CASE STUDY - Transforming AI-related risks into organisational resilience

From risk to resilience: securing a B2B legal assistant

Company
A professional services firm

Region
Europe

Intertek Solutions
AI Red Teaming engagement

"As a leading provider of AI Red Teaming and Agentic Assurance, Intertek helps organisations transition from risk to resilience. Our technical expertise ensures that high-stakes AI integrations, particularly those involving sensitive data, align with global security standards such as the EU AI Act, ISO 42001 and the OWASP Top 10 for LLM Applications."

Intertek recommended a multi-layered defence strategy, including pattern-based filtering for untrusted inputs and rigorous output sanitisation.

These targeted recommendations, paired with our hands-on support, empowered the client to harden the assistant's defences and launch the service with full confidence in its operational integrity.

By committing to ongoing monitoring and change-triggered assessments, the client ensures that the assistant remains secure against evolving threats and model drift, maintaining a resilient posture as the technology and its use case expand.

The challenge

The customer solution used a web-based front end and a Retrieval Augmented Generation (RAG) architecture to retrieve specialised data from a secure knowledge base.

Integrating Large Language Models (LLMs) into existing web functionality present significant security risks, and primary concerns included:

- Preventing malicious manipulation of the AI's output through Prompt Injection or Jailbreaking.
- Allowing only authorised users to interact with the model and its underlying data.
- Minimising the risk of the AI generating harmful or unethical content that could damage the firm's brand.

The solution

Intertek's security assessment included:

- LLM Adversarial Payload Testing: Using proprietary tooling to simulate real-world attacks like Prompt Injection and Jailbreaking.
- Web Application & API Testing: Evaluating the front-end interfaces to ensure traditional security flaws didn't compromise the AI environment.
- Cloud Configuration Review: Analysing the supporting infrastructure to ensure robust data sovereignty and security.

The result

The client had taken positive steps by limiting the AI to a defined user group and a standalone instance. However, Intertek identified high-severity vulnerabilities that required detailed remediation:

- Instruction Integrity Controls: Testing revealed that the system's handling of user-supplied input could be exploited to influence model behaviour beyond intended boundaries, including exposure of internal configuration.
- Input Validation Gaps: The content filtering mechanisms demonstrated insufficient robustness against varied input patterns, enabling certain categories of restricted content to pass through undetected.
- Context Window Integrity: Absent integrity controls over session context meant that interaction history could be manipulated in ways that affected the reliability of subsequent outputs.
- Configuration Exposure: Internal system configuration was retrievable through adversarial interaction patterns, revealing settings that did not align with security best practice.
- Through real-time reporting and structured retesting, identified issues were remediated efficiently - enabling the client to launch their service with confidence.

For more information

nta-sales-dept@intertek.com

Contact form

intertek.com/ai