Cyber **security**

AI PENETRATION TESTING SERVICES

THREAT READY





LRQA.COM

LRQA

Your Risk Management Advantage

CONTENTS

Our Accreditations 03

Our AI Penetration Testing Services 04

AI Services and Technologies 05

Our Approach, Reports, and Deliverables 06

OUR ACCREDITATIONS

As a leading penetration testing company, LRQA holds the most coveted accreditations across the world.

- We are the only organisation in the world with a full suite of CREST accreditations;
- We are a proud member of the UK Government's NCSC CHECK scheme. Our team of testers includes CHECK team leaders within infrastructure and web applications, as well as CHECK team members; We are an ISO 9001, ISO 14001, and ISO 27001 certified organisation, and as such, we guarantee to provide our clients with top-quality services from within a rigorously controlled environment;
- LRQA are certified by a variety of governing bodies for the work that we regularly perform within a range of highly regulated industries, including the financial and payment card sectors;
- We are an accredited supplier of CREST and an approved provider of STAR testing services. We also deliver services for both TIBER and iCAST in heavily regulated financial services environments;
- Our security consultants hold a variety of different certifications across a multitude of different areas of information security. Consultants possess penetration testing-specific certifications such as those provided by Offensive security (OffSec) (OSCP, OSWE, OSEP, OSED, OSEP, etc.)
- The LRQA security testing team also includes those provided by Cyber Scheme, such as Cyber Scheme Team Leader (CSTM) and Cyber Scheme Team Leader (CSTL).
- All of our penetration testers have been fully background- checked;
- In addition, our team is comprised of industry-recognised consultants and published authors who have been acknowledged by the media and the cybersecurity community.

OUR AI PENETRATION TESTING SERVICES

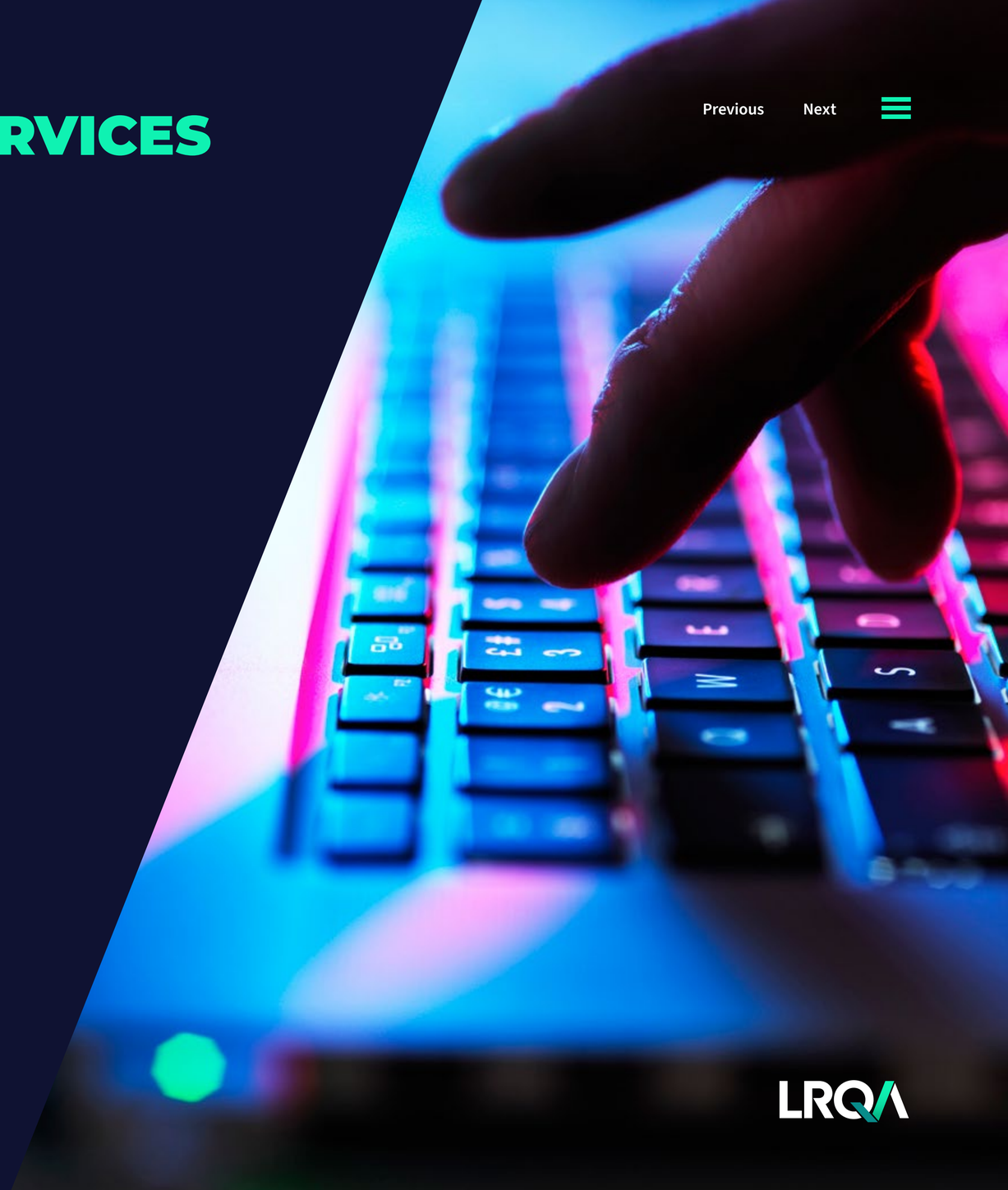
INTRODUCTION

Advancements in Artificial Intelligence (AI) are reshaping how organisations harness technology. With machine learning algorithms and large language models (LLMs), businesses are deploying retrieval augmented generation (RAG) systems, chatbots, and agentic workflows powered by Model Context Protocol (MCP) to transform decision making and customer engagement.

Yet this transformation introduces new security challenges. AI systems expand the attack surface beyond traditional application security: prompt injection on chatbots and UI endpoints, MCP authentication failures, over permissioned agent tools, and data leakage through misconfigured RAG pipelines. The deep integration of LLMs with enterprise infrastructure magnifies the risk, making robust AI security essential.

Securing AI requires a lifecycle approach. From data governance and model validation to chatbot hardening, backend access controls, and continuous monitoring, every stage must be protected. Key measures include safeguarding RAG retrieval integrity, enforcing MCP authentication and agent permissions, sanitising inputs to block prompt injection, and preventing unauthorised data disclosure. As LLMs connect to critical systems, ongoing evaluation and adaptation of security controls is vital.

LRQA's AI Security Review delivers comprehensive assessments across these dimensions—covering models, chatbots, infrastructure, RAG pipelines, MCP configurations, and access controls. By proactively identifying and mitigating vulnerabilities, organisations can protect their AI investments while ensuring data integrity and confidentiality.



WHY CONDUCT AI PENETRATION TESTING?

AI penetration testing is imperative to gauge the resilience of AI-driven applications and the infrastructure underpinning them. As organisations increasingly integrate AI technologies into their operations, ensuring the security of these systems becomes paramount. Given the unique security challenges posed by AI systems, proactive measures are essential to effectively mitigate risks.

WHAT PROBLEMS DOES IT ADDRESS?

AI security testing serves to tackle various critical concerns, including:

- **Preventing data breaches:** AI systems process sensitive data and connect to critical applications. Penetration testing uncovers AI-specific weaknesses such as prompt injection, insecure RAG pipelines, and over-permissioned agents that can lead to breaches, financial losses, and reputational damage. Aligned controls with MITRE ATLAS and OWASP's LLM guidance strengthen detection and mitigation across the AI lifecycle.
- **Meeting compliance requirements:** Beyond HIPAA, PCI DSS, and GDPR, organisations must align with EU AI Act requirements on robustness and cybersecurity. NIST's AI RMF and ISO/IEC 42001 provide structured governance and risk management across design, deployment, and monitoring.
- Ensuring secure DevOps practices: Integrating AI penetration testing into CI/CD enables early detection and remediation. Embedding controls mapped to NIST AI RMF and ISO/IEC 42001 supports secure MLOps/DevSecOps, addressing vulnerabilities before production and post-release.
- Mitigating threats: Using MITRE ATLAS and OWASP's GenAI/LLM Top 10, teams proactively identify and remediate risks like data poisoning, model extraction, and prompt injection. Continuous testing reduces attack surface and protects AI investments.



AI SERVICES AND TECHNOLOGIES

TYPES OF SERVICES OFFERED:

AI Security Testing

In recent years, the proliferation of Artificial Intelligence (AI) has reshaped industries and consumer experiences alike. Among its myriad applications, chatbots have emerged as pivotal tools in customer service, support systems, and various interactive platforms. To highlight its significance, prompt injections have recently been ranked Number 1 in the OWASP LLM Top 10, a list that catalogues the most pressing vulnerabilities in Large Language Models like chatbots.

LRQA specialises in evaluating and fortifying the security of AI systems, including Large Language Models (LLMs). Our comprehensive approach involves utilising a range of prompt manipulation and exploitation techniques to rigorously assess the robustness of your LLM-based systems.

Examples of such vulnerable applications include, but are not limited to:

1. Chatbots and Virtual Assistants
2. Sentiment Analysis Tools
3. Recommendation Engines
4. AI Translation Systems
5. Image Recognition Systems

By partnering with LRQA you can ensure that your AI systems are fortified against the risks posed by prompt injections, safeguarding both your data and your customers' experiences.

Security of AI Systems

As organisations deploy AI systems publicly and internally, threat models differ fundamentally. Public-facing systems face direct external attack; custom deployments require deep infrastructure assessment. Our comprehensive penetration testing services evaluate both scenarios following OWASP LLM Top 10 guidelines and emerging AI governance frameworks.

Public-Facing AI Systems

Your chatbots, APIs, and recommendation engines operate as direct attack surfaces. We conduct adversarial testing to extract training data through prompt injection and jailbreak attempts, identify hallucinations and factual accuracy failures that damage brand, assess bias and discrimination across protected characteristics, evaluate model extraction resistance, and uncover output validation weaknesses leading to downstream exploitation.

Custom Model Deployments

When you control the infrastructure, assessment runs deeper. We interrogate models directly to extract embeddings and fine-tuning data, evaluate RAG pipeline security and training data integrity, test MCP server authentication and agent privilege escalation, identify integration attack chains targeting databases and internal systems, and validate compliance with GDPR, HIPAA, NIST AI RMF, and ISO/IEC 42001.

Our testing includes:

1. Model extraction and embedding reversal
2. Hallucination and guardrail bypass
3. Bias and fairness assessment
4. Data pipeline integrity
5. Compliance framework alignment

LRQA identifies and remediates vulnerabilities before attackers do—protecting your AI investments, customer trust, and operational integrity.

OUR APPROACH, REPORTS AND DELIVERABLES

OUR APPROACH TO PENETRATION TESTING

LRQA has a robust testing methodology that extends across infrastructure and application testing engagements. Although every penetration test is tailored to our client's individual needs, we follow the same proven methodology so as to maintain a consistent set of results.

- **Phase 1:** Scoping
- **Phase 2:** Reconnaissance and Enumeration
- **Phase 3:** Mapping and Service Identification
- **Phase 4:** Vulnerability Analysis
- **Phase 5:** Service Exploitation
- **Phase 6:** Pivoting and General Post-Exploitation
- **Phase 7:** Reporting and Debrief

TESTING REPORT AND DOCUMENTATION

You will receive a high-level management report and an in-depth technical review document for each engagement. The technical report outlines the engagement details and provides comprehensive vulnerability findings, including descriptions, severity ratings, affected components, contextual impact, screenshots, and detailed remediation guidance, with supplementary documents included for larger environments. The management report offers an executive-level overview of the engagement, summarising key results, overall risk, major themes, and both strengths and concerns, and concludes with a prioritised list of recommended actions.

TEST DEBRIEF

LRQA ensures that all tests have a full debrief at the end of the engagement. If required, LRQA can deliver this debrief in a face-to-face manner. During this process, we will provide a presentation of critical and high-level vulnerabilities, along with guidance on remediation and countermeasures.

When a face-to-face debrief is not required, LRQA conducts debriefs via a video conference. Through this approach, we are still able to have a comprehensive presentation of vulnerabilities and areas identified as being high risk.

ABOUT LRQA

LRQA is a leading global risk management partner.

Through our connected risk management solutions, we help you navigate an evolving global landscape to keep you one step ahead.

From certification and cybersecurity, to safety, sustainability and supply chain resilience, we work with you to identify risks across your business. We then create smart, scalable solutions, tailored to help you prepare, prevent and protect against risk.

Through relentless client focus, backed by decades of sector-specific expertise, data-driven insight and on-the-ground specialists across assurance, certification, inspection, advisory and training, we support over 61,000 organisations in more than 150 countries.

LRQA – Your risk management advantage.

GET IN TOUCH

Visit **LRQA.com** for more information or email **enquiries.uk@lrqa.com**

LRQA
1 Trinity Park
Bickenhill Lane
Birmingham
B37 7ES
United Kingdom

Care is taken to ensure that all information provided is accurate and up to date; however, LRQA accepts no responsibility for inaccuracies in or changes to information.

LRQA
Your Risk Management Advantage