

PRACTICE QUESTIONS

AI CERTs

AI+ Security

SHARED BY [u/quickbadger2001](#) on aiexamcerts

EDITION	Free Set 1
QUESTIONS	10
SHARED	October 15, 2026
EXAM	AI CERTs AI+ Security

Shared by [u/quickbadger2001](#) on aiexamcerts

10 practice questions for the AI CERTs AI+ Security exam, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/ai-certs/ai-security. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. Which goal describes an evasion attack against a deployed classifier?

- A Causing one crafted input to be misclassified at inference time, without changing the model
- B Planting a hidden trigger in the training records so the model misbehaves after deployment
- C Copying the model's behavior by querying the public API until a clone can be trained
- D Reading the training records directly by exploiting weak access controls on the storage bucket

Answer: A

An evasion attack perturbs an input at inference time so the deployed model returns the wrong label, and the model itself is left untouched. Planting a trigger in training data is a poisoning attack, and cloning through queries is model extraction. The storage-bucket option is an ordinary data breach rather than an adversarial machine learning attack.

2. Loading a pickle file from an untrusted source can allow arbitrary code execution, a risk known as insecure _____.

Answer: deserialization (also accepted: deserialisation)

Insecure deserialization happens when loading an object runs code that the attacker controls. Checksums and safer formats reduce this risk.

3. Why is a defense that only blocks one known attack method considered weak?

- A Known attacks are never a real risk, so blocking them has no value at all for the system
- B Blocking methods slow the model down, which users notice more than any attack that the security team has reported to date
- C An adaptive attacker can change the method, so the defense must be tested against attacks designed to bypass it
- D Any defense that uses a published method is automatically prohibited under the EU AI Act

Answer: C

Attackers adapt to the defenses they meet, so a filter tuned to one published attack can be bypassed with a small change. Robustness claims therefore need adaptive evaluation. The other options dismiss the risk or invent legal and performance reasons that do not apply.

4. A model with high accuracy on clean data can still be fooled by small input perturbations.
- A True
 - B False

Answer: A (True)

High clean-data accuracy measures typical performance and says little about worst-case inputs, which attackers choose deliberately. Small, carefully crafted perturbations can flip predictions in many models. The distractor is equating good average accuracy with robustness.

5. Which two practices help detect poisoned records in a fine-tuning dataset? (Choose two.)
- A Shuffling the dataset so that poisoned rows are spread evenly through the file
 - B Provenance tracking for each source, with records from unknown sources quarantined
 - C Deduplicating only exact copies of each record in the dataset
 - D Adding a larger validation set collected by the same contributors as before
 - E Statistical outlier and label-consistency checks run before every training job

Answer: B, E

Provenance tracking shows where each record came from, and outlier and label checks catch records that do not fit the rest of the data. Shuffling, exact deduplication and a larger set from the same contributors do not reveal poisoning. Those steps leave the attacker's records in place.

6. Which OWASP LLM risk covers attacks where crafted user or document text overrides the model's instructions?
- A LLM03 Supply Chain, which covers vulnerable third-party models, datasets and plugins used by the application
 - B LLM01 Prompt Injection, which includes direct and indirect manipulation of model instructions
 - C LLM09 Misinformation, which covers outputs that are plausible but false and unsupported by evidence
 - D LLM10 Unbounded Consumption, which covers runaway usage of compute, memory and API calls

Answer: B

LLM01 is Prompt Injection, and it includes direct input and instructions hidden in retrieved content. Supply chain, misinformation and unbounded consumption are separate entries in the same list. Matching the risk to its entry is the skill the question tests.

7. What is the main risk when a retrieval-augmented assistant treats retrieved web content as trusted instructions?
- A The assistant becomes slower when it reads long web pages, which increases the cost per answer
 - B The retrieval index grows too large, which means the team must archive older documents every year
 - C An attacker can plant instructions in a page the assistant reads and steer its actions
 - D The embedding model may produce vectors with fewer dimensions than the database expects to store

Answer: C

Retrieved pages are written by third parties, so treating them as instructions lets an attacker steer the assistant through content it reads. The result is indirect prompt injection with real actions attached. Latency, index size and vector dimensions are operational issues, not security risks of this kind.

8. The NIST framework with the functions GOVERN, MAP, MEASURE and MANAGE is the AI Risk _____ Framework.

Answer: Management

The NIST AI RMF is the AI Risk Management Framework, and its four core functions structure how organizations manage AI risk. Governance is the function that sets the policies and accountability for the rest.

9. What does mutual TLS add to an inference API?
- A Faster response times, because the handshake skips the certificate checks on both sides of the call
 - B A guarantee that the model's outputs are free of harmful content for every single request
 - C Automatic removal of personal data from every request body that the client application sends
 - D Both the client and the server present certificates, so each side verifies the other's identity

Answer: D

Mutual TLS authenticates both ends of the connection, so only trusted clients reach the service and the client confirms the server it is talking to. It does not speed up responses, filter content or remove personal data. Those are separate controls with separate purposes.

10. Which NIST AI RMF function sets the policies, roles and accountability for managing AI risk?
- A MAP, which identifies the context, intended uses and potential impacts of an AI system at the start
 - B GOVERN, which sets the policies, roles and culture for managing AI risk across the organization
 - C MEASURE, which assesses identified risks with metrics, tests and evaluations of the system
 - D MANAGE, which allocates resources to treat, monitor and respond to the risks that were measured

Answer: B

GOVERN is the function that establishes accountability, policy and culture, and it applies across the other three functions. MAP, MEASURE and MANAGE each work on a specific part of the risk cycle. Confusing MANAGE with GOVERN is the most tempting slip because both sound administrative.