Certified AI Security Engineer Training

Duration: **3 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Overview

- **Learn how to identify, test, and defend against real-world AI security threats.**
- **Designed for security engineers and AI developers responsible for building or protecting LLM-powered applications.**
- **Gain practical experience through hands-on labs covering prompt injection, jailbreaks, model theft, and other AI attack techniques.**

The Certified AI Security Engineer course provides a practical and comprehensive introduction to the security challenges surrounding modern artificial intelligence systems. Participants examine how both predictive and generative AI models can be targeted by attackers and how those attacks can affect applications, data, infrastructure, and organisational assets.

The programme explores common AI attack techniques including prompt injection, model jailbreaks, denial of service, model theft, and data poisoning. Rather than focusing only on the model itself, the course considers the complete AI attack surface, covering training datasets, APIs, integrations, applications, supply chains, and deployed infrastructure.

Participants also learn how to defend AI systems using structured prompt protections, security APIs, secure integration patterns, data integrity controls, and resilient infrastructure designs.

Through realistic scenarios and extensive practical exercises, learners develop the skills required to build AI capabilities that are responsible, reliable, and secure while protecting organisational information and maintaining trust in AI-assisted systems.

Prerequisites

Participants should ideally have:

- A basic understanding of artificial intelligence concepts such as neural networks and model lifecycle stages
- Fundamental knowledge of cybersecurity principles and common attack techniques
- Experience working with applications that use AI or LLM functionality
- Access to a development environment suitable for practising AI integration

Experience with AI-enabled applications and access to a development environment are recommended rather than mandatory.

Certified AI Security Engineer Training

Duration: **3 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Who Should Attend

This course is suitable for:

- Technology professionals responsible for deploying, integrating, or securing AI solutions
- Security practitioners who want a deeper understanding of AI-specific threats
- Developers building applications powered by large language models or generative AI
- Organisations seeking to improve resilience against AI-related security risks

What You Will Learn

By the end of the course, learners will be able to:

- Describe different types of AI systems and explain their security weaknesses
- Identify and mitigate attacks including prompt injection, model jailbreaks, visual prompt manipulation, and denial of service
- Use defensive techniques and security API tooling to strengthen AI systems
- Assess the security of training data, model integrity, and supply-chain dependencies
- Integrate large language models securely into applications
- Apply appropriate LLM trust boundaries
- Evaluate responsible AI, ethical considerations, reliability, and explainability
- Investigate model behaviour and identify signs of misuse
- Apply structured threat modelling to AI-enabled workflows
- Design safer human-AI interactions that reduce hallucinations, misuse, and sensitive data exposure

Outline

Introduction to AI Security

This section establishes the core concepts and boundaries of AI security.

Topics include:

- Defining AI and defining security
- Scope and boundaries of AI security
- Types of AI systems including neural networks, models, and integrated systems
- How organisations use AI systems
- What secure AI means
- Responsible, reliable, explainable, and aligned models
- Human-AI interactions
- Risks created by uncensored or malicious models
- Real-world misuse including deepfakes, voice cloning, and social engineering
- The spread of misinformation through AI-generated content
- Practical exploration of uncensored models and image watermarking

The AI Security Landscape

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

This section examines the attack surface of an AI system throughout its lifecycle.

Topics include:

- AI attack surfaces across the model lifecycle
- Components of AI pipelines
- Why supply-chain security matters
- Models accessed through APIs
- APIs accessed by AI models
- Traditional non-AI attack vectors that remain relevant
- OWASP ML Top 10
- OWASP LLM Top 10
- Applying these risks to modern AI environments
- Threat modelling for AI-integrated applications
- Common findings in AI-powered workflows

Exercise: Threat modelling an LLM-integrated application using a realistic data flow.

Prompt Injection

This section explores how attackers manipulate instructions provided to large language models.

Topics include:

- Prompt injection concepts
- Direct prompt injection
- Indirect prompt injection
- Prompt-based social engineering
- Phishing opportunities
- Using SudoLang to represent attack logic
- How LLM integration choices can create vulnerabilities

Exercises:

- Translating prompts into SudoLang
- Retrieving passwords in Level 1 and Level 2 scenarios

Model Jailbreaks

This section examines techniques designed to bypass model restrictions.

Topics include:

- How jailbreak techniques work
- Common jailbreak approaches
- DAN prompts
- AutoDAN
- Tree of Attacks with Pruning (TAP)

Exercises:

- Level 3
- Level 4
- Level 5

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

scenarios focused on retrieving restricted information.

Prompt Extraction

This section covers attempts to extract hidden instructions and sensitive information from AI systems.

Topics include:

- System prompt extraction
- Private data extraction
- Discovering model boundaries
- Techniques used in security challenges and real applications

Exercises:

- Level 6
- Level 7

scenarios involving prompt and boundary extraction.

Defending AI Systems

This section focuses on intermediate and advanced approaches for protecting AI applications.

Topics include:

- Intermediate and advanced defence strategies
- ReBuff
- Llama Guard
- Lakera
- Similar security API technologies
- Exploits observed in public security challenges
- Reverse psychology techniques
- Manipulation techniques
- Categorising injection attacks
- Implementing robust protections
- Structured defensive approaches such as the Bergeron method

Exercise: Attempting to bypass protections used in Level 8 and Level 9 scenarios.

Visual Prompt Injection

This section examines how visual inputs can be used to manipulate multimodal AI models.

Topics include:

- How visual prompt attacks work
- Simple visual manipulation examples
- Advanced adversarial attacks
- Attacks affecting self-driving systems
- Image classifier vulnerabilities
- Exercises using OpenAI vision capabilities
- Creating adversarial samples

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- Protections against visual attacks
- Dataset security considerations

Denial of Service

This section examines resource exhaustion and availability attacks against AI systems.

Topics include:

- Denial of service in LLM environments
- DoS risks affecting chatbots
- Prompt routing challenges
- Resource exhaustion
- System-level mitigation approaches

Exercise: Designing prompts intended to halt or degrade model behaviour.

Model Theft

This section examines techniques used to reproduce or extract model behaviour.

Topics include:

- The model extraction threat landscape
- Dataset exploration risks
- Query-based model stealing
- Cloning fine-tuned models
- Using API parameters to reproduce model behaviour
- Model confidentiality protections
- Rate limiting
- Advanced monitoring techniques

Exercise: Using API parameters to imitate model behaviour.

LLM Integration

This section focuses on integrating large language models securely into applications.

Topics include:

- Understanding the LLM trust boundary
- Classical integration risks in new AI workflows
- Treating LLM output as untrusted user input
- Exchange formats
- Secure function calling
- Risks associated with custom GPTs
- Identity flow
- Cross-application access
- SQL injection
- XSS payload generation
- Invalid parameter passing
- Privilege escalation
- Applying secure coding principles to AI systems

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- Bishop, Saltzer, and Schroeder principles
- Designing privilege boundaries around AI components

Exercise: Breaking out of an AI sandbox.

Training Data Manipulation

This section examines the importance of trusted and reliable training data.

Topics include:

- Dataset integrity
- Dataset reliability
- Training data poisoning
- Dataset cards
- Model cards
- Dataset analysis
- Reviewing dataset objectives

Exercise: Constructing and analysing malicious datasets.

Secure Supply Chain

This section focuses on protecting AI models throughout development and deployment.

Topics include:

- Proving model integrity
- Emerging cryptographic methods
- Hardware-assisted attestation
- Model verification
- Risks throughout the model-building lifecycle
- Risks throughout the deployment lifecycle

Human-AI Interaction

This section examines risks that arise when users rely on AI-generated outputs.

Topics include:

- Overreliance on LLM output
- Hallucinations
- Validating AI-generated information
- Sandboxing
- Safe API patterns

Exercise: Verifying LLM output in realistic scenarios.

Secure AI Infrastructure

This section examines the infrastructure requirements needed to operate AI systems securely.

Topics include:

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

- Monitoring
- Observability
- Traceability
- Confidentiality
- Integrity
- Availability
- Privacy
- Real-world examples such as the Samsung data leak
- LangSmith
- Safe AI evaluation
- BlindLlama
- Emerging AI evaluation tools

Exercise: Using LangSmith to explore safer AI evaluation practices.

Exams and Assessments

The independent Certified AI Security Engineer examination is taken after the course.

Candidates access the exam using an exam voucher code through the APMG proctor platform.

Exam details:

- **Duration:** 60 minutes
- **Questions:** 60
- **Question Type:** Multiple choice
- Each question contains 4 possible answers, with only 1 correct answer
- **Pass Mark:** 50%