EXAM 3 PRACTICE - ANSWER KEY

# Practice Answer Key

**CIS 350: AI for Cybersecurity - Smart Defense for the Digital Business**

Colorado State University - Computer Information Systems

**About this document**

Answers and brief explanations for the Exam 3 practice questions. Use this to self-check AFTER attempting the practice exam. Understanding WHY an answer is right matters more than the letter.

## Part A - Multiple Choice: answers

1:C 2:A 3:D 4:B 5:A
6:D 7:B 8:C 9:D 10:A
11:C 12:B 13:A 14:C 15:B
16:D 17:C 18:A 19:D 20:B

## Part A - explanations

**1. C - A tiny, crafted change to an input that flips the prediction**
It is a small, crafted perturbation - often invisible - that crosses a decision boundary.

**2. A - Poisoning that plants a hidden trigger for secret misbehavior**
A backdoor is stealthy poisoning: normal behavior until the secret trigger appears.

**3. D - Whether a specific record was in the training set**
Membership inference reveals if a particular person's data was used to train the model.

**4. B - Defense in depth**
No single defense is perfect - layer defenses, detect, and plan to respond.

**5. A - Heavily querying a model to clone its behavior**
An attacker queries a lot and trains a copy that mimics the target model.

**6. D - Cannot reliably separate trusted instructions from untrusted data**
Instructions and data share one text channel, so attacker text can win.

**7. B - Get the model past its safety rules**
Jailbreaks coax the model into ignoring its safety guidelines.

**8. C - Reveals its hidden system prompt**
Leakage exposes the developer's hidden instructions, aiding further attacks.

**9. D - Least privilege**
Least privilege limits the blast radius if the LLM is manipulated.

**10. A - Checks the model's response before it is shown or acted on**
It is an independent check on the output - e.g., block replies that leak secrets.

**11. C - Can hide per-group harm**
A high average can mask that one group is treated much worse.

**12. B - True-positive rate across groups**
Equal opportunity requires each group's real positives be caught at the same rate.

**13. A - Risk level (higher risk -> stricter duties)**
It is risk-tiered: unacceptable, high-risk, and limited/minimal-risk categories.

**14. C - Biased outputs become tomorrow's training data**
The model's biased decisions generate biased data that reinforces the bias.

**15. B - Can conflict, so fairness needs a deliberate choice**
When base rates differ, the common definitions cannot all hold at once, so fairness is a deliberate choice.

**16. D - More privacy but usually lower accuracy**
Smaller epsilon = more noise = more privacy and less accuracy.

**17. C - Only the model updates**
Raw data stays on the device; only learned updates are shared and combined.

**18. A - Compute on encrypted data without decrypting it**

You can process ciphertext; only the owner decrypts the result.

**19. D - Combining other fields can re-identify people**

Quasi-identifiers (zip, birth date, device) can re-identify individuals.

**20. B - Helps detection but raises privacy risk**

More signals aid detection but create a sensitive-data honeypot - a trade-off.

## Part B - True / False: answers

**21. TRUE**

Poisoning happens at training time by tampering with the data.

**22. FALSE**

It is small and crafted, often imperceptible.

**23. TRUE**

It tops the OWASP LLM list.

**24. TRUE**

Averages hide per-group harm - disaggregate to see it.

**25. FALSE**

A clever attacker can talk past it; you need independent guardrails.

**26. FALSE**

Proxy features (e.g., zip code) reintroduce it.

**27. TRUE**

Smaller epsilon = more noise = more privacy (less accuracy).

**28. FALSE**

It keeps raw data on devices and shares only model updates.

**29. FALSE**

No perfect defense - defend in depth.

**30. TRUE**

The same tool protects users and hides malicious traffic.

## Part C - Short Answer: model answers

**31. What is data poisoning, and how does it differ (in timing) from an adversarial/evasion attack?**

Poisoning corrupts the TRAINING data (training time) so the model learns the wrong thing; an adversarial/evasion attack crafts an input to fool the ALREADY-DEPLOYED model (test time). Same goal, different when and different access.

**32. What is prompt injection, and name two defenses for an LLM application.**

Attacker input that overrides the app's instructions. Any two defenses: input/output guardrails, least privilege, separating data from instructions, monitoring, and human approval for high-impact actions.

**33. Why can a model be accurate overall but unfair, and name one fairness definition.**

Overall accuracy averages the groups and hides per-group harm, so you must disaggregate. One definition: demographic parity (equal selection rate), equal opportunity (equal TPR), or equalized odds (equal TPR and FPR).

**34. What does epsilon control in differential privacy, and what is the trade-off?**

Epsilon is the privacy budget: it sets the noise scale (sensitivity/epsilon). Smaller epsilon = more noise = more privacy but lower accuracy; larger epsilon = the reverse. That is the privacy/accuracy trade-off.

### 35. Compare federated learning and encrypted computation in one sentence each.

Federated learning keeps raw data on each device and shares only model updates. Encrypted computation (e.g., homomorphic encryption) computes on data while it stays encrypted, so the processor never sees it in the clear.

## Part D - Applied Scenarios: model answers

### 36. A hospital readmission model (10 pts)

(a) Compare per-group true-positive rate (equal opportunity) or false-negative rate; the harmed group is caught less often. Fix by rebalancing/relabeling data, group-aware thresholds, then re-audit.

(b) Model inversion or membership inference; defenses: differential privacy, limit exposed detail, rate-limit queries.

(c) Federated learning (train locally, share only updates), optionally combined with DP.

(d) A model card with per-group fairness, a bias audit, human oversight/appeals, or board sign-off.

### 37. A company AI code assistant (10 pts)

(a) Prompt injection; indirect injection could hide instructions inside a document the assistant reads.

(b) The model can't separate trusted instructions from untrusted content, so it can be talked past the prompt - guardrails are still needed.

(c) Guardrail: validate input and output (block secret leakage/dangerous commands). Least privilege: no destructive commands; read-only access; approval for anything high-impact.

(d) Log prompts/outputs/commands and require human approval for high-impact actions.

### 38. A malware detector under attack (10 pts)

(a) Evasion / adversarial examples; defenses: adversarial training, behavior/anomaly detection, input checks, low-confidence detection.

(b) Data poisoning; defenses: vet data and track provenance, anomaly-check the training set, test for backdoors.

(c) Telemetry is sensitive personal data (privacy/breach risk); middle ground: data minimization, anonymize/aggregate, short retention, access controls, or on-device/federated processing.

(d) Attacks and defenses co-evolve and defenses trade off cost/accuracy; defend in depth - raise attacker cost, detect, keep a human in the loop, and plan to recover.