CSE 6392: Advanced Topic: Seminar in LLM Security, Privacy & Ethics

Course Code / Term: CSE 6392
Instructor: Faysal Hossain Shezan
Class Format: Weekly seminar, 3 hrs, discussion + student presentations
Prerequisites: Graduate-level background in ML, security and privacy (or instructor permission)

Course Description

This seminar examines the security, privacy, and ethical implications of large language models (LLMs) and related AI technologies. We begin with foundational adversarial machine learning concepts, then progress into contemporary vulnerabilities in LLMs (jailbreaking, prompt injection, agentic AI), modalities (deepfakes), platform implications (social media, software security), and finish with ethics and governance. Students engage deeply with cutting-edge research papers, lead seminar discussions, and produce a final project exploring a novel problem or defence mechanism.

Learning Outcomes

- Critically analyse adversarial attacks and defences in classical ML and LLM settings.
- Identify and categorise novel vulnerabilities in LLM-based systems (e.g., jailbreaks, prompt-injection, agentic behaviours).
- Design or critique guardrail/mitigation strategies for LLMs and AI systems.
- Assess the social, platform, and governance implications of AI deployment on privacy, misinformation, and ethics.
- Conduct a small research project (experiment, or design proposal) and present findings in writing and orally.

Weekly Schedule and Content

**We have listed potential papers; but we will update the list later.

**We also plan to invite our industry collaborators and connections to give invited talk on some topic based on their expertise (remotely/in-person).

Week	Topic	Required Readings	Assignments / Activities
1	Introduction & Course Overview: Landscape of ML security, LLMs & ethics	Course syllabus & logistics; Short overview of adversarial ML + LLMs	Student introductions + identify project interests
2	Classic Adversarial ML – Attacks	Explaining and Harnessing Adversarial Examples (arXiv 1312.6199); Towards Deep Learning Models Resistant to Adversarial Attacks (arXiv	Discussion: mechanisms of adversarial examples, threat models

		1412.6572)	
3	Classic Adversarial ML – Defences	Adversarial Training Methods for Deep Networks (arXiv 1707.08945); Universal Adversarial Perturbations (arXiv 1706.06083); Towards Evaluating the Robustness of Neural Networks (arXiv 1608.04644)	Discussion: robustness bounds, limits of defences; brief reflection write-up
4	LLM Security & Privacy (Foundations)	LLM Security and Privacy: An Overview (USENIX Security 2024 talk); Arxiv 2307.15043; elder-plinius/L1B3RT4S (GitHub project)	Discussion: 'What makes LLMs different from classical ML in terms of security/privacy?'
5	Jailbreaking of LLMs	Jailbreaks of Large Language Models: Taxonomy & Cases (arXiv 2312.02119); Misuse of LLMs via Jailbreaking (arXiv 2404.02151); Advanced Jailbreak Techniques in LLMs (arXiv 2404.16873)	Seminar discussion: ethics of red-teaming, possible defences; mini-assignment: find a recent jailbreak instance and summarise
6	Prompt Injection & Retrieval-Augmented Systems	EmbraceTheRed blog: 'Prompt Injection in LLMs'; Prompt Injection Attacks on LLM-based Systems (arXiv 2403.03792)	Discussion+In-class exercise: craft a hypothetical prompt-injection attack and propose mitigation
7	Agentic AI & Multi-Step Reasoning Attacks	Agentic AI: Risks and Opportunities (arXiv 2409.11295); Security Implications of Agentic Large Language Models (arXiv 2501.089702); Tool-Using LLMs and Attack Surfaces (arXiv 2410.09024)	Group discussion: vulnerabilities in agentic systems; begin project brainstorming + Project Status Check
8	GuardRails, Alignment & Monitoring	GuardRails for LLM Systems: Design and Challenges (arXiv 2406.04313); Real-World Alignment Frameworks (arXiv 2309.00614v2)	Debate: trade-offs in alignment (usability vs safety); project idea check-in
9	Deepfakes & Multimodal Threats	Generating and Detecting Deepfakes: A Survey (arXiv 2311.03191); When ChatGPT Detects Deepfakes (CVPR 2024 W); Deepfake Generation via LLM-Fusion (arXiv 2406.05946)	Case study presentations: deepfake attack + detection strategy

10	Social Media, Platform Abuse & Misinformation	AI-Driven Misinformation Campaigns on Social Media (arXiv 2310.19181); Platform Security and AI in Social Media Ecosystems (USENIX Security 2025)	Round-table: policy, accountability, technical defences
11	Software Security, LLMs in DevOps & Code Generation	Security Vulnerabilities in LLM-Generated Code (arXiv 2307.02192); AI Code Generation and Supply Chain Risks (USENIX Security 2023); Emerging Software Security Challenges (arXiv 2506.025483)	Discussion: Review a code snippet generated by an LLM; identify potential flaws; project mid-point check
12	Ethics, Governance & Responsible AI Deployment	Ethics of Large Language Models (ACL 2023 Long); AI Governance, Fairness, Transparency in LLMs (CL 2024 3.8)	Discussion: roles of researchers, deployers, regulators; final project prep
13	Student Project Presentations	—	20-minute project presentations + peer feedback
14	Wrap-up, Reflection & Future Directions	—	Final reflection submissions; discussion of open problems

Assessments & Grading

- Participation & Seminar Discussion – 20%
- Short assignments – 20%
- Seminar lead – 20%
- Final Project (paper/report + presentation) – 30%
- Attendance – 10%
- Bonus: up to 10% extra credit

Final Project Options

- Literature review of an emergent vulnerability in LLMs or agentic systems.
- Design and (optionally) prototype a mitigation/guardrail mechanism.
- Empirical study (e.g., applying jailbreak or prompt-injection technique and analysing outcomes).
- Policy/governance white-paper focusing on an ecosystem (social media, software supply chain, deepfakes).
- We will provide more details about project and grading later.

Deliverables

Proposal Due Week 3;

Project Status Update (Phase 1) Week 8;

Project Status Update (Phase 2) Week 11;

Final Report + Presentation Week 14.

Policies

- Late assignments penalized unless prior approval given.
- Academic integrity and citation required.
- Accessibility accommodations available upon request.