AI / LLM Security

AI / LLM Security

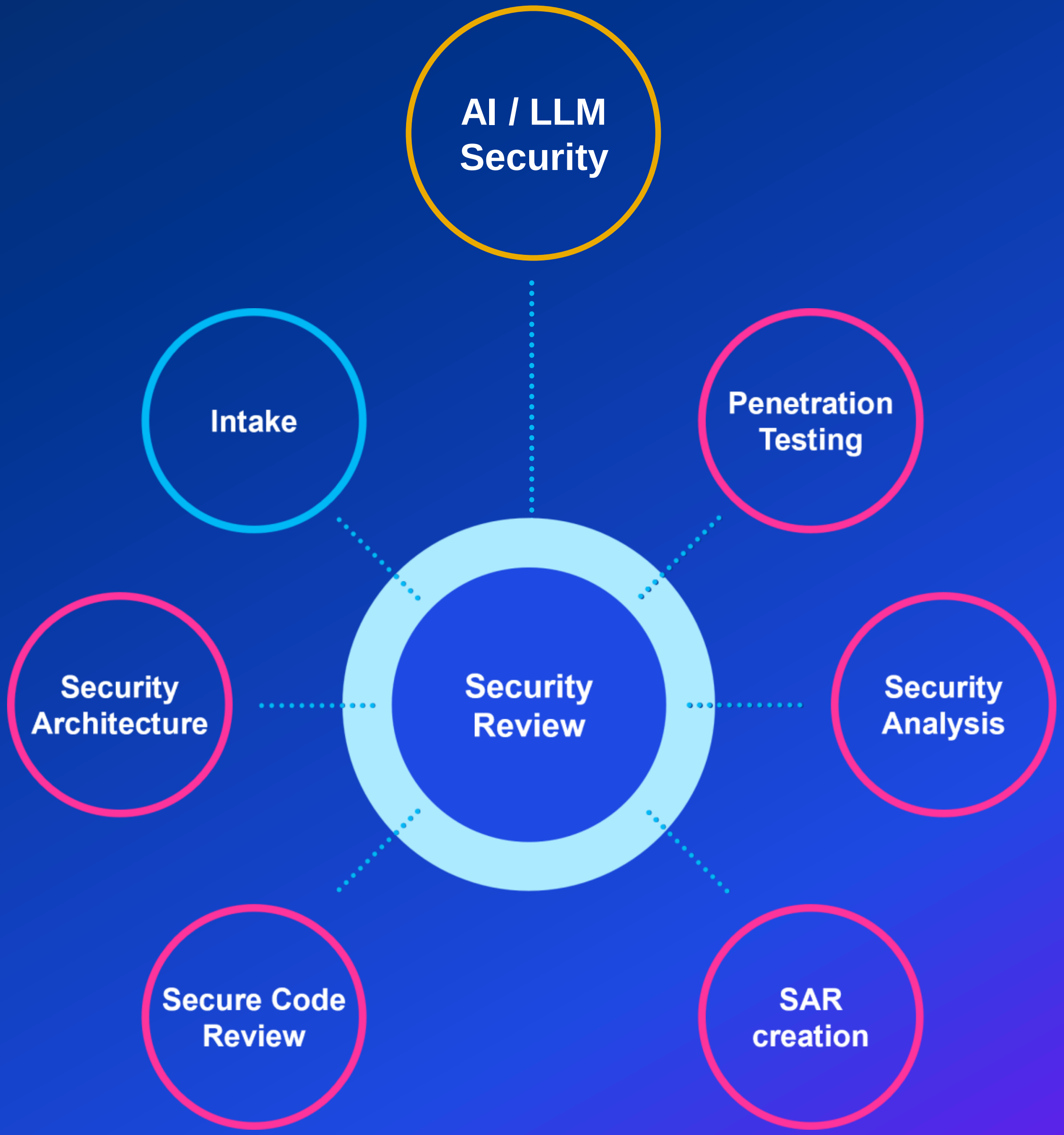
As more and more solutions are developed or updated with new AI functionalities, Member Firms must keep up with the security challenges of LLM and AI, and update their own processes of the solution reviews.

The utilization of LLMs in applications can introduce a variety of complex and unique security threats, that may affect internal resources, company reputation, client data, intellectual property, or give a foothold to further attacks.

Penetration testing and code reviews might not be enough anymore, and AI should be considered as a scope component too.

KPMG Hungary’s services can be use to extend the solution review with special focus on a solution’s AI component, which are the followings:

1. LLM pentesting - adversarial testing and attack simulation
2. AI Compliance
3. AI review – AI/ML development best practices



Adversarial testing and attack simulation

We developed a novel methodology for security testing that focuses on Generative AI and Large Language Models from an attacker perspective. This approach is anchored in frameworks developed by reputable organizations, such as OWASP Top 10 for LLM and MITRE ATLAS™ and is continually refined with insights from the latest research and developments in AI security.

Our testing covers direct and indirect prompt injections, system prompt retrieval, leaking training data, denial of service, cross-site scripting and code execution attacks, jailbreaking and bypassing access control implemented through system prompting. We leverage our expertise in web security testing in context of AI security to reveal the impact of common design and logic flaws of web applications on the behavior and functionality of AI solutions.

During penetration testing of genAI applications we cover the following elements of OWASP LLM Top 10 (2025):

- prompt injections (direct & indirect)
- sensitive information disclosure
- improper output handling
- excessive agency
- system prompt (initial instructions) leakage
- vector and embedding weaknesses
- misinformation
- unbounded consumption
- in special cases: data poisoning

The attack simulation includes the following elements of MITRE ATLAS threat matrix:

- obtain capabilities
- develop capabilities
- in special cases: poison training data
- evade ML model
- LLM Prompt injection (direct & indirect)
- AI Model Inference API access
- Command and Scripting Interpreter
- LLM Plugin compromise
- LLM Prompt self-replication
- LLM Jailbreak
- LLM Meta Prompt Extraction
- LLM Data Leakage
- Cost harvesting

Compliance and regulatory

KPMG HU can help you operationalize AI Governance Target Operating Model Components

Regulatory Compliance	Ethics & Responsible AI	Model Governance and Ownership	Risk Management	Communication & Transparency
Compliance Assessment	Ethical Guidelines	Model Strategy	Risk Assessment	Communication Strategy
Compliance Framework	Ethical Review and Impact Assessment	Model Documentation	Risk Mitigation	Reporting and Accountability
Ongoing Compliance Monitoring	Stakeholder Engagement	Model Ownership	Risk Monitoring	
		Model Lifecycle Management		
		Model Incident Management		

Compliance and regulatory

How can we help you ensure your compliance with AI regulations?

AI Compliance is an evergreen process and will need to be agile to adapt to changes in leading industry practices, regulations, and your organization. KPMG HU can help you both during the initial implementation and anytime later during the lifecycle.

Org Strategy & TOM

- Define roles and responsibilities
- Identify Regulatory compliance requirements
- Determine Organization model structure

Defining AI Principles

- Establish AI Oversight Committee
- Define and communicate AI principles around ethics, trust, accountability, compliance (align with a framework such as NIST

Policies & Standards

- Develop policies based on the AI principles
- Incorporate AI concepts in existing policies, as needed
- Identify regulations that might influence policies
- Develop a change management process to review and update policies

Training

- Conduct risk training for all employees to understand the risks with utilizing AI across different levels and lines of business

Risk Assessments

- Identify risks, categorize on rankings, and prioritize accordingly, define actions, project planning
- Ensure compliance with regulatory bodies
- Continually reassess identified risks

Model Lifecycle & Control

- Determine organization and deployment level controls and strategy
- Identify key risks across the model lifecycle and design controls to mitigate risks according to requirements (including regulations)

Metrics, Monitoring, Reporting

- Create KPIs/KRIs to baseline AI benefits/compliance/ROI
- Create controls based on the type of AI model deployment that was utilized
- Report both internally and to regulatory bodies

Maintenance & Evolution

- Regularly review and update the AI Governance program
- Leverage technology to maintain and evolve AI governance

Compliance and regulatory

Some specific examples in details

Governance and accountability: Establishing a robust governance framework to address AI risks.

Service: Creating policies and guidance documents

- Definition of internal and external values and norms in relation to bias
- Definition of standards for the entire AI lifecycle, including:
  - AI Principles
  - Clear governance structure
  - Risk assessment processes
  - Metrics, monitoring and regulatory reporting
  - Ensuring human oversights

Regulatory compliance: Staying updated with regulatory requirements and ensure compliance with AI-specific regulations like the EU AI Act or other standards such as standards defined by NIST, OECD, ENISA…

Service: Conducting a gap analysis
Identify gaps and enhancements with company policies, standards, and SOP documentation against EU AI Act regulation (in case of users and stakeholders from EU), international standards and good practices.
If required, defining action plans and project planning as well as implementation project management is also ensured.

Controls: Ensuring that the AI models operate ethically, securely and in alignment with the company's goals and regulatory requirements.

Service: Creating a control catalogue
Gather an inventory of controls in place and recommend potential improvements.

Roles and responsibilities: Ensuring accountability, enhancing collaboration, and mitigating risks associated with the deployment and management of AI models.

Service: Creating a RACI matrix
Present the AI model related roles, organizational relationships, competencies and skill requirements.

Compliance with AI/ML development best practices

- **An AI/ML project delivery has various obstacles from idea to delivery and each has to be tackled to have a successful product**
- **Neglecting these steps would result in a technical debt which accumulates over the product lifecycle and prevents its scale up**

Feature engineering and data selection

Selecting and transforming the relevant data is essential and it is the backbone of our operation.

Model selection, training and/or transfer learning

During development we need to consider several aspects when we choose from a wide variety of applications and train them on our usecase. The most important tradeoff is explainability vs performance.

Deployment and MLOps

The development does not end when our model is ready, it has to reside in an environment where it is safe to run and automatically retrain itself when it is necessary.

Documentation

The operational and functional guidelines should be adequate for 3rd party developers and audit teams to understand the functionality and push required changes effectively.

Bias and fairness testing

The goal is to ensure that these models treat different groups fairly and do not produce harmful, discriminatory, or unethical responses.*

01

Bias Detection

- Demographic Bias: Checking if the model disproportionately favors or discriminates against certain groups based on race, gender, age, nationality, etc
- Social and Cultural Bias: Identifying stereotypes or prejudiced viewpoints embedded in model outputs
- Toxicity and Harmful Content: Ensuring the model doesn't generate offensive, harmful, or misleading content.
- Prompt Sensitivity: Testing if the model reacts differently to similar prompts when slight demographic variations are introduced (e.g., "auditor" vs. "female auditor").

02

Fairness Evaluation Methods

- Counterfactual Testing: Providing inputs with minor demographic variations (e.g., swapping names or genders) to check if responses change unfairly.
- Bias Benchmarks: Using datasets like WinoBias, StereoSet, or the Bias Benchmark for QA (BBQ) to evaluate biases in NLP models
- Adversarial Testing: Crafting edge cases to trigger potential biased outputs.
- Statistical Fairness Metrics: Measuring fairness through metrics like equalized odds, demographic parity, and disparate impact.

03

Bias Mitigation Techniques

- Data Curation: Filtering, balancing, and augmenting training data to reduce biases.
- Model Fine-Tuning: Using curated dataset to update/guide the model toward more neutral and ethical responses.
- Post-Processing Filters: Adding rule-based or AI-driven moderation layers to detect and filter biased outputs.
- Human-in-the-Loop (HITL): Incorporating human reviewers to validate and adjust model behavior based on fairness guidelines.

**KPMG HU is currently finetuning its AI bias and fairness testing methodology*