THE HUMAN ELEMENT OF AI SECURITY

EXPERT TESTING FOR INTELLIGENT SYSTEMS

WHY THIS MATTERS TO YOUR ORGANIZATION

Whether you're building with OpenAI, Anthropic, Meta, or custom LLMs, AI vulnerabilities are not theoretical. They affect:

- **Trust** – Users expect safe, helpful, truthful responses
- **Compliance** – Output must align with regulatory frameworks (e.g., EU AI Act, FTC guidelines)
- **Brand Safety** – One viral screenshot of toxic or illegal content can damage years of reputation

You need assurance that your AI won't say, suggest, or generate the wrong thing — no matter how it's asked.

THE NEW CHALLENGE: TESTING INTELLIGENCE, NOT INTERFACES

As AI-powered applications rapidly evolve, so do the risks. But here's the catch: you can't secure large language models (LLMs) and generative AI systems using traditional penetration testing methods.

Why? Because LLMs don't behave like software. They behave like people.

Standard security tools and static payloads fail to account for the unpredictable, language-driven, context-sensitive nature of AI. Scanners might catch an open port, but they can't simulate a manipulative conversation or an adversary who speaks in metaphor.

Testing AI is not about code injection. It's about conversation exploitation. And that requires expert human testers with a deep understanding of both cybersecurity and language dynamics.

THE RISKS GO BEYOND PROMPTS

Organizations deploying LLMs today face a unique blend of technical and reputational risk:

- **Prompt injection** that bypasses business logic or content filters
- **Toxic or biased outputs** that could damage brand credibility
- **Misinformation** about legal, financial, or medical topics
- **Leaked sensitive data** from training artifacts or backend systems
- **Circumvented guardrails** via subtle context manipulation

These aren't just bugs. They're liabilities with consequences for customers, compliance, and public trust.

WHAT AUTOMATION MISSES

Today's off-the-shelf security tools have fundamental limitations when applied to AI systems. While they can detect known prompt injection vectors, they fall short in critical areas of language manipulation and social engineering.

For example, an automated tool can't effectively craft a series of positive toned messages to manipulate an LLM into generating prohibited output, nor can it trick a chatbot into revealing protected customer information through multilingual paraphrasing techniques. Similarly, automation can't use nuanced social engineering language patterns the way a human can, or subvert system guardrails by exploiting context window dynamics over extended conversations. These sophisticated attack vectors require human creativity, adaptability, and an understanding of both language and psychology that machines cannot yet replicate.

WHAT EXPERT TESTERS DO DIFFERENTLY

To be most effective, AI can't be tested like software. Instead, it should be tested like a person, applying adversarial prompt exploitation techniques grounded in human behavior and social engineering. Expert testers employ a sophisticated toolkit of conversational techniques that machines can't replicate. These include:

- **Emotional Preloading** – Building trust through benign dialogue before pivoting into malicious territory
- **Narrative Leading** – Starting with false facts to confuse retrieval-augmented generation (RAG) systems
- **Negative Casing** – Framing restricted content as examples of what not to do, bypassing filters
- **Content-Adjacent Prompting** – Describing dangerous or banned content in indirect terms the AI reassembles
- **Language Nesting** – Translating prompts through multiple languages to evade English-trained guardrails

This testing is creative, high-skill engagement built around how adversaries actually behave.

REAL-WORLD EXAMPLE: MANIPULATING THE MODEL

In one recent assessment, our team successfully bypassed a commercial AI safety system through a carefully orchestrated sequence of interactions.

The process began by establishing rapport through friendly, professional dialogue that set a tone of trust and cooperation. Once this foundation was laid, our testers gradually pivoted the conversation toward political themes, introducing them in ways that didn't trigger content filters. This contextual preparation created the conditions to prompt the system to generate a complete memetic warfare campaign designed to destabilize a Western democracy.

Despite the supposed protective controls in place, the system responded with detailed messaging strategies, influencer budget plans, and campaign slogans—demonstrating a significant failure of safety guardrails. This kind of nuanced, context-aware failure simply wouldn't be caught with automation or payload lists alone.

HOW BISHOP FOX CAN HELP

The security landscape for AI and LLMs is changing fast and so are the techniques to test and defend them. At Bishop Fox, we don't claim to have all the answers in a space that's evolving by the day. But we do bring the experience, curiosity, and rigor needed to keep pace with real-world threats.

Our offensive security experts are not just skilled in traditional penetration testing, they're constantly experimenting with new adversarial techniques, sharing what we learn with the broader community, and refining our approach as the technology shifts.

Here's what that means for you:

REALISTIC THREAT SIMULATION
We replicate how actual adversaries manipulate language, context, and system behavior.

HUMAN-LED CREATIVITY
Our testers use dynamic, psychological methods to uncover hidden behavior, not just known exploits.

DEFENSIBLE, CONTEXT-RICH REPORTING
We deliver clear findings with full conversational transcripts, caveats where needed, and remediation guidance grounded in your AI stack.

COLLABORATIVE KNOWLEDGE BUILDING
We partner with our clients to stay ahead of emerging threats, and contribute to community learning.

If your organization is investing in AI, we're here to help you test responsibly, learn rapidly, and defend proactively—so you can innovate with confidence.

ABOUT BISHOP FOX

Bishop Fox is the leading authority in offensive security, providing solutions that help organizations secure their most critical assets against sophisticated cyber threats. Since 2005, we've partnered with Fortune 100 enterprises and high-growth innovators to deliver high-impact security testing and advisory services. Our comprehensive service offerings include tech-enabled, human-driven continuous threat exposure management, red teaming, and penetration testing for applications, cloud, networks, IoT, and AI/LLM.

LEARN MORE AT BISHOPFOX.COM
FOLLOW US ON X