LLLM Security:

Securing Large Language Models Before They Cost You.

subrosa

Contents

subrosa

Artificial Intelligence (AI) has become a valuable tool for businesses, making tasks easier and faster. One of the most popular forms of AI today is Large Language Models (LLMs), like ChatGPT, which help generate text, answer questions, and automate conversations. However, along with their many benefits come significant security risks. This white paper will explore the hidden dangers of LLMs and show you practical ways to keep your AI tools safe and effective.

Understanding Large Language Models

Large Language Models are advanced computer programs that can understand and create human-like text.

Companies use LLMs for customer service, content creation, marketing, and even data analysis. These AI systems learn from reading lots of information from the internet and other sources. Once trained, they can respond to requests, produce reports, or even write software code.

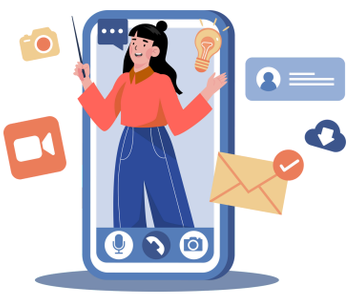
Because they understand context and intent, LLMs can perform complex tasks, from drafting documents to analyzing customer feedback. Their ability to automate routine interactions frees employees to focus on more critical responsibilities, significantly improving business efficiency.

LLMs in Action

LLMs power virtual assistants, automated email replies, content creation tools, customer service chatbots, and data analysis platforms used by businesses globally.

Why People Love LLMs...

Streamlined customer interactions and support



Enhanced content creation and productivity

Improved efficiency through automation of repetitive tasks

Accelerated decision-making through fast data processing

Attack Focus: Prompt Injection

Prompt injection involves attackers manipulating LLM responses by crafting malicious or misleading prompts.

This technique can exploit the AI's responses, leading it to reveal sensitive data or behave unpredictably. For instance, an attacker might deceive a chatbot into disclosing confidential financial information or customer details.

These attacks pose significant threats, including misinformation, reputational harm, regulatory penalties, and compromised customer trust. As businesses increasingly rely on LLM-driven customer interactions, understanding and mitigating prompt injection risks become essential to secure operations.

Red Flag Alert:

Beware of suspiciously detailed or unusual queries to your AI systems, as these could indicate attempts at prompt injection.

Common Prompt Injection Tactics

Overly detailed or repeated prompts

Prompts designed to confuse or mislead the AI

Requests for sensitive information disguised as legitimate queries

Case Example: A Realistic LLM Attack

SubRosa was engaged by this client to evaluate the security of a large language model (LLM) integrated into their internal knowledge management system. The firm relied on the AI assistant to draft case briefs, summarize filings, and surface past precedent. Our testing revealed critical vulnerabilities that could have led to data leakage and reputational harm.

1

Prompt Injection Discovery
During testing, SubRosa simulated adversarial prompts designed to override the assistant's instructions. The model revealed details from privileged internal memos — information that should have been inaccessible to end-users.

Adversarial Input Handling
Specialized test inputs were crafted to manipulate the assistant into delivering misleading summaries of case law. If left unchecked, these could have resulted in lawyers relying on inaccurate or adversarially biased information.

2

3

Immediate Client Engagement
Within hours of identifying the risk, SubRosa met with the firm's IT and compliance officers. Together, we reproduced the vulnerabilities and demonstrated how an external actor could exploit them through routine queries.

Preventive Measures and Safeguards
SubRosa recommended new guardrails: stricter input validation, role-based access controls, and fine-tuning the model with updated policy constraints. Forensic review confirmed no prior breaches, but the client implemented ongoing monitoring to detect and block injection attempts in real time.

4

The Result

SubRosa's timely intervention prevented the law firm's AI assistant from exposing confidential client information and misdirecting case research, averting both immediate reputational harm and long-term compliance consequences.

Risk Exposure Eliminated

The firm's LLM assistant no longer exposed privileged client data or internal memos. New guardrails and access controls stopped prompt injection attempts before they could yield sensitive results.

Operational Integrity Preserved

By mitigating adversarial input risks, attorneys could rely on the AI assistant for case research without fear of manipulation or misinformation. Productivity gains from AI adoption were preserved without introducing hidden liabilities.

Compliance Alignment

Remediation steps were mapped against the firm's regulatory obligations (client confidentiality, GDPR, ABA guidelines). This reduced potential liability in the event of regulatory review or legal discovery.

Client & Stakeholder Trust

By demonstrating proactive AI security testing, the firm reinforced its reputation for safeguarding sensitive information — a differentiator when winning and retaining high-value clients.

Foundation for Continuous Security

With monitoring and policy enforcement in place, the firm established a repeatable process for evaluating future AI integrations — ensuring that as the technology evolves, so does their security posture.

Key Takeaways

LLMs Expand the Attack Surface

Unlike traditional IT systems, large language models can be manipulated through natural language itself. That creates new, often overlooked risks for sensitive industries.

Real-World Impact is Tangible

Without testing, LLMs can expose confidential data, distort outputs, or be hijacked for malicious use. In legal, healthcare, and finance, these risks translate directly into liability and reputational damage.

Specialized Testing is Essential

Traditional pen testing stops at networks and endpoints. SubRosa's proprietary LLM penetration testing methodology closes this gap by simulating prompt injection, adversarial inputs, data leakage, and access control exploits.

Business Value is More than Security

Securing LLM deployments preserves productivity gains, ensures compliance, and strengthens client trust — delivering ROI beyond technical defense.

Continuous Monitoring is the Future

AI systems evolve constantly. One-off testing isn't enough. Building ongoing assessment and red-teaming into your security program is the only way to keep pace.

Our Company

We deliver cutting edge security technology solutions and services to our Clients so that they are prepared to tackle the ever growing cyber threat. Through our extensive expertise, we deliver threat intelligence, security controls validation and incident alerting and response. We employ and partner with some of the leading risk and security experts in the industry, enabling us to deliver effective services and software solutions to our clients of all sizes, across the globe.

More resources available at: https://subrosacyber.com

SUBROSA
2000 Auburn Dr
Ste 200 #366
Beachwood OH
44122
888.893.1983
info@subrosacyber.com

Since 2017 SubRosa proudly delivers cyber risk advisory and related services to public, private and government clients.

This brochure is to be used as a reference for general information only, and no content contained herein is designed to provide cyber risk professional advice or services.