EXAM 4 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 4 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 - Model inversion (reconstructing training data)
Model inversion, membership inference, and model theft are privacy attacks; the others are normal ML steps.

2. A - Match learned patterns and are brittle near boundaries
Models match statistical patterns, so a tiny nudge across a boundary flips them.

3. D - Vetting the training data and tracking provenance
Poisoning is a data problem - vet and track where training data comes from.

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

5. A - The hidden system prompt
Leakage reveals the developer's hidden instructions, enabling further attacks.

6. D - No wall between trusted instructions and untrusted data
Instructions and data share one text channel, so attacker text can win.

7. B - True-positive AND false-positive rates across groups
Equalized odds demands equal TPR and equal FPR across groups.

8. C - Purpose, performance, fairness, and limits
A model card is a short 'nutrition label' for a model - a transparency tool.

9. D - Biased outputs become tomorrow's training data
The model's biased decisions generate data that reinforces the bias.

10. A - Compute on encrypted data without decrypting it
You can process ciphertext; only the owner decrypts the result.

11. C - Quasi-identifiers can re-identify people
Zip + birth date + device can re-identify individuals, so name-removal is weak.

12. B - Less privacy (and usually more accuracy)
Larger epsilon = less noise = less privacy but more accuracy.

13. A - How the system overall works and its limits
Transparency is system-level openness; explainability is about one decision.

14. C - Oversight and the ability to override
Human oversight catches errors and preserves accountability for high-impact calls.

15. B - Prepare
Prepare (plans, roles, tools) comes before detect/analyze and the rest.

16. D - Accountability
Accountability means humans - not 'the algorithm' - are responsible.

17. C - A strong pretrained base (transfer learning)
A good pretrained representation lets a model learn from a handful of examples.

18. A - Differential privacy

DP on the shared updates strengthens federated learning's privacy.

19. D - Signature-only detection

It mutates to slip past signatures - use behavior/anomaly detection.

20. B - Raise attacker cost, detect fast, and adapt

There is no permanent win - defense in depth: raise cost, detect, adapt.

Part B - True / False: answers

21. TRUE

Poisoning tampers with the training data (training time).

22. TRUE

Averages mask harm to a group - disaggregate to see it.

23. FALSE

They are small, crafted, and often invisible.

24. TRUE

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

25. FALSE

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

26. FALSE

It keeps raw data local and shares only model updates.

27. TRUE

Data drift silently degrades accuracy after launch.

28. FALSE

No perfect defense - defend in depth.

29. TRUE

A pretrained base means your small dataset goes much further.

30. FALSE

Explainability is the reason for a specific decision, not speed.

Part C - Short Answer: model answers

31. What is an adversarial example and why does it work? When does data poisoning happen instead (timing)?

An adversarial example is a small, crafted input change that flips the prediction (often invisibly), because models match patterns, not meaning. Data poisoning happens at TRAINING time (corrupting the data); adversarial/evasion is at TEST time.

32. What is prompt injection, and name two layered defenses for an LLM app.

Attacker input that overrides the app's instructions. Two defenses (any): 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? Name one fairness definition.

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

34. What does epsilon control in differential privacy (which way is more private)? Name one other privacy technique.

Epsilon is the privacy budget: smaller epsilon = more noise = more privacy (less accuracy). Another technique: federated learning (keep data local, share updates) or encrypted computation (compute on ciphertext).

35. Name two trustworthy-AI pillars (explain one), then one advanced technique and the problem it addresses.

Two of: transparency, explainability, robustness, accountability (e.g., robustness = holds up under attack/drift, monitored over time). Advanced technique: transfer/few-shot/federated learning - all address label scarcity (lots of data, few labels).

Part D - Applied Scenarios: model answers

36. A loan-approval model (10 pts)

(a) Per-group false-positive/selection rate or equal-opportunity gap; reduce via relabeling, dropping proxy features, or group-aware thresholds, then re-audit.

(b) Explainability - give the applicant the top reasons for the decision in plain language.

(c) Model inversion / membership inference or a breach; defense: differential privacy, minimize/protect data, access controls.

(d) A model card with per-group fairness, a bias audit, human oversight, or an appeal path.

37. An LLM email assistant under attack (10 pts)

(a) Prompt injection; indirect injection can hide instructions inside an email/document it reads.

(b) Guardrail: validate input and output (block data leakage). Least privilege: draft-only, no send without human approval, scoped read access.

(c) Robustness - a new failure after an update suggests drift/regression; it must be monitored and tested.

(d) Prepare (plan/roles ready) then Detect & Analyze (confirm, scope, triage) before containing.

38. A malware detector adapting to attackers (10 pts)

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

(b) Poisoning injects mislabeled/malicious data that shifts the boundary or plants a backdoor; defense: vet data & track provenance, anomaly-check the training set.

(c) Few-shot (or transfer) learning - recognize a new attack from a handful of examples via a pretrained base.

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