

Cyber Security Agentic AI and LLM Test and Assurance

A QinetiQ Cyber Security Service

Key Benefits

- Enables Trusted AI Use in Mission Critical Environments
- Test and assure Artificial Intelligence (AI) Agents and underlying Large Language Models (LLM).
- Can be extended to include hosting infrastructure or to collaborate with those testing and assuring related infrastructure for a holistic test and assurance service.
- Highlights security threats and risks.
- Allows identification of mitigating controls to allow the management of residual risks.

Artificial intelligence is increasingly being embedded in Defence, Intelligence, and other highly regulated systems, delivering clear operational benefits but also introducing new and unfamiliar risks. Large language models and agentic AI systems behave probabilistically, adapt over time, and can be influenced in ways that are difficult to predict or fully control using traditional systems assurance approaches. For organisations operating in high consequence environments, this creates a pressing need for independent, evidence-based assurance of AI behaviour.

QinetiQ's AI Test & Assurance Service is designed to meet this need. It provides customers with a practical means to understand how AI systems behave, where risks may arise, and what controls or mitigations are required to enable safe, secure, and compliant deployment.

The QinetiQ Approach

Our service delivers a structured, test-driven assessment of AI systems as they are intended to be used in practice. Rather than relying solely on vendor claims or high-level policy reviews, we focus on observing real system behaviour under expected, stressed, and adversarial conditions.

Key elements of the service include:

- AI model behaviour assessment, examining the safety, reliability, robustness, and security characteristics of language models and similar AI components.
- Evaluation of susceptibility to misuse and manipulation, including prompt injection, jailbreak techniques, unsafe output generation, and information leakage.
- Comparative testing of models and configurations, enabling informed choices during procurement and system design.
- Assessment of Acceptable Use Policies and Terms of Service, helping customers understand provider limitations, assumptions, and residual risks.

Assurance of Agentic and Tool Using AI Systems

Where AI systems extend beyond a single model into multiagent or tool-enabled architectures, QinetiQ provides system level assurance that considers the full operational context. These systems introduce additional complexities due to autonomy, delegation, and multistep decision making.

Our assurance activities focus on:

- How agents are orchestrated and supervised.
- How tools are selected, invoked, and controlled.
- How context, memory, and state are managed across task steps.

QINETIQ IN CONFIDENCE

- How errors, unexpected outputs, or unsafe decisions propagate through workflows.
- Ensuring risks introduced by autonomy and automation are understood and appropriately governed.

Adversarial Testing and Standards Aligned Reporting

A core differentiator of the QinetiQ service is structured adversarial testing. Drawing on recognised attack patterns, including MITRE ATLAS, we simulate realistic misuse, manipulation, and failure scenarios to test system resilience. This provides insight into risks that are unlikely to be identified through functional testing alone.

Findings are translated into clear assurance evidence and mapped to recognised frameworks and standards, including:

- NIST AI Risk Management Framework (AI RMF).
- ISO/IEC 42001 (AI Management Systems).
- OWASP Top 10 for LLM Components, OWASP Top 10 for Agentic AI.
- Relevant regulatory and governance obligations, such as the EU AI Act.
- This supports both internal assurance and external audit or regulatory engagement.

Clear, Decision Focused Outputs

The service produces clear and actionable outputs designed for both technical teams and senior decision makers. Reporting focuses on explaining what was tested, what risks were identified, and why they matter in an operational context. Recommendations prioritise practical mitigations, design improvements, and governance controls rather than abstract compliance statements.

Customers typically receive:

- A clear summary of AI risks and limitations.
- Evidence to support deployment, procurement, or assurance decisions.
- Prioritised recommendations for control enhancement and risk reduction.

Other QinetiQ Cyber Services

QinetiQ combines deep experience in Defence, Security, and system assurance with hands on expertise in AI testing, adversarial evaluation, and assurance frameworks. Our approach is grounded in practical testing and operational realism, enabling customers to deploy AI with confidence rather than uncertainty.

This service forms part of a wider service portfolio, which helps organisations mature their digital resilience and to build and evidence confidence in their ability to deal with cyber-attacks through a holistic approach.

This service is a sub-service of Cyber Security Advisory.

.....

QINETIQ

Why QinetiQ Cyber

- Unmatched team of vetted, expert cyber consultants, architects, engineers and human performance scientists
- Unique experience and patented intellectual property in securing the world's most complex, safety critical environments
- UKCSC registered cyber professionals, NCSC Assured Services

Copyright QinetiQ Ltd 2026 | CYBER SECURITY RISK ASSESSMENT SERVICES QINETIQ/20/01806

For further information please contact:

Malvern Technology Centre St
Andrews Road, Malvern
Worcestershire, WR14 3PS
United Kingdom
+44 (0)1252 392000
cyberenquiries@QinetiQ.com
www.QinetiQ.com