CASE STUDY - Transforming AI-related risks into organisational resilience

Safeguarding the digital customer journey in financial services

Company
A leading financial services provider, offering a diverse portfolio of car, home, travel, and health insurance products

Intertek Solutions
AI Red Teaming

As a trusted cybersecurity partner, Intertek delivers specialised AI Red Teaming and Agentic Assurance to enable enterprise and government clients to operate securely within the evolving AI landscape.

Our approach integrates proprietary AI-powered tooling with expert manual investigation to simulate sophisticated adversarial attacks.

By mapping our findings to the OWASP Top 10 for LLM Applications, we provide a comprehensive evaluation of the chatbot or other AI interface's susceptibility to compromise.

"To modernise their user experience, our customer sought to embed an AI-powered chatbot into their website. This solution utilised a Retrieval Augmented Generation (RAG) architecture to provide anonymous users with instant, accurate information regarding the firm's specific insurance offerings."

The challenge

Integrating customer-facing AI presents a unique set of cybersecurity hurdles. The client needed to manage several critical risks to ensure a successful deployment:

- Data Safeguarding: Protecting both corporate intelligence and customer data from unauthorised access.
- Brand Reputation: Preventing "hallucinations" or unexpected AI behaviours that could tarnish the firm's professional image.
- Complex Attack Vectors: Addressing novel threats like Prompt Injection and Jailbreaking that traditional security measures often miss

The solution

The client engaged Intertek to perform an intensive, two-week AI Red Teaming exercise. Operating from the perspective of a standard anonymous user, our specialists conducted a "black-box" assessment to identify security flaws across the LLM application.

The result

Intertek identified several high-impact vulnerabilities, providing the client with a clear roadmap.

- Access Control Gaps: Client-side validation weaknesses could permit unauthorised access to restricted functionality, with implications for content integrity and regulatory compliance.
- Knowledge Base Integrity: Reliance on automated external data ingestion introduced risk, whereby unverified content could influence AI outputs. Remediation focused on transitioning to an internally governed, sanitised document repository.
- Guardrail Robustness: Despite reasonable defensive controls, testing identified techniques capable of circumventing safety boundaries, resulting in policy-violating outputs and unintended external referrals.
- Interface Boundary Controls: Insufficient domain restrictions allowed the AI interface to be rendered within untrusted external environments, outside its intended security context.
- Output Injection Risk: Adversarial inputs caused the model to produce responses containing malicious markup, with potential to facilitate credential harvesting - bypassing standard input-layer defences

The outcome

The engagement delivered a clear remediation roadmap, enabling confident launch and operation of the client's AI assistant while managing exposure to emerging threats. Intertek continues to support the client through ongoing testing cycles as their model evolve.

For more information

nta-sales-dept@intertek.com

Contact form

intertek.com/ai