CSCE 585: Machine Learning Systems

University of South Carolina — Spring 2025
Instructor: Prof. Pooyan Jamshidi

Course Description

Machine Learning (ML) has transformed nearly every industry, but deploying ML models in real systems is far more complex than training them in isolation. This course explores the emerging discipline of **Machine Learning Systems (MLSys)** — the intersection of ML, systems, and software engineering. Students will learn to critically read MLSys papers, design reproducible experiments, and implement scalable, efficient, and trustworthy ML systems.

We will cover topics ranging from **ML in production pipelines**, **causal reasoning for robustness**, and **experiment design**, to modern frontiers such as **LLM serving & scaling**, **compound/multi-agent LLM architectures**, **hardware efficiency & sustainability**, and **privacy, safety, and security**. The course emphasizes **hands-on replication of MLSys research**, engagement with state-of-the-art frameworks (e.g., vLLM, LangChain, Ray Serve, MLPerf), and a **capstone project**.

Learning Objectives

By the end of the course, students will be able to:

1. Critically evaluate MLSys research using structured reading and critique methods.
2. Design and analyze reproducible ML experiments with appropriate metrics.
3. Replicate and extend results from cutting-edge MLSys papers.
4. Understand challenges in production ML systems (MLOps, feature stores, retraining).
5. Engage with modern MLSys frontiers: LLM systems, compound architectures, sustainability, and privacy/safety.
6. Communicate effectively in written reports and oral presentations.

Course Format

- Weekly lectures and paper discussions
- Replication & lab assignments (short, hands-on experiments)
- Capstone project (team-based, with milestones and final presentation)

Grading Breakdown

Reading & Critique Assignments	15%
Replication & Lab Assignments	30%
Capstone Project	40%
Participation & Discussions	15%
Total	100%

Tentative Weekly Schedule

Part I: Foundations (Weeks 1–4)

1. Course Overview & Motivation
2. How to Read and Critique MLSys Papers
3. Designing & Motivating Experiments (InferLine case study)
4. Project Milestone 1: Problem Motivation

Part II: Core MLSys Challenges (Weeks 5–8)

5. ML in Production & MLOps
6. Designing ML Systems
7. Causal AI & Robust ML Systems
8. Guest Lecture(s): Robotics Causality / Causal Bayesian Optimization

Part III: Modern Frontiers (Weeks 9–11)

9. LLM Systems: Serving & Scaling
10. LLM Systems: Compound & Multi-Agent Architectures
11. Hardware & Sustainability in ML Systems

Part IV: Trustworthy MLSys (Weeks 12–13)

12. Privacy, Safety, & Security in ML Systems
13. Reproducibility & Benchmarking in MLSys

Part V: Capstone (Weeks 14–15)

14. What Makes an Impactful MLSys Paper
15. Capstone Project Presentations

Assignments

- **Paper Reading & Critique** (weekly short writeups)
- **Replication & Lab Assignments:**
  - InferLine latency vs. cost trade-offs
  - vLLM serving efficiency
  - RAG pipeline evaluation (accuracy vs. latency)
  - Energy profiling of inference

  - Privacy & adversarial robustness experiment

- **Capstone Project:**
  - Milestone 1: Problem Motivation
  - Milestone 2: System Design + Initial Experiments
  - Final Report (8–10 pages)
  - Final Presentation

Participation

Active participation in paper discussions and peer feedback is expected. Students should come prepared to ask questions, challenge ideas, and contribute insights from both research and industry.

Academic Integrity

All students must adhere to the University of South Carolina's academic integrity policies. Collaboration is encouraged for discussions, but submitted work must reflect individual understanding unless explicitly team-based.