

PRACTICE QUESTIONS

AI CERTs

# AI+ Security

SHARED BY [u/rusticracoon1982](#) on aiexamcerts

|           |                       |
|-----------|-----------------------|
| EDITION   | Free Set 2            |
| QUESTIONS | 10                    |
| SHARED    | October 15, 2026      |
| EXAM      | AI CERTs AI+ Security |

---

Shared by [u/rusticracoon1982](#) 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](https://www.aiexamcerts.com/ai-certs/ai-security). Spotted a mistake?  
 Email [support@aiexamcerts.com](mailto:support@aiexamcerts.com) with the exam name and the question number.

---

1. Which training-time defense is most directly aimed at improving resistance to adversarial perturbations?
  - A Adversarial training, which includes perturbed examples in the training data
  - B Data augmentation with random crops that only changes the framing of each image
  - C Early stopping based on the validation loss of a clean holdout set
  - D Weight decay regularization applied to every layer in the network architecture

**Answer: A**

Adversarial training exposes the model to perturbed inputs during learning so it learns to resist them, which directly targets the threat. Random crops and early stopping address generalization, not adversarial inputs. Weight decay is a regularization method that is not designed for perturbation resistance.

---

2. A system prompt guarantees that the model will never follow instructions that appear inside user-supplied text.
  - A True
  - B False

**Answer: B (False)**

The system prompt is guidance the model weighs, not an enforced boundary, so a crafted input can still override it. Robust designs enforce limits outside the model, such as permissions and approvals. The distractor treats the system prompt as a security control.

---

3. What is a backdoor attack on a model?
  - A Gaining administrator access to the server room where the GPUs are racked up in the data center
  - B Training the model so that a specific trigger in an input produces an attacker-chosen output, while normal inputs behave correctly
  - C Sending a flood of requests that exhausts the rate limit of the inference service each day, which drains the quota for paying customers
  - D Copying the model file to an external drive without the owner's permission for later use

**Answer: B**

A backdoor pairs a hidden trigger with an attacker-chosen output, so ordinary evaluation can look clean while the trigger still works. Physical access, request floods and file copying are real threats but not backdoors. Testing with suspected triggers is how backdoors are found.

---

4. A tensor file format designed to store model weights without any executable code is called \_\_\_\_\_.

**Answer: safetensors** (also accepted: safe tensors)

Safetensors stores tensor data without the code execution path that pickle files have, which makes loading safer. Teams commonly prefer it when distributing weights from public hubs.

5. Which two measures reduce the risk of model extraction through a public inference API? (Choose two.)

- A Rate limiting and per-client quotas on query volume
- B Publishing the model architecture so that customers can verify what they are calling
- C Returning only the final label instead of full probability scores
- D Returning full logits and confidence values in every response to aid debugging
- E Allowing unauthenticated users unlimited requests so that latency stays low for everyone

**Answer: A, C**

Extraction needs many queries and rich output signal, so quotas slow it and label-only responses withhold the signal. Publishing the architecture and returning logits both hand the attacker information. Unlimited anonymous access removes the very friction that protects the model.

6. Which control most limits the damage a prompt-injected agent can do?

- A Least-privilege tool access, with human approval for actions that move money or change data
- B A longer system prompt that repeats the rules several times in capital letters for emphasis
- C A higher temperature setting so the agent varies its responses and avoids predictable patterns
- D Storing the agent's API keys inside the prompt so that it always has the tools it needs at hand

**Answer: A**

Limiting what the agent can do and requiring approval for consequential actions caps the harm of any successful injection. A longer prompt or higher temperature does not enforce anything, and embedding keys in the prompt exposes them. Controls that sit outside the model are what hold under attack.

7. Which activity belongs in an AI security risk assessment before deployment?
- A Choosing the color palette for the public chat widget and the logo placement in the page header
  - B Counting how many employees attended the launch event and recording their attendance
  - C Modeling threats to the system, identifying likely attackers and rating the impact of each misuse
  - D Writing the press release for the launch so that marketing can review its wording early

**Answer: C**

A security risk assessment models threats, adversaries and impacts so that controls can be chosen against them. Visual design, attendance and press releases have no bearing on the system's exposure. The assessment is the input that justifies each control.

---

8. The core functions of the NIST AI Risk Management Framework include GOVERN, MAP, MEASURE and MANAGE.
- A True
  - B False

**Answer: A (True)**

The AI RMF organizes its core work into GOVERN, MAP, MEASURE and MANAGE, with GOVERN applying across the other three. Each function holds categories and subcategories that teams use to plan and report on AI risk.

---

9. Which signal best suggests that an attacker is probing a classifier for its decision boundary?
- A A single long request that arrives once per day from a well-known internal service account
  - B A rise in average response length during the company's monthly newsletter send window
  - C A stream of queries from one client with small, systematic changes to a single input
  - D A higher number of users who ask about the pricing page on a weekday afternoon in the spring

**Answer: C**

Small systematic variations of one input from one client are how boundary probing looks in logs, and the pattern is detectable. Routine internal calls, seasonal traffic and popular pages are normal variation. Good detection focuses on the shape of queries rather than their volume alone.

---

10. What should an incident responder preserve first when a model is suspected of compromise?
- A** Logs, prompts, outputs and model artifacts with hashes, before any rollback or retraining changes them
  - B** The old chat transcripts, which can be deleted after the first day to comply with the storage limits set by finance
  - C** The model itself, which should be retrained on new data to erase the influence of the attack
  - D** The public status page, which should be updated with a full technical breakdown of the attack vector

**Answer: A**

Evidence must be captured before rollback or retraining overwrites the state that shows what happened. Hashes prove the artifacts were not altered after capture. Deleting transcripts, retraining in place and publishing details all destroy evidence or disclose too early.