Google Cloud

# Model Armor: Securing AI Deployments

This course explains how to use Model Armor to protect AI applications, specifically large language models (LLMs).

The curriculum covers Model Armor's architecture and its role in mitigating threats like malicious URLs, prompt injection, jailbreaking, sensitive data leaks, and improper output handling.

Practical skills include defining floor settings, configuring templates, and enabling various detection types. You'll also explore sample audit logs to find details about flagged violations.

DURATION
2 hours 30 minutes

LEVEL
Introductory

FORMAT
Instructor-led

## What you'll learn

- Explain the purpose of Model Armor in a company's security portfolio.
- Define the protections that Model Armor applies to all interactions with the LLM.
- Set up the Model Armor API and find flagged violations.
- Identify how Model Armor manages prompts and responses.

| | |
|---|---|
| Overview | 6 modules · 4 videos · 1 lab |
| Who this course is for | Security engineers, AI/ML developers, cloud architects |
| Products | Security |
| Prerequisites | • Working knowledge of APIs<br>• Working knowledge of Google Cloud CLI<br>• Working knowledge of cloud security foundational principles<br>• Familiarity with the Google Cloud console |

## Module 01 Course overview

**Topics**

What's in it for me?

**Objectives**

Recall the course learning objectives.

## Module 02 Model Armor overview

**Topics**

- About Model Armor
- LLM security risks

**Objectives**

- Explain the purpose of Model Armor in a company's security portfolio.
- Identify the subset of top 10 OWASP LLM vulnerabilities that Model Armor addresses.
- Identify Model Armor key concepts and architecture.
- Map Model Armor features to the security risks they mitigate.

**Activities**

- Knowledge check
- Quiz

## Module 03 Customize Model Armor

**Topics**

- About customization
- Floor settings
- Guard rails and confidence levels
- Templates

**Objectives**

- Define the protections that Model Armor applies to all interactions with the LLM.
- Describe floor settings and explain how they work.
- Explain the purpose of a template and how it works with the API.
- Configure the four types of detections in the template.

**Activities**

- Knowledge check
- Quiz

## Module 04 Use Model Armor

**Topics**

- About setup
- API setup
- Flagged violations

**Objectives**

- Set up the Model Armor API and find flagged violations.
- Explain the prerequisites that are required to work with the API.
- Describe how to enable the API.
- Set up logging in the template, explore types of audit logs, and find them in SCC.
- Explain how to find floor setting violations in SCC and resolve them.

**Activities**

Quiz

## Module 05 Put it all together

**Topics**

- Prompts and responses
- Application code

**Objectives**

Identify how Model Armor intercepts and manages prompts and responses.

ELO:

- Explain how Model Armor reviews prompts and reports findings based on content safety flags.
- Explain how Model Armor reviews LLM responses and updates them according to template settings.
- Execute various commands for sanitizing user prompts against different security features.

**Activities**

Quiz

## Module 06 Course conclusion

**Topics**

What did I learn?

**Objectives**

Summarize the course learning objectives.