EXAM 3 STUDY GUIDE - WEEKS 10-13

Attacks on AI, LLM Security, Ethics & Privacy

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

Colorado State University - Computer Information Systems

How to use this study guide

This guide pulls together everything Exam 3 covers (Weeks 10-13): attacks on ML systems; prompt injection & LLM security; AI ethics, bias & fairness; and privacy-preserving ML. Study actively - cover the answers and recall them, then redo the self-check questions. Exam 3 is online, 90 minutes, 100 points, open-notes, individual, no heavy math or coding.

Estimated review time: 45-60 minutes. Pair this with the Practice Exam and Handouts 8-11.

What Exam 3 Covers

Exam 3 is non-cumulative and covers Weeks 10-13 - one connected story: how AI systems are attacked, and how we make them safe, fair, and privacy-respecting. The exam rewards CONNECTING these ideas and applying them to business security scenarios (Part D), not memorizing definitions.

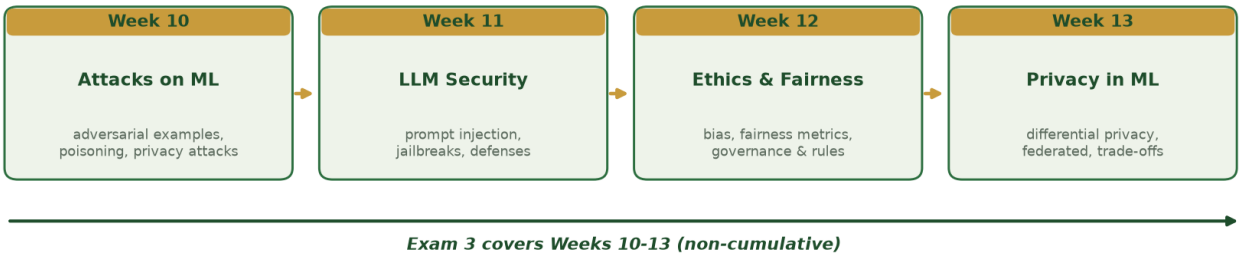

Figure 1. The arc: attacks on ML (W10) -> LLM security (W11) -> ethics & fairness (W12) -> privacy (W13).

Exam logistics

100 points, 90 minutes, online via Canvas (proctored with LockDown Browser). Format: 20 multiple-choice (40), 10 true/false (10), 5 short-answer (20), 3 applied scenarios (30). Open-notes and individual - no collaboration, no AI. No heavy math or coding.

Week 10 - Attacks on ML Systems

- Adversarial example: a tiny, crafted input change that flips a prediction (EVASION, at test time).
- Data poisoning: corrupt the TRAINING data (training time); a backdoor plants a hidden trigger.
- Privacy attacks: model inversion (reconstruct data), membership inference (was X in the set?), model theft.
- Why it works: models learn patterns, not meaning - and can memorize training data.
- Defenses: adversarial training, data vetting/provenance, detection, differential privacy - defense in depth.

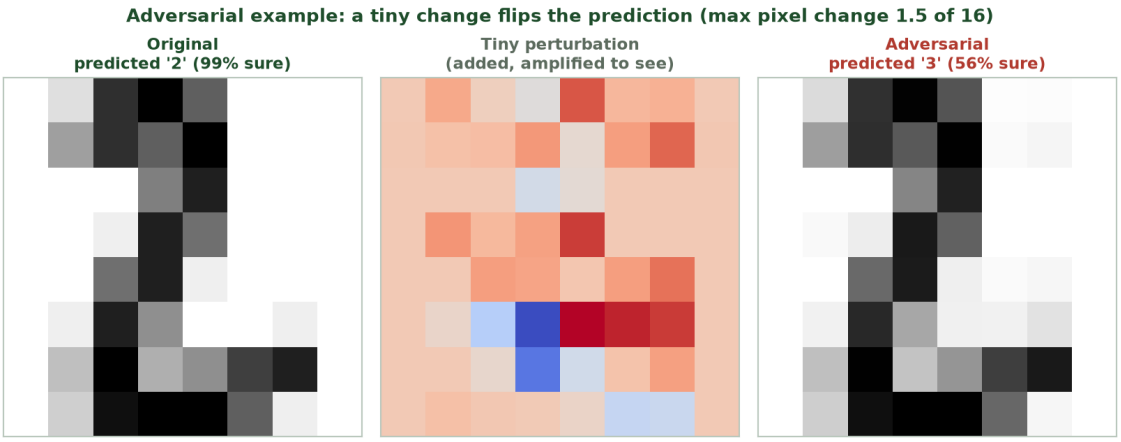

Figure 2. An adversarial example: a tiny change flips the prediction while looking identical.

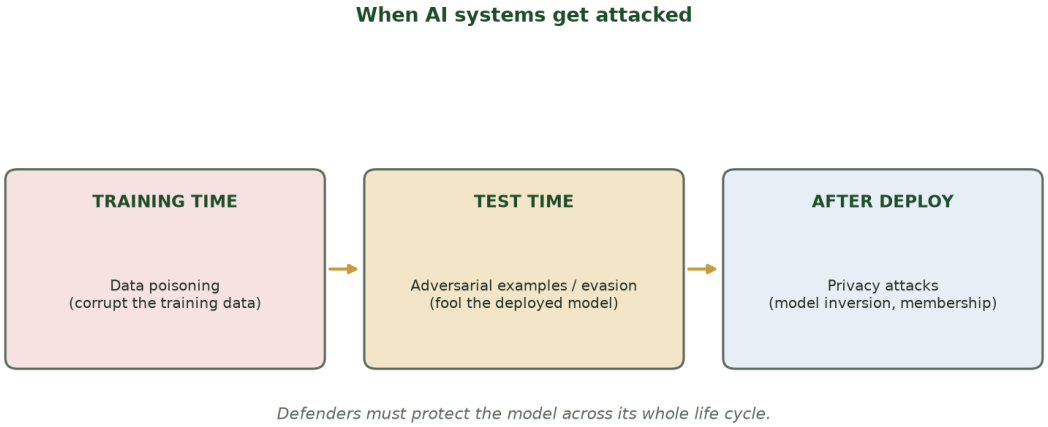

Figure 3. Attacks by phase: poisoning (training), evasion (test time), privacy attacks (after deployment).

Week 11 - Prompt Injection & LLM Security

- An LLM predicts the next word and FOLLOWS instructions in its input - it doesn't verify or understand.
- Prompt injection (OWASP LLM risk #1): attacker text overrides the app's rules.
- Variants: direct, indirect (hidden in a document the LLM reads), jailbreak, prompt leakage.
- Root cause: no wall between trusted instructions and untrusted data.
- Defenses: input & output guardrails, least privilege, human oversight - layered, no single fix.

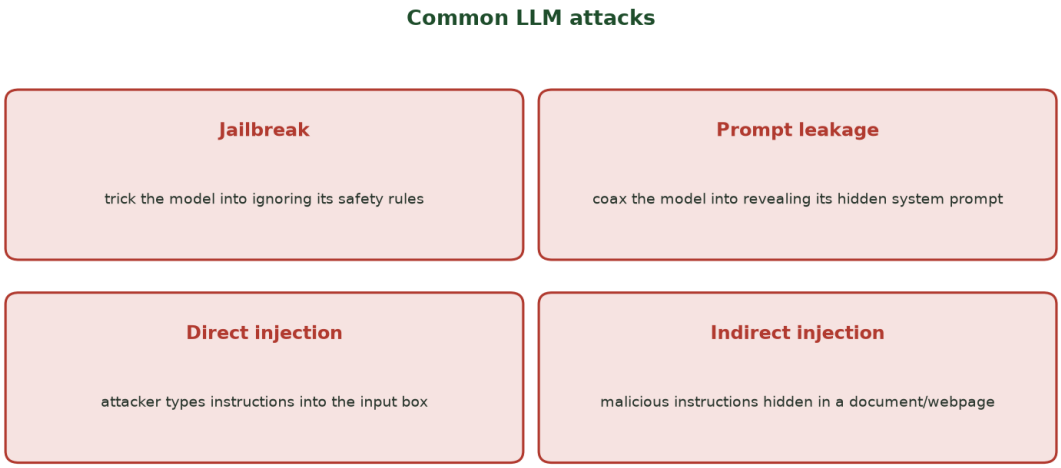

Figure 4. Four LLM attacks: direct/indirect injection, jailbreak, prompt leakage.

Figure 5. Defense in depth for an LLM app: strong prompt + guardrails + least privilege + oversight.

Week 12 - AI Ethics, Bias, Fairness & Governance

- A model can be accurate OVERALL yet unfair - always disaggregate metrics BY GROUP.
- Bias enters via data, model design (proxy features), deployment, and feedback loops.
- Fairness definitions (demographic parity, equal opportunity, equalized odds) can CONFLICT - you choose.
- Governance: GDPR, the EU AI Act (risk tiers), US/sector rules; model cards and audits.
- Mitigation: fix data, drop proxies, group-aware thresholds - then re-audit.

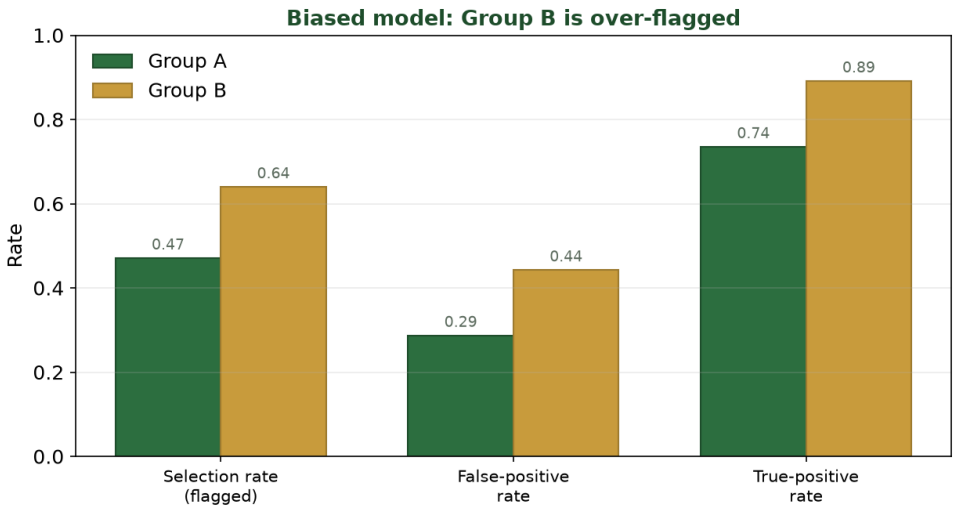

Figure 6. Per-group metrics reveal bias: the flagger over-flags Group B (higher FPR).

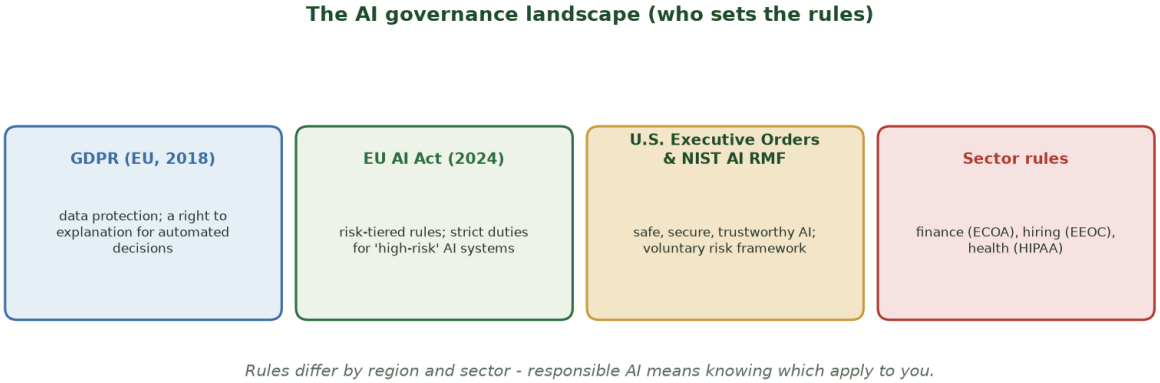

Figure 7. The governance landscape: GDPR, the EU AI Act, US/NIST, and sector rules.

Week 13 - Privacy-Preserving ML & Trade-Offs

- Models can leak data; 'just remove names' is weak (re-identification).
- Differential privacy: add calibrated noise; epsilon = privacy budget (smaller = more privacy, less accuracy).
- Federated learning: data stays on-device; only model updates are shared.
- Encrypted computation: compute on ciphertext - very strong, but slow.
- Trade-offs: privacy vs. accuracy (epsilon); security vs. privacy (balance, don't maximize).

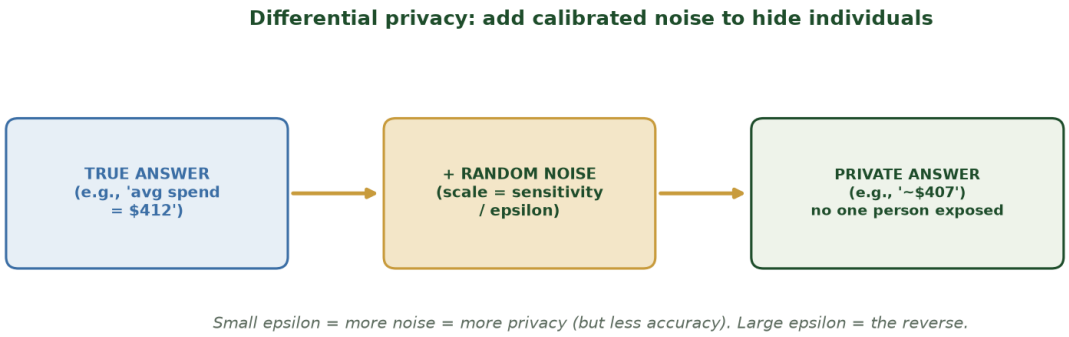

Figure 8. Differential privacy: true answer + calibrated noise -> a private released value.

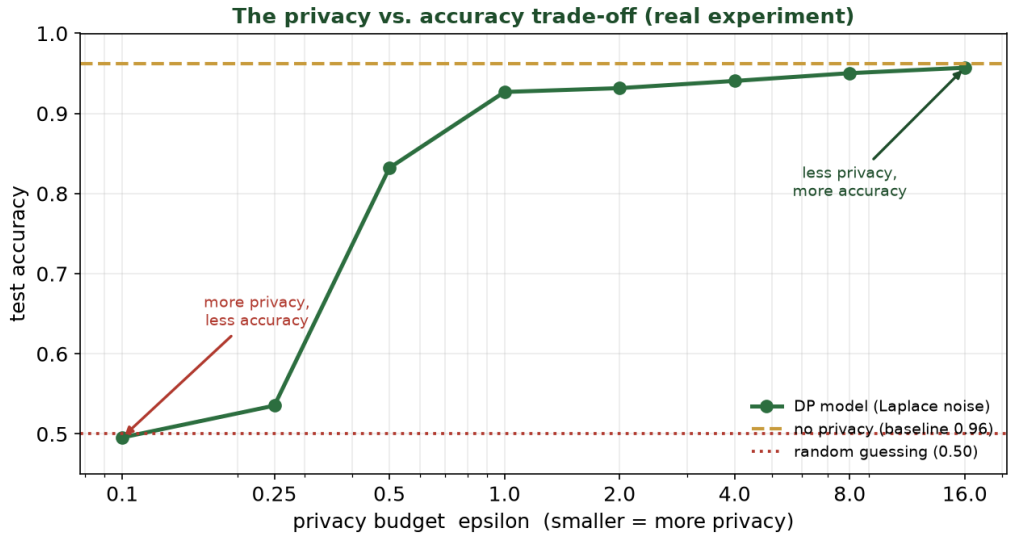

Figure 9. Privacy vs. accuracy: tiny epsilon -> ~0.50 (guessing); epsilon=1 -> ~0.93 vs a 0.96 baseline.

Key Terms to Know Cold

- Adversarial example; evasion; data poisoning; backdoor.
- Model inversion; membership inference; model extraction/theft.
- LLM; prompt injection (direct/indirect); jailbreak; prompt leakage; guardrail; least privilege.
- Fairness; bias; proxy feature; demographic parity; equal opportunity; equalized odds.
- Governance; GDPR; EU AI Act; model card; bias audit.
- PII; differential privacy; epsilon; federated learning; encrypted computation.
- Security-vs-privacy trade-off; defense in depth; human in the loop.

Self-Check Questions

1. How do poisoning and evasion differ (when does each happen)?
2. Name two privacy attacks on a model and what each reveals.
3. What is prompt injection, and why isn't a strong system prompt enough?
4. Why can a model be accurate overall but unfair? Name one fairness definition.
5. What does epsilon control in differential privacy, and what's the trade-off?
6. How does federated learning protect privacy, and how is it different from encrypted computation?

7. Give one security-vs-privacy trade-off and a balanced response.
8. Why is there 'no perfect defense,' and what should defenders do instead?

If you can answer all eight from memory and apply them to a new scenario, you are ready. Remember: AI assistance and collaboration are not permitted on the exam.

References

- NIST AI 100-2 (Adversarial ML); NIST AI RMF 1.0; NIST SP 800-226 (differential privacy).
- OWASP Top 10 for LLM Applications; MITRE ATLAS.
- S. Barocas, M. Hardt, A. Narayanan, 'Fairness and Machine Learning' (fairmlbook.org).
- EU GDPR (2018) and the EU AI Act (2024).
- C. Dwork & A. Roth, 'The Algorithmic Foundations of Differential Privacy' (2014).
- H. B. McMahan et al., federated learning (2017); Handouts 8-11.