AI Security Labs – Trainee Report Template

Participant

- Name:
- Date:
- Team/Function:

Environment

- OS / Version:
- Python Version:
- CPU/GPU:
- Anything unusual:

Lab 1 – Prompt Injection

- Prompts attempted & outcomes:
- Observations:
- Controls/Mitigations:

Lab 2 – Adversarial Examples

- Epsilon runs & success:
- Attach adv_result.png
- Observations:

Lab 3 – Model Extraction

- Baseline & Mitigated results:
- 401/429 logs:
- Mitigations to adopt:

Lab 4 – Data Poisoning Detection

- Parameters & accuracies:
- Observations:

Framework Mapping

- NIST AI RMF: GV.3 GV.4 MP.3 MP.4 MP.5 MS.1 MS.2 MS.3 MG.2 MG.3 MG.4 MG.5
- MITRE ATLAS: Prompt Injection, Adv. Examples, Model Extraction, Data Poisoning, Model Integrity Compromise

Final Notes & Next Actions

- Production hardening ideas:
- Follow-up experiments: