PRACTICE QUESTIONS

EC-Council

Certified Offensive AI Security Professional (COASP)

PUBLISHED BY aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
ISSUED	November 8, 2026
EXAM	EC-Council Certified Offensive AI Security Professional (COASP)

Shared by **aiexamcerts** on aiexamcerts

10 practice questions for the EC-Council Certified Offensive AI Security Professional (COASP) exam, each followed by its answer and a short explanation. More free sets and the full practice exam: **aiexamcerts.com/ec-council/certified-offensive-ai-security-professional-coasp**. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

---

1. A tester should keep exploiting a confirmed prompt injection on a production system after the finding is verified, to collect more examples for the report.

A True

B False

Answer: B (False)

Once a finding is verified, the tester should stop, record the minimum reproduction needed, and report it, since further exploitation on production adds risk and can expose real data. Collecting more examples sounds thorough but goes beyond what the rules of engagement usually allow. Any extra proof should come from a staging replica that the owner has agreed to.

---

2. During reconnaissance, a tester finds that an inference API adds the exact model build and framework version to every response header. Why does this matter to the owner?

A It lets an attacker match the service to known weaknesses in that exact build, so it should be removed

B It proves the model carries a valid license, so the header must stay in every response

C It causes legitimate users to be rate limited, so it should be removed for better uptime

D It tells the billing system which customer to charge, so removing it would disrupt invoicing

Answer: A

Exact version strings help an attacker choose known issues for that build, so reducing this detail is a standard hardening step. The option claiming the header proves licensing is the most tempting misreading, but a response header plays no part in license validation. Rate limits and billing work from authenticated identity, not from a version string, so the other two claims are also false.

---

3. A team wants to test whether an LLM application's JSON replies always parse, across thousands of generated prompts. Which kind of test fits this goal best?

A Property-based testing that asserts every reply validates against the expected schema

B A single manual review of five replies chosen by the developer, checked against the spec

C A benchmark measuring tokens generated per second

D A physical penetration test of the data center badge readers, door locks, and server cages

Answer: A

Property-based testing checks an invariant, here schema validity, across many generated inputs, which catches rare failures that a handful of hand-picked replies miss. A manual review of five replies is the most tempting shortcut, but it samples too few cases to support a claim about every reply. Throughput benchmarks and badge-reader tests do not check whether replies match the schema.

---

4. Crafted framing, such as asking the model to play a role that has no rules, that overrides a model's safety behavior is called a _____.

Answer: jailbreak

A jailbreak is an input that gets a model to set aside its safety behavior, and role-play framing is one common pattern. It differs from prompt injection, which hijacks the application's instructions or actions through untrusted content. Defenses include layered safety evaluation and output checks rather than trusting the model to refuse on its own.

---

5. A team wants to publish a medical image classifier, but it worries that detailed confidence scores could expose patients in the training set. Which defense addresses this most directly?

A Add more hidden layers so the network can represent finer detail in its learned features and outputs

B Enlarge the validation set so accuracy estimates are more stable

C Train with differential privacy and return only the top label, not full confidence values

D Lower the learning rate so training converges to a smoother solution

Answer: C

Differential privacy limits what training records can reveal, and withholding full confidence scores shrinks the signal that privacy attacks rely on. Adding layers is the most tempting response because it sounds like a capacity fix, but it does nothing for leakage. A larger validation set and a lower learning rate change accuracy estimates and convergence, not what the published outputs disclose.

---

6. Which description best matches a clean-label poisoning attack?

- **A** The attacker changes the labels of most training records to random classes so accuracy collapses
- **B** The attacker sends corrupted files to the inference endpoint during peak traffic hours
- **C** The attacker deletes the test set from the repository so that evaluation is skipped
- **D** The poisoned examples keep correct labels, but their content is crafted to make the model misbehave

Answer: D

Clean-label poisoning keeps the labels correct, so the records pass a label audit, while their content shapes what the model learns. Random label corruption is the most tempting confusion because it is the more familiar attack, but it is not clean-label. Corrupting inference traffic and deleting a test set are different attacks, one aimed at serving and one at evaluation.

---

7. It is safe to assume that a tool call an LLM requests reflects the user's intent, so the call can run without validation.

- **A** True
- **B** False

Answer: B (False)

The model can request a call because of injected text, a misread instruction, or a hallucinated argument, so the request is not proof of user intent. Tool calls need schema validation, authorization checks against the current user, and approval for sensitive actions. Assuming intent is the core weakness that injection exploits.

---

8. Which practice most directly supports the integrity of an ML build pipeline's artifacts?

- **A** Generate provenance attestations for build outputs and verify them before deployment
- **B** Turn off build logs to save disk space on the runner
- **C** Accept whichever dependency version the public index lists as newest at the moment of each build
- **D** Let every developer push models straight to production without any review step at all

Answer: A

Provenance attestations record how an artifact was built, and verifying them before deployment confirms that only artifacts from the expected pipeline run are used. Accepting the newest dependency version is the most tempting shortcut, because it sounds current, but it enables dependency confusion and unvetted changes. Disabling logs and skipping review remove the evidence and checks that integrity depends on.

---

9. Which action is an example of hardening rather than detection?

- **A** Removing unused tools, endpoints, and debug routes from the model-serving stack to shrink what is exposed
- **B** Writing an alert that fires when a tool call fails several times in a row, for on-call staff
- **C** Reviewing a weekly sample of conversations for signs of policy violations
- **D** Tracking the daily refusal rate of the assistant on a monitoring dashboard

Answer: A

Hardening reduces the attack surface by removing exposure, as deleting unused endpoints does. Alerts, conversation reviews, and refusal dashboards are detection controls that observe activity rather than reduce what can be attacked. The alert option is the most tempting, since monitoring feels protective, but it does not remove anything.

---

10. Which two actions support a defensible AI forensic investigation? (Choose two.)

- **A** Keep a chain-of-custody record for every exported log and model artifact
- **B** Edit the original logs to remove personal data before any investigator sees them
- **C** Clear the inference server's caches so the bug cannot reappear before the review
- **D** Work from forensic copies so the only production copy is never altered during analysis
- **E** Share the raw incident export with the press before the legal and security teams review it

Answer: A, D

Chain-of-custody records show who handled each artifact and when, and working from forensic copies protects the original evidence. Editing originals is the most damaging wrong option, because it breaks integrity and undermines every finding built on them. Clearing caches and sending raw exports to the press destroy the live state or expose sensitive material.