SERVICE DEFINITION
AI / LLM PENETRATION TESTING

Our address
Unit 3E-3F, 33-34 Westpoint,
Warple Way, Acton, W3 0RG

Give us a call
0333 939 8080

Send us a message
hello@jumpsec.com

Find out more
www.jumpsec.com

About JUMPSEC

JUMPSEC is a UK-based cyber security consultancy, established in 2012, with a proven track record in offensive testing, incident response, managed detection, and advisory services for both public and private sector organisations. Our work is led by skilled professionals with extensive experience in red teaming, security engineering, and incident response, ensuring that our services are rooted in real-world knowledge and practical expertise.

We place a premium on assurance. Our teams include CHECK and CREST-qualified penetration testers, as well as NCSC CIR-approved incident responders and NCSC CIE-accredited Incident Managers, enabling us to deliver a wide range of government-level security engagements. We have deep experience deploying and managing Microsoft security technologies such as Microsoft Defender and Sentinel, particularly within regulated environments. Our approaches and documentation are consistent with recognised standards including the NIST Cyber Security Framework, CIS Controls and CAF, and our reporting is clear and purposeful, catering to both technical specialists and executive decision-makers.

Clients use JUMPSEC to achieve three outcomes:

- **to reduce the risk of significant security incidents**
- **to speed up the detection and containment of threats, and**
- **to demonstrate robust governance to stakeholders such as boards, auditors, and insurers**

Our services are transparent and collaborative, as we provide our clients with clear scopes, well-defined engagement terms, daily updates during testing, and strong evidence to support their security posture. We prioritise remediation advice based on business impact, not just automated scan results.

Protecting client data is at the heart of our operations. JUMPSEC acts as a data processor under UK GDPR and DPA, adheres to principle-of-least-privilege access, and only retains evidence for periods agreed with clients. Where necessary, our personnel hold BPSS or SC clearance, and we can provide reports and artefacts in accessible formats to meet WCAG 2.1 AA standards.

Our public sector experience covers central and local government, healthcare, education, and critical national infrastructure. In the private sector, we support clients in finance, manufacturing, logistics, and technology. We often deliver projects alongside complementary services, such as red and purple teaming, detection engineering, and incident response retainers, to drive measurable improvements and tangible returns.

We continually refine our services to address the evolving needs of our clients, frequently developing tailored solutions to unique challenges. Our approach is adaptable and shaped by each client's specific situation, so the best way to understand how we can help is to discuss your needs directly with us before RFP release.

This overview outlines JUMPSEC's broad capabilities and demonstrates how our expertise can be applied to deliver practical, threat-focused security for your organisation. To find out more about how we can address your cyber security requirements, please get in touch for a straightforward conversation.

Example use cases

AI productivity platform — Prompt injection to SaaS connectors

JUMPSEC found that a workplace assistant connected to email, calendar, storage, and ticketing exposed broad tool affordances. Adversarial prompts embedded in shared docs coerced the model to summarise hidden content, then trigger create-event and file-share tools with attacker-supplied parameters. Guardrails relied on post-generation moderation and generic “do not” policies.

JUMPSEC’s recommendations hardened constrained tool intents with strict parameter schemas, added user-present confirmation checks, and enforced resource-scoped access tokens. Retrieval filters stripped embedded instructions; outputs passed through allow-listed renderers. A CI prompt suite and safe tool stubs verified blocked injection chains, correct consent flows, and refusal to act on out-of-context instructions.

AI-enabled diagnostic device — RAG leakage and unsafe decision support

JUMPSEC identified that a clinical decision support workflow used retrieval-augmented generation (RAG) over protocols and device logs to explain ultrasound outputs. By crafting specific inputs, we demonstrated how context could pivot to archived validation datasets, exposing study snippets and prompting off-label adjustments. Our assessment revealed tooling that exposed a calibration function without clinician-in-the-loop checks.

To remediate these risks, JUMPSEC isolated per-patient contexts, enforced corpus allow-lists with query rewriting, and mandated signed clinician confirmation before any calibration or setting change. We made provenance labels and citation thresholds compulsory in explanations. Our regression tests confirmed no cross-record leakage, ensured safe refusal for off-label prompts, and verified that all actions were audited and clinician-approved.

Local authority — cross-tenant leakage via vector store mis-scoping

JUMPSEC guided a multi-tenant knowledge assistant serving several service lines. Embedding and retrieval operated within a shared index using soft namespace tags. JUMPSEC demonstrated how adversarial queries and wildcard tokens could extract passages from another service’s documents. To resolve these issues, JUMPSEC transitioned to hard index isolation, enforced tenant filters at the gateway, and implemented semantic allow-lists for each collection. Query rewriting was used to prevent wildcard escalation. Continuous integration tests confirmed tenant isolation and successfully blocked cross-collection queries while preserving relevant answers.

Service Overview

JUMPSEC delivers a manual-first security assessment of AI-enabled applications and LLM integrations to evidence exploitability and land practical fixes. The service covers system and user prompt layers, guardrails, retrieval-augmented generation (RAG), tools/functions and orchestration, vector stores and data pipelines, API gateways, and downstream actions. Testing focuses on the attack chains that matter in production: prompt injection and jailbreaking, insecure output handling and tool abuse, RAG leakage and document boundary bypass, supply chain weaknesses across models, embeddings and plugins, and privacy or safety failures that create real business impact. Findings translate into precise, low-friction controls, with replayable artefacts and verification tests that survive future releases.

Objectives

- Evidence if the application can be induced to disclose sensitive data, execute unsafe actions, or route around intended controls.
- Stress guardrails and tool-use boundaries to confirm safe orchestration under realistic adversarial prompts.
- Harden retrieval, context-building, and output handling to prevent cross-tenant or cross-document leakage.
- Deliver exact policy, code, and platform changes with acceptance criteria and CI-ready regression tests.

Approach

Our engagements start with scoping to confirm models, orchestration framework, retrieval stack, tools/plugins, data classifications, and acceptable use boundaries. JUMPSEC analyses system prompts and safety policies, then exercises adversarial inputs that chain injection with tool calls, RAG pivots, or downstream actions. Testing deliberately targets model overreliance and excessive agency, validates output sanitisation and content provenance, and inspects the supply chain for model, embedding, or plugin trust issues. Work remains non-destructive and uses synthetic or redacted data wherever possible. Recommendations are sequenced for safety, include rollback, and align to your change windows.

Method

- **Plan** — agree scope, artefacts, data handling, and safety gates for tool execution; define success measures and rollback.
- **Discover** — review system prompts, guardrails, routing rules, tool definitions, vector store schemas, retrieval filters, and content moderation.
- **Val**idate — exercise adversarial scenarios against prompts, tools, and RAG: injection chains, context pivoting, output handling, function abuse, and privacy leaks; confirm exploitability with constrained payloads.

- **Analyse** — link each weakness to realistic business impact; define exact fixes (prompt hardening, tool policy, retrieval filters, content sanitisation, gateway controls) with acceptance criteria.
- **Debrief** — present actions, owners, sequencing, verification tests, and a short retest plan; agree CI integration for regression..

Deliverables

- Executive briefing that explains material risks, affected user journeys, and expected risk reduction after change.
- AI/LLM Security Report with evidence for each finding, minimal reproducible prompts/calls, impact narrative, exact remediation steps (policies, code, config), rollback notes, and acceptance criteria.
- Guardrail and Orchestration Hardening Pack covering system prompt patterns, tool allow/deny and parameter validation, routing constraints, moderation thresholds, output sanitisation, and provenance/labelling.
- RAG Safety Pack including retrieval filters, document boundary enforcement, query rewriting controls, vector store tenancy isolation, and leakage tests with verification artefacts.
- Regression bundle for CI/CD: prompt suites, safe tool stubs, negative tests for leakage and unsafe actions, and guidance to integrate into build pipelines.
- Standards and alignment embedded in reporting: OWASP Top 10 for LLM Applications (LLM01–LLM10), OWASP ASVS mappings for auth and output handling, OAuth/OIDC for delegated access, and NCSC secure design principles for UK buyers

Outcomes and benefits

- Lower probability of sensitive information disclosure or unsafe tool execution.
- Guardrails and orchestration that resist prompt injection while preserving usability.
- Retrieval and multi-tenant boundaries that survive adversarial queries.
- Repeatable tests that keep protections in place through future model or policy changes

Pricing model

Fixed price per application/integration, or time-and-materials for complex estates. Price drivers include number of tools/functions, RAG complexity, tenant/role matrix, third-party plugins, and verification depth.

Timelines

- Planning and access preparation: 3–5 working days.
- Testing and analysis: 5–10 working days per application and role/tool set.
- Report issue: 5–10 working days after testing completion.

- Optional re-test and CI integration support by agreement.

Team and roles

- Engagement Lead for governance, cadence, and quality.
- AI Security Tester(s) for adversarial prompting, orchestration analysis, and RAG testing.
- Platform Engineer for gateway/policy hardening and CI integration.
- Coordinator for logistics, artefacts, and evidence handling.

Buyer responsibilities

- Nominate a product owner and technical contact for the AI stack.
- Provide non-production access, system prompts/policies, tool definitions, and synthetic datasets.
- Align change windows and owners for prompt/policy/code updates.

JUMPSEC responsibilities

- Operate within agreed safety gates and authority.
- Use non-destructive techniques and minimise data handling.
- Provide precise, auditable fixes with verification artefacts and rollback.

Data protection and privacy

The buyer is data controller; JUMPSEC acts as processor. Only minimum necessary evidence is collected and retained. Secure transfer, strict access control, and agreed retention periods apply. Artefacts are returned or securely destroyed on completion.

Constraints and assumptions

- No live customer data unless explicitly approved and masked.
- No denial-of-service of model endpoints; load is controlled.
- Third-party plugin or model testing requires written consent.
- Personnel with BPSS/SC available where required
- Reports available in accessible formats (WCAG 2.1 AA)

Related services

- API Penetration Testing
- Secure SDLC / DevSecOps / Pipeline Assessment
- Threat Modelling

Unit 3E-3F, 33-34 Westpoint,
Warple Way, Acton, W3 0RG

0333 939 8080

hello@jumpsec.com

www.jumpsec.com