Assessment Instruction

SEC-301: Security Challenges in Modern AI Systems

Spring 2026

General Information

Competency Code: SEC-301

Competency Title: Security Challenges in Modern AI Systems

Semester: Spring 2026

Instructor Information: Charnon Pattiyanon, Ph.D. (charnon@cmkl.ac.th)

Assessment Overview

As AI-enabled systems become a hype and easier to access for end users, they are also being targeted by security threats that seek to exploit them for adversaries' benefit. This competency offers you the opportunity to explore security challenges and AI-specific attacks, enabling you to gain awareness of such threats and learn how to protect your AI system.

To ensure that, as a student, you have built proper awareness and acquired the skills needed to develop AI systems that are safe from security threats, this assessment will ask you to apply the knowledge, techniques, and tools discussed during the lectures to a real AI system. The assessment also requires you to get hands-on experience by implementing a real attack against your developing system and then developing protections against that attack.

Assessing Skills

- **[SEC-301:00010] Analyze AI Security Risks** – Successful students must be able to assess potential security risks in modern AI systems.
- **[SEC-301:00020] Apply Analysis Techniques to AI Security Threats –** Successful students must be able to use analytical methods to identify security threats in modern AI systems.
- **[SEC-301:00030] Analyze AI-Specific Attack Scenarios –** Successful students must be able to thoroughly analyze attack scenarios that can be exploited in modern AI systems.
- **[SEC-301:00040] Understand AI Safety in Academia –** Successful students must be able to demonstrate an understanding of current trends, methodologies, and the research landscape in AI safety.

Pre-Cautions

- **Express your answers from your own ideas and perspective.** Plagiarism is unacceptable. You must cite referenced sources properly to acknowledge their originality and must not copy partial or entire ideas from your peers. If content or ideas are found to be remarkably similar between two or more submissions, or if original material is copied from other works without proper citation, all students will receive a score deduction as a consequence of disciplinary action.
- **Demonstrate deep understanding through critical analysis and original insight.** Overreliance on AI-generated content without substantial original thought will negatively impact the assessment score.

- **Justifications** should explain a decision or finding in a "why" style, providing adequate technical and valid rationale. For example: *"I believe that this security threat is potentially dangerous to the system because a common model-training framework is forked from a public repository where anyone can contribute to it."* There will be no one-size-fits-all solution or criticism for writing a justification; your skill will be evaluated on the clarity of your justifications.
- **Inquiries:** Students are encouraged to ask instructors any questions about the assessment or competency content via email or other agreed channels. However, students are not allowed to submit an assessment report and ask for feedback; such a submission will be treated as a report submission.
- **Optional questions** may be provided in this assessment with a clear indicator. Students may omit them from the report without affecting the final grade. However, optional questions may be considered in cases where a student receives a borderline score between two mastery levels or fails the competency. The optional question can contribute to the final score but will not exceed 10 % of the overall score, at the instructor's discretion.

Assessment Instruction

The total score for this assessment is 400 points, with each skill contributing 100 points. Please carefully follow the instruction below:

For SEC-301:00010, SEC-301:00020, and SEC-301:00030:

1. Each student must pair up with another student (teams of **two**); if the enrollment is odd, one team may have three members.
2. Each team must select an AI system to develop. The system may be implemented independently using publicly available datasets from sites such as HuggingFace.co or Kaggle.com to develop the AI system
3. [5 Points] Each team must submit the first deliverable on Sunday, February 8, 2026, before 11:59 PM, which includes:
    a. Your team-member list (name, nickname, and email address).
    b. Your chosen AI system title (one concise sentence).
    c. A description of your AI system, detailing its purpose, inputs, and expected output.
    d. A list of datasets you will use, with links.
4. Each team must develop the selected AI system according to the functionalities described in 3. The deliverable may be a simple CLI program that accepts inputs and produces outputs, or a Jupyter Notebook that documents the model-training process.
5. Each team must conduct a security-threat analysis using techniques discussed in Lecture 2, identifying potential security threats based on the current architecture.
6. Each team must select one AI-specific risk or attack from the identified threats in 5, then exercise the attack procedure on the developed AI system and observe the resulting behaviour.
7. Each team must research for proper protection and implement security measures and controls to protect the AI system against the chosen attack. The system's output should be re-tested to confirm that the attack can no longer be exploited.
8. Each team must write a full report on the work performed during the assessment activity. The report template, provided separately to this instruction, outlines the minimum requirements that every team should meet, but teams are free to add additional details beyond this baseline. Any extra information that enhances clarity may positively influence the final score.
9. [395 points] Each team must submit the full report described in 8 on Friday, May 1, 2026, before 11:59 PM on Canvas. Late submissions are not accepted, as the deadline is set by the university.