Deloitte.

Together makes progress

Trusted & Secure AI (TSAI)

AI Application Risk Management

The TSAI GenAI Implementation Guidebook provides a curated process to help improve the safety, security, reliability, and robustness of AI deployments by supplying a structured framework to help assess, address, test, and monitor AI risks. Our guidebook assists in addressing risks related to privacy, transparency and explainability, fairness, responsibility, accountability, robustness and reliability, and safety and security.

A step-by-step process for embedding trust and security into GenerativeAI (GenAI) deployments

01

Assess AI risk profile

Systematically evaluate AI deployments to identify risks related to safety, privacy, and security. This process includes providing recommendations for controls that help organizations identify, mitigate, and monitor these risks throughout the AI life cycle.

02

Implement controls to safeguard AI systems

Provide accelerators and guidelines for prompt engineering, guardrail instantiation and configuration, and protection of both structured and unstructured data. These measures support large language model (LLM) applications in meeting established design and security requirements.

03

Conduct Testing on Output Quality and in testing on output quality and risk

Perform automated point-in-time, pre-deployment testing on AI outputs to determine their quality and risk profiles. Use testing results to adjust application configurations, enhancing output quality and minimizing potential risks.

04

Enable trusted and secure AI operations

Implement continuous testing, real-time monitoring, and comprehensive reporting to deliver ongoing insights into AI system performance. These practices support the proactive identification and remediation of emerging vulnerabilities within AI operations.

Trusted and Secure AI. **AI built with confidence.**

Deloitte's TSAI GenAI Implementation Guidebook provides tailored LLM application risk management to unlock GenAI at scale. From inception through implementation, these seven capabilities, including tools and accelerators, can help provide an effective approach to identifying, mitigating, and managing GenAI risks.

Risk assessment

Evaluate LLM application risk to identify effective controls and testing needed to mitigate and manage AI risk.

Capability: Streamlined risk assessment survey designed to target unique use case risks. Customized risk assessment scorecard presenting risk profile and tailored recommendations for risk mitigation strategies.

Application risk management

Configure prompt engineering and Guardrails to confine LLM applications to operate within design specifications, while protecting data and enhancing output accuracy.

Capability: Automated instantiation and configurations for cloud service providers (CSPs) or vendor guardrails, customized to mitigate risk identified through the risk assessment and scorecard process.

Unstructured data protection

Scanning of vector database content to identify, anonymize, omit or metatag sensitive information for retrieval augmented generation (RAG) tasks.

Capability: Automated scripts utilized to scan vector database chunks for prohibited (PII, classified, encryption & API keys, code, foreign languages, etc.) or malicious data (prompt injections).

Risk acceptance testing

Evaluation of AI systems inputs and outputs to identify risks and vulnerabilities, identify errors, and provide recommendations to balance security, safety, trustworthiness, and reliability.

Capability: Automated prompt testing designed to identify risks and vulnerabilities across 16 risk dimensions, helping to balance safety, security, accuracy, and performance.

Output quality testing

Iterative testing and refinement procedures designed to enhance LLM performance and produce consistent, transparent, and trustworthy outputs.

Capability: Automated prompt testing designed to evaluate output quality, identifying potential inaccuracies or inconsistencies that can be addressed through RAG variable adjustments.

Monitoring & reporting

Logging and monitoring of relevant data and events across LLM-enabled applications and infrastructure in near real time to detect, alert, and respond.

Capability: LLM application-specific KPIs and recommendations utilizing CSP-specific logging and monitoring capabilities to enable real-time monitoring, reporting, and auditability.

Continuous testing

Ongoing, periodic testing of AI system vulnerabilities by simulating threat actor tactics, combining manual and automated methods across cyber, human, and physical dimensions.

Capability: Automated and manual procedures designed to evaluate LLM-enabled application's performance and security as it operated in an emerging threat environment.

Get in touch

Joe Conti
Trusted & Secure AI Leader
Deloitte & Touche LLP
joconti@deloitte.com

Ram Regunathan
Trusted & Secure AI Leader
Deloitte & Touche LLP
rregunathan@deloitte.com

Copyright © 2026 Deloitte Development LLC. All rights reserved.