EXAM 4 STUDY GUIDE (FINAL) - WEEKS 10-16

Attacks, LLM Security, Ethics, Privacy, Trust & Beyond

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 4 covers (Weeks 10-16): attacks on ML; prompt injection & LLM security; AI ethics, bias & fairness; privacy-preserving ML; trustworthy AI & incident response; and advanced topics & emerging threats. Study actively - cover the answers and recall them, redo the self-check questions, then take the Practice Exam timed. Exam 4 is online, open-notes, 100 points, non-cumulative (Weeks 10-16), no heavy math or coding.

Estimated review time: 60-75 minutes. Pair with the Practice Exam and Handouts 8-13.

What Exam 4 Covers - the Whole Second Half

Exam 4 is non-cumulative and covers Weeks 10-16 (Week 16 is review, so 10-15 content). The arc: how AI systems are attacked, and how we make them fair, private, robust, and trustworthy - plus what's on the horizon. The exam rewards CONNECTING these ideas (especially in the applied scenarios), not memorizing definitions.

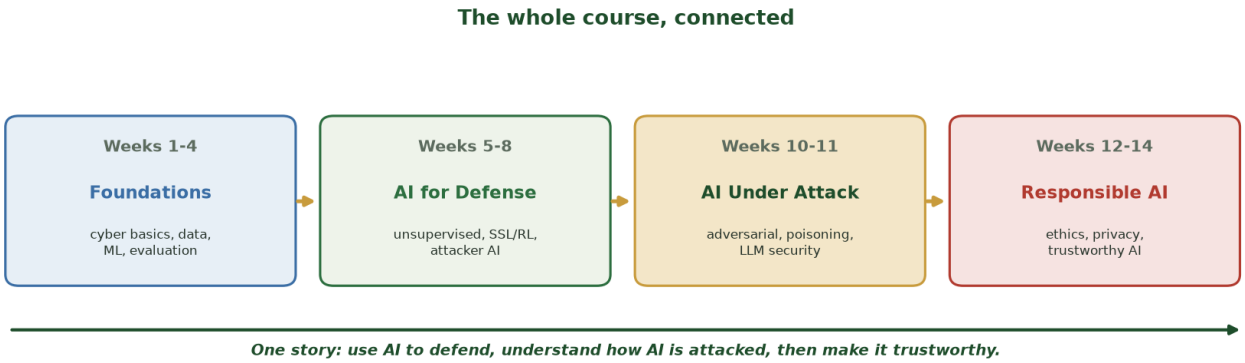

Figure 1. The whole course as one arc; Exam 4 covers the second half (Weeks 10-16).

Exam logistics

100 points, online via Canvas (LockDown Browser), during the final-exam window. 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, test time).
- Data poisoning: corrupt the TRAINING data (training time); a backdoor plants a hidden trigger.
- Privacy attacks: model inversion, membership inference, model theft.
- Defenses: adversarial training, data vetting/provenance, detection, DP - defense in depth.

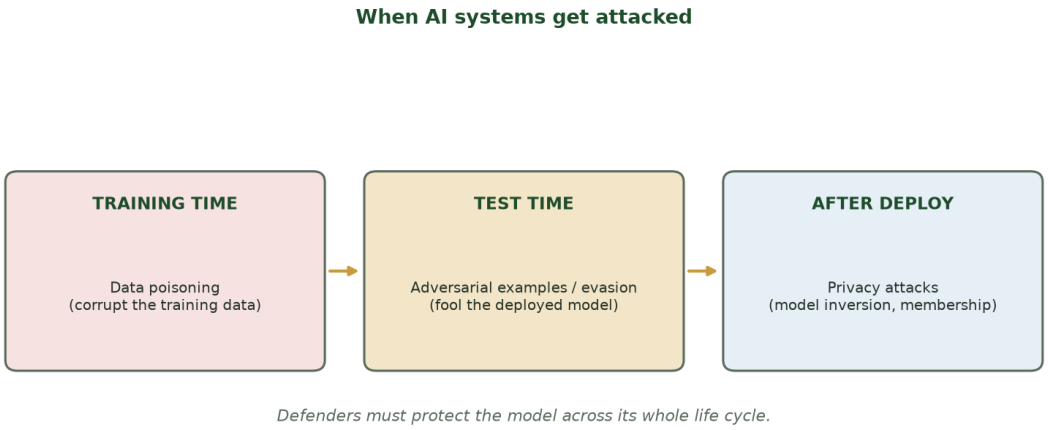

Figure 2. Attacks by phase: poisoning (training), evasion (test time), privacy attacks (after deployment).

Week 11 - Prompt Injection & LLM Security

- LLMs predict the next word and FOLLOW instructions in their input.
- Prompt injection (OWASP LLM #1): attacker text overrides the app's rules.
- Variants: direct, indirect (hidden in a document), jailbreak, prompt leakage.
- Defenses: input & output guardrails, least privilege, human oversight - layered, no single fix.

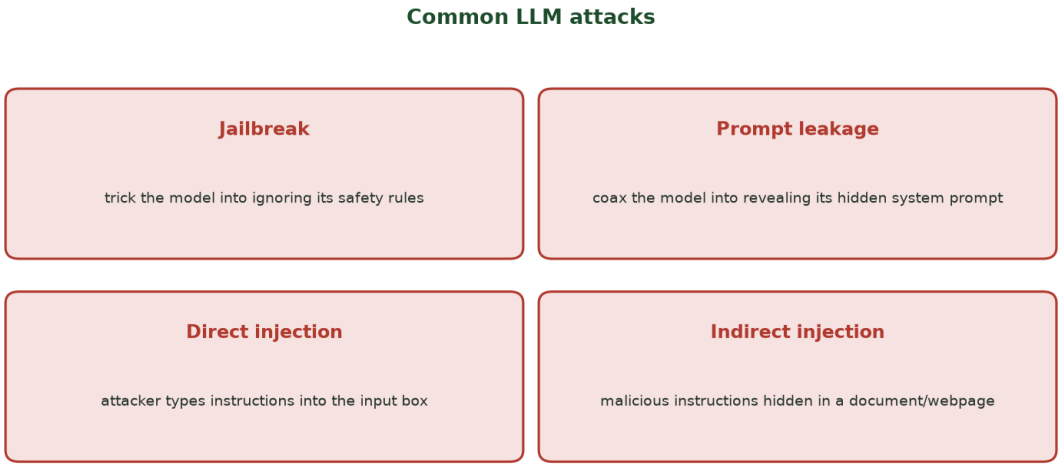

Figure 3. Four LLM attacks: direct/indirect injection, jailbreak, prompt leakage.

Week 12 - AI Ethics, Bias, Fairness & Governance

- A model can be accurate OVERALL yet unfair - disaggregate metrics BY GROUP.
- Bias enters via data, model design (proxies), deployment, and feedback loops.
- Fairness definitions (parity, equal opportunity, equalized odds) can CONFLICT - you choose.
- Governance: GDPR, EU AI Act (risk tiers), US/sector rules; model cards & audits.

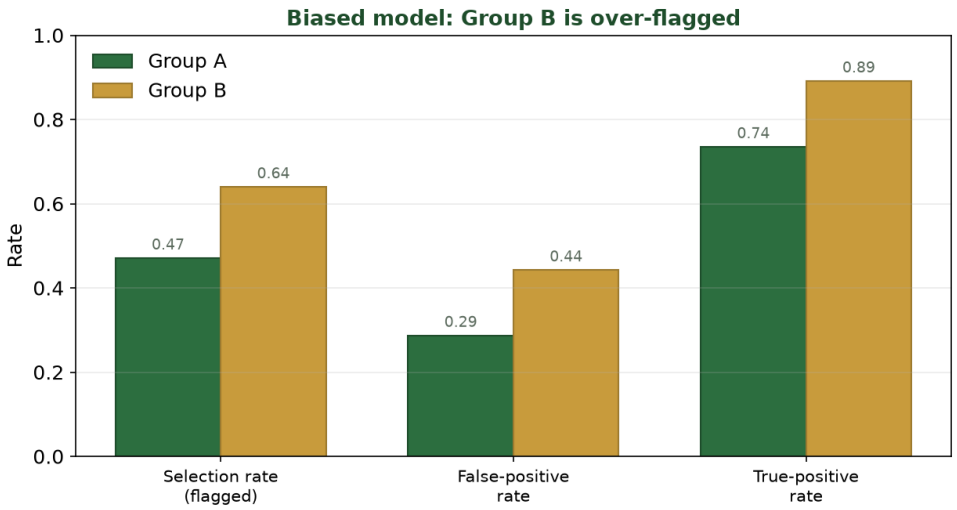

Figure 4. Per-group metrics reveal bias: the flagger over-flags Group B (higher FPR).

Week 13 - Privacy-Preserving ML

- Models can leak data; 'just remove names' is weak (re-identification).
- Differential privacy: add calibrated noise; epsilon = privacy budget (smaller = more privacy).
- Federated learning: data stays on-device; only model updates are shared.
- Encrypted computation: compute on ciphertext; trade-offs: privacy vs. accuracy, security vs. privacy.

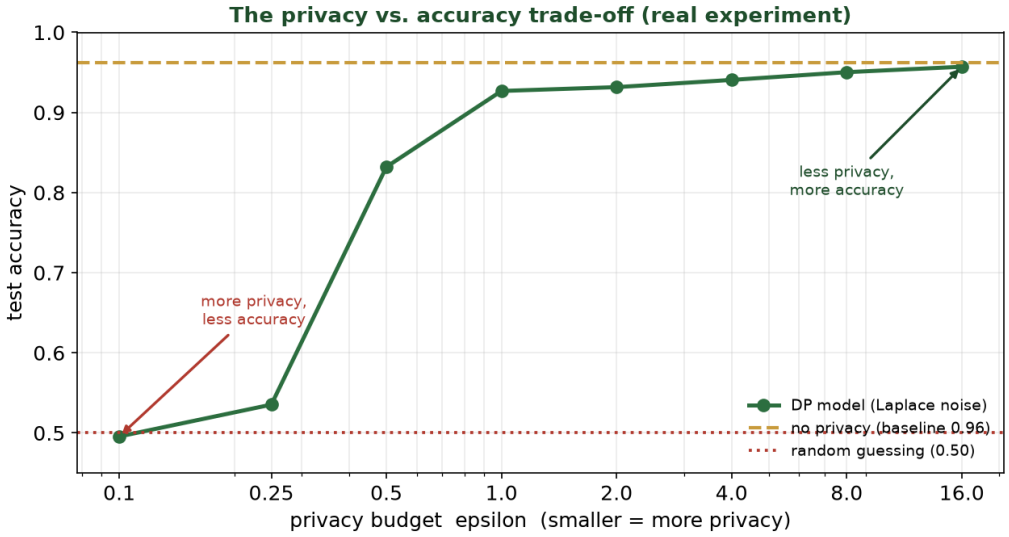

Figure 5. Privacy vs. accuracy: tiny epsilon -> guessing; epsilon=1 -> ~0.93 vs a 0.96 baseline.

Week 14 - Trustworthy AI & Incident Response

- Four pillars: transparency, explainability, robustness, accountability.
- Transparency = how the system works; explainability = why THIS decision.
- Robustness must be MONITORED (data drift); accountability = owners, override, appeals.
- IR lifecycle: prepare; detect & analyze; contain/eradicate/recover; post-incident (hotwash).

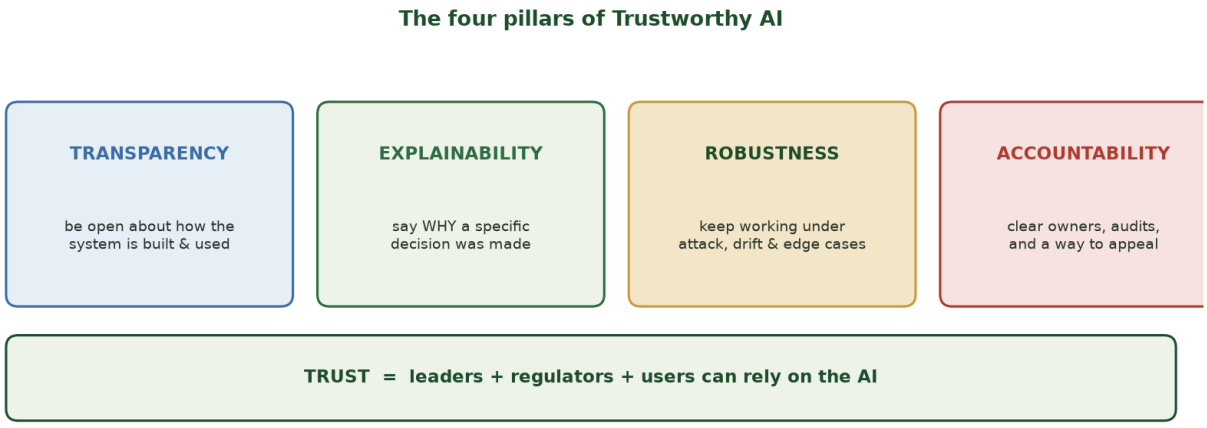

Figure 6. The four pillars of trustworthy AI.

Week 15 - Advanced Topics & Emerging Threats

- Transfer, few-shot, and federated learning all fight LABEL SCARCITY.
- Transfer reuses a pretrained model (far fewer labels needed).
- Emerging threats: AI-generated malware, poisoning at scale, deepfakes, autonomous agents.
- It's an arms race - no permanent win; raise cost, detect, adapt (defense in depth).

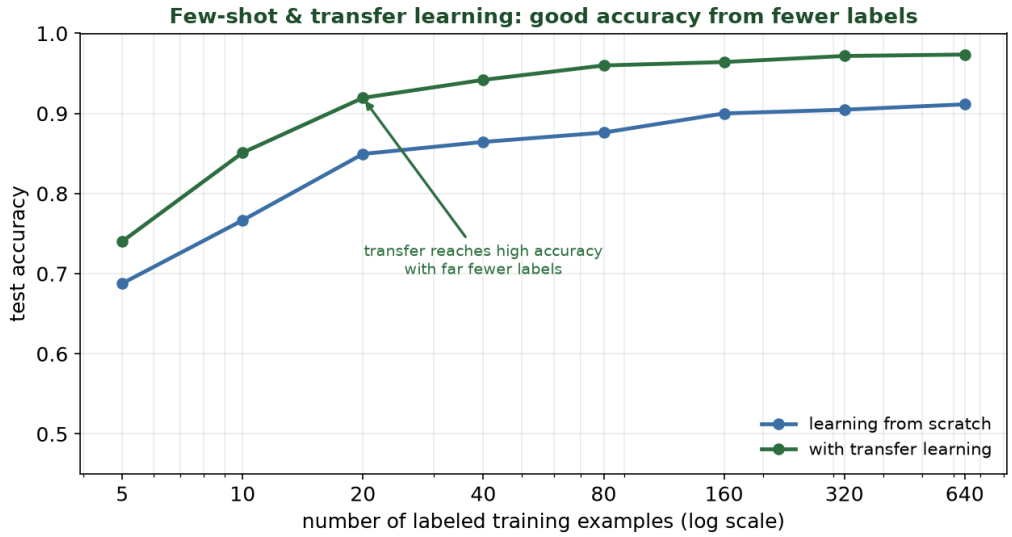

Figure 7. Transfer/few-shot learning: good accuracy from far fewer labels.

Fix These Common Misconceptions

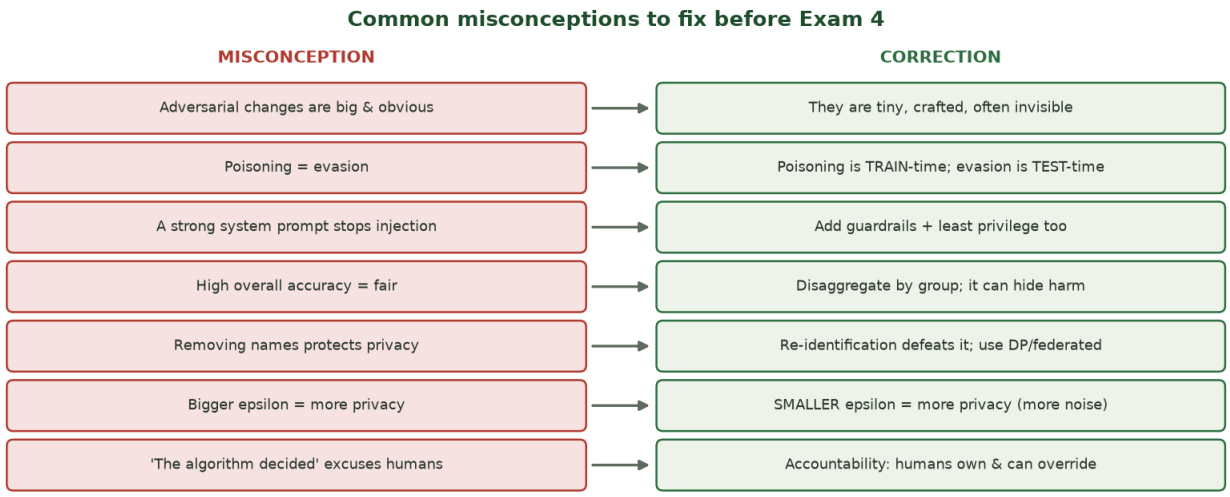

Figure 8. The most common exam mix-ups (left) and the corrections (right).

Exam Game Plan

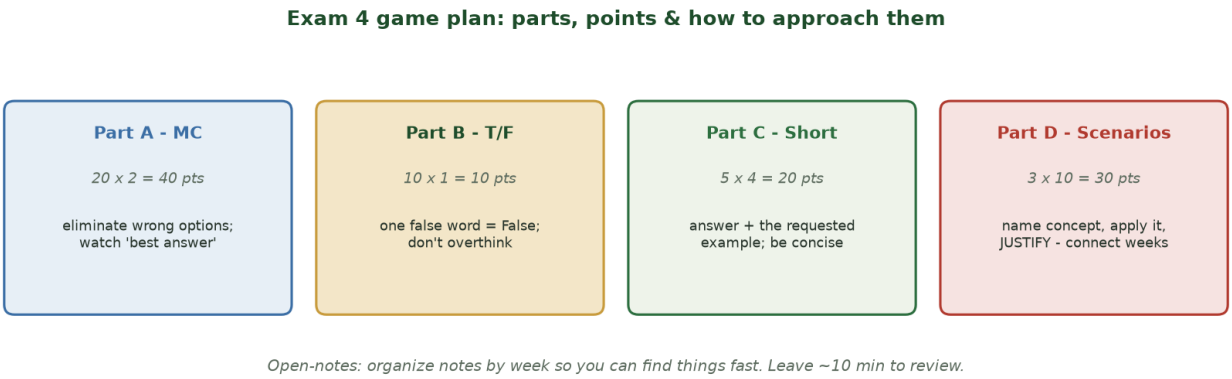

Figure 9. How to approach each part - and leave ~10 minutes to review.

Self-Check Questions (all should be answerable from memory)

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 which way is more private?
6. How does federated learning protect privacy?
7. Name the four trustworthy-AI pillars and the four IR-lifecycle phases.
8. What problem do transfer, few-shot, and federated learning all address?
9. Give one emerging threat and a defense; why is there 'no perfect defense'?

If you can answer all nine 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 (DP); NIST SP 800-61 (IR).
- 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 (differential privacy, 2014).
- Pan & Yang (transfer learning); McMahan et al. (federated learning); Handouts 8-13.