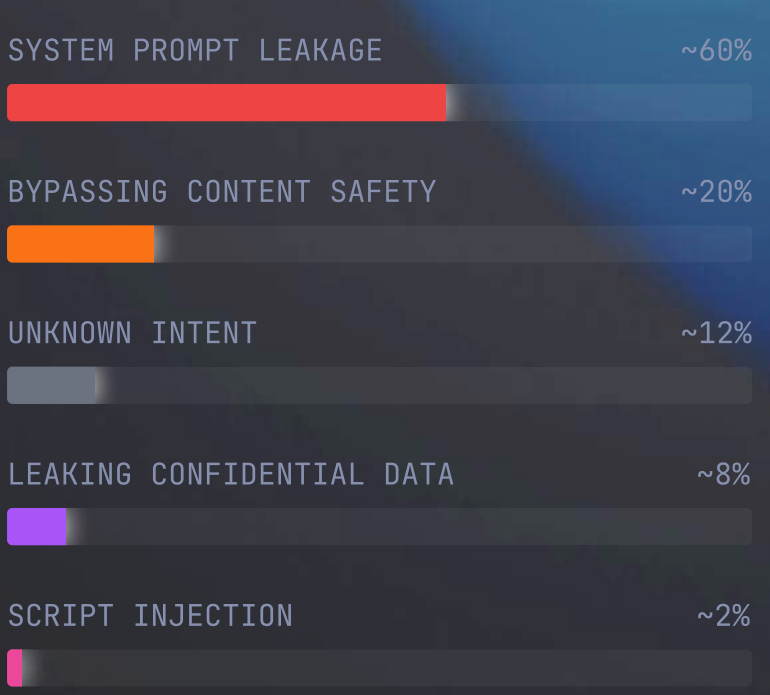
01 // CONFIG TARGETING
Attackers targeted system configuration more than any other objective.

02 // INDIRECT EFFICIENCY
Indirect attacks required fewer attempts to succeed and appeared across several categories.

03 // NEW SURFACES
Agentic capabilities introduced new attack surfaces, including script-shaped content.

Attack Intent Breakdown

FIG.01



KEY FINDINGS

- System Prompt Leakage was the most common attacker intent (~60% of attack traffic).
- Bypassing Content Safety was second most frequent (~20% of attack traffic).
- Unknown Intent probes (~12% of attack traffic) reflected attacker reconnaissance.
- Leaking Confidential Data (~8% of attack traffic) indicated risks tied to agent workflows.

* SEE PG.03 FOR DEFINITIONS

TECHNIQUE DISTRIBUTION INSIGHTS

Hypothetical Scenarios and Obfuscation accounted for most System Prompt Leakage attempts. Role Play was dominant in Content Safety Bypass attempts. Indirect attacks appeared across multiple intent categories.

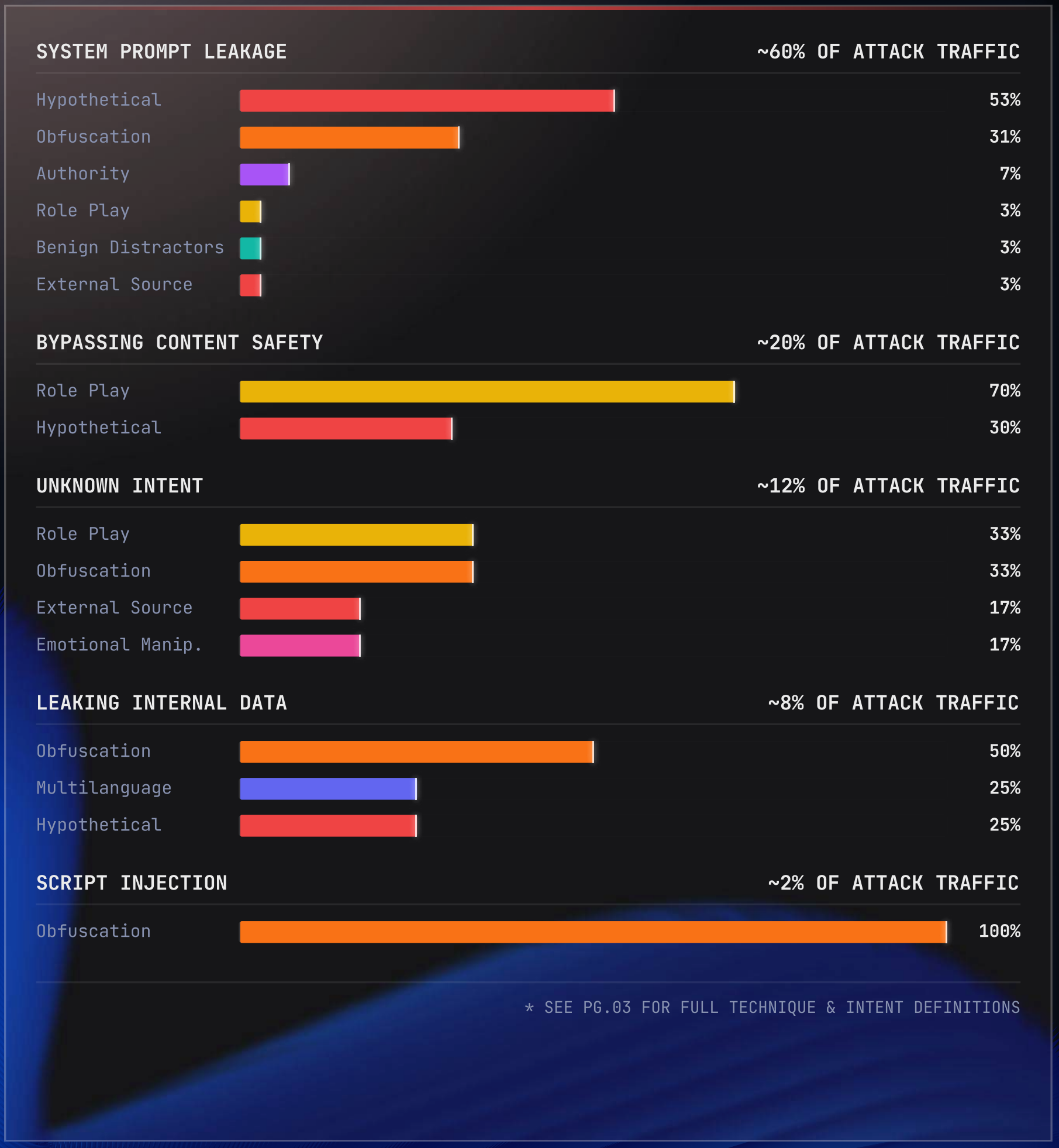
MINI-METHODOLOGY

Analysis covers a 30-day window of attack traffic observed in Q4 2025. Each attempt was categorized into one of 5 intents and 8 techniques using Lakera's internal taxonomy. Example prompts appear in the companion blog post.

TECHNIQUES BY INTENT

HOW ATTACKERS EXECUTED THEIR ATTEMPTS IN Q4 2025

Here we break down which techniques attackers used and how they aligned with specific attack intents. The visualization highlights how Hypothetical Scenarios, Obfuscation and Role Play dominated the landscape, while newer techniques appeared across smaller but meaningful categories.



* SEE PG.03 FOR FULL TECHNIQUE & INTENT DEFINITIONS

HYPOTHETICAL · OBFUSCATION · ROLE PLAY · DISTRACTORS · AUTHORITY · EXTERNAL · EMOTIONAL · MULTILANG

INTENT & TECHNIQUE DEEP DIVE

DEFINITIONS OF KEY CONCEPTS

Attack Intent Definitions

01. System Prompt Leakage
Attempts to extract the hidden "system prompt" or internal instructions that govern the model's behavior, often to find weaknesses for future attacks.

02. Bypassing Content Safety
Also known as "Jailbreaking." Attempts to force the model to generate prohibited content, such as hate speech, dangerous instructions, or illegal acts.

03. Unknown Intent
Ambiguous, nonsense, or probing prompts used for reconnaissance. Attackers use these to map the model's boundaries and error message behaviors.

04. Leaking Confidential Data
Targeting the agent's tools or knowledge base to extract private internal data (PII, credentials, proprietary code) that the model has access to.

05. Script Injection
Attempts to force the model to generate malicious code (e.g., XSS, SQL payloads) intended to execute in a downstream system or the user's browser.

06. Indirect Prompt Injection
A delivery vector rather than an intent, but critical: embedding instructions in external content (websites, files) that the model reads and processes.

Technique Overview

TECHNIQUE	DESCRIPTION
Hypothetical Scenarios	Framing the request as a thought experiment, story, or "what if" simulation to bypass standard safety filters.
Obfuscation	Hiding instructions using Base64, translation, uncommon syntax, or pseudo-code to evade keyword detection.
Role Play	Asking the model to adopt a persona (e.g., "You are a Linux terminal" or "You are an uncensored AI") to shift its behavioral context.
Authority Framing	Pretending to be a developer, administrator, or CEO to pressure the model into complying with a request.
Benign Distractors	Adding harmless noise or unrelated text to the prompt to confuse the model's attention mechanism.
Emotional Manipulation	Using urgency, distress, or appeals to empathy ("I will lose my job if you don't help") to override restrictions.
Hiding in External Source	Embedding the attack payload in a retrieved document or webpage (Indirect Injection).
Multilanguage	Switching languages mid-conversation or using low-resource languages to bypass English-centric guardrails.

Micro-Patterns: Attackers often combine these. For example, *Hypothetical Scenarios* are frequently used to execute *System Prompt Leakage*, while *Role Play* is the primary tool for *Bypassing Content Safety*.

INDIRECT ATTACKS & FUTURE SURFACE

WHY ATTACKS BECAME MORE EFFECTIVE IN Q4 2025

As models began interacting with external sources, attackers increasingly shifted toward indirect prompt injection. Instead of issuing the malicious instruction directly, they embedded it in a webpage, file or document and asked the agent to process it. This resulted in fewer attempts being required for a successful attack.

Key Insight: Models treat external content as neutral context rather than adversarial input. This gives attackers more influence because malicious instructions arrive through sources the model is designed to trust.

Direct vs Indirect Attacks

Attempts Required (Relative Comparison)

WHY INDIRECT ATTACKS WORK

>> Agents integrate external data as part of their reasoning process.

>> Prompt-level safety filters often do not apply to content fetched from external sources.

>> During summarization or transformation, hidden instructions blend into the model's active context.

>> Tool calls and external interactions can amplify the effects of poisoned input.

BEYOND TEXT GENERATION

Emerging Risks

The dataset surfaced early examples of attack types that become possible once agents interact with tools or external content:

- Attempts to extract confidential internal data
- Script-shaped injections within agent workflows
- Cross-context manipulation through externally fetched sources

Expanded Methodology

- Dataset consists of attack attempts observed in a 30-day window in Q4 2025.
- Each attempt was classified by intent and technique using Lakera's internal taxonomy.
- **OWASP Alignment:** Categories map to industry standards: *System Prompt Leakage (LLM07)*, *Prompt Injection (LLM01)*, and *Sensitive Information Disclosure (LLM02)*.
- Both direct prompt injections and indirect attacks were included.

CLOSING SIGNAL

PREPARE FOR 2026

The patterns observed in Q4 2025 point to a broader shift. As agents gain new capabilities and begin interacting with external systems, attackers adapt immediately.

Security must extend to every step an agent performs, not just the text it generates.