STOR 890: TRUSTWORTHY MACHINE LEARNING, SPRING 2026

Course Information

This topics course is designed for graduate students who are interested in the emerging area of AI security and trustworthy machine learning. Modern machine learning systems are increasingly deployed in high-stakes applications, making it essential to understand their vulnerabilities, limitations, and the principles needed to ensure reliability. The goal of this course is to help students develop a solid conceptual and practical foundation in the security and trustworthiness aspects of machine learning by studying core threat models, analyzing state-of-the-art research papers, and working on research-oriented projects. Upon completion of this course, students will be able to:

- Understand major threat models in machine learning, including adversarial examples, data poisoning, backdoor attacks, and privacy leakage.
- Analyze and evaluate defense strategies such as adversarial training, robustness certification, backdoor detection, privacy-preserving learning, and uncertainty quantification.
- Build familiarity with deep learning architectures (e.g., convolutional networks, transformers, generative models) as the basis for understanding security vulnerabilities.
- Critically read, present, and synthesize recent research papers in trustworthy machine learning.
- Carry out a research-oriented or survey-style final project on a topic related to AI security.

Lecture:	TTH 8:00am - 9:15am, Hanes 125
Course Website:	`https://liyao880.github.io/stor890/`

Instructor

Instructor:	Yao Li
Office:	Hanes 334
Email:	yaoli@email.unc.edu
Office Hours:	W 11:00am - 12:00pm.
Website:	`https://liyao880.github.io/yaoli/`

Prerequisites

Students are expected to have prior coursework in machine learning or deep learning (such as STOR 565 or STOR 566, or an equivalent course). Familiarity with Python is assumed. Prior exposure to machine learning research papers is helpful but not required.

Topics

This course will cover both foundational and cutting-edge topics in AI security and trustworthy machine learning. We will discuss major threat models, theoretical insights, and practical defenses across classical deep learning and modern large models. Tentative topics include:

- Backdoor
- Data poisoning
- Privacy attacks
- Adversarial robustness

- Robustness certification
- Explainable and interpretable ML
- Secure and trustworthy federated learning
- Uncertainty quantification for safety: calibration, detection of distribution shift
- Security challenges in large language models: jailbreak attacks, safety alignment failures, prompt vulnerabilities, hallucination analysis

Additional topics may be added depending on student interest and recent developments in the field.

Grading

Paper Reviews	Final Project	Paper Presentation	Participation	Total
15%	40%	20%	25%	100%

Paper Reviews

- Students will receive full credit for this component by submitting reviews for *half* of the assigned papers. Additional reviews are welcome but will not further increase the grade.
- Each review should demonstrate genuine engagement with the paper. At minimum, it should (i) summarize the main idea in the student's own words, (ii) identify strengths and weaknesses, and (iii) raise at least one question for discussion.
- Outstanding reviews that show exceptional insight, clarity, or depth of analysis may receive **extra credit**, at the discretion of the instructor.
- Reviews must be submitted before the corresponding lecture; late submissions will not be accepted.
- Reviews that rely heavily on large language models without meaningful synthesis or critique, or that otherwise demonstrate poor engagement, may receive **no credit**.

Final Project

This course includes a final project in lieu of a final exam. Projects may be completed **individually or in groups of two**. Groups of more than two are not permitted. The final project consists of:

- Project proposal (10%)
- Project presentation (40%)
- Project paper (50%)

Project Topics: I will meet with each student or group to discuss potential project topics. Suitable topics include, but are not limited to:

- Conducting a careful empirical study comparing state-of-the-art methods;
- Reproducing an influential research paper and analyzing its limitations;
- Developing a small methodological or algorithmic extension;
- A structured survey of a focused sub-area in trustworthy machine learning.

Project Proposal: The proposal is limited to 2 pages (excluding references) and should include:

- The problem you aim to address;
- A brief review of related work;
- The method(s) you plan to use or compare;
- Evaluation metrics and expected outcomes;
- References.

Please use the LaTeX template at link for the proposal.

Project Presentation: Presentations will take place during the final 2–3 lectures of the semester. Each student or group will give a short presentation (length announced later) summarizing the problem, approach, results, and conclusions. Attendance is required for all presentations.

Project Paper: Students must submit a written final report in PDF format. The report must use the NeurIPS LaTeX style files and should be no more than 8 pages excluding references (there is no minimum length requirement). The report may include a discussion of possible future extensions.

Paper Presentation

- Each student will present one to two research papers during the semester, depending on enrollment. Presentations are expected to be completed individually. In exceptional circumstances, certain papers may be presented in pairs (e.g., papers that are unusually long or technically demanding).
- Students who present more than two papers will receive **extra credit** toward the final course grade (up to 5 additional percentage points).
- Presentations should clearly communicate the problem, methodology, main results, and limitations, and should facilitate classroom discussion.
- Students are expected to prepare 2–3 discussion questions and lead the Q&A for their assigned paper.
- Slides must be submitted before class, and late submissions will receive penalties.

Participation

- Active participation is an essential component of this topics course. Students are expected to contribute to class discussions by asking questions, offering insights, commenting on papers, and engaging respectfully with peers.
- To receive full credit for the participation component, each student should make approximately 10 meaningful contributions over the semester. A meaningful contribution may include a thoughtful question, a critique, a clarification, or a connection to related work.
- A shared Google Sheet will be provided for students to self-report their participation after each class. Students should record the date and a brief note (one sentence) describing their contribution. The instructor will verify all entries after each class.
- Students are expected to read the assigned paper(s) before class and be prepared to participate.

Academic Integrity and AI tools

All homework and analysis assignments must be completed individually. Assistance from other students, AI tools (e.g., ChatGPT), or using previously uploaded work from other sources (e.g., CourseHero) is strictly prohibited. This policy also applies to project work; AI tools are not allowed to aid in the completion of any projects. Violations of this policy will result in a grade of 0 for the assignment or project. Additionally, any alleged violations will be reported to the University of North Carolina (UNC) for further review and potential disciplinary action.

Notes

The Instructor reserves the right to make any changes she considers academically advisable.

Attendance

Regular class attendance is a student obligation, and a student is responsible for all the work, including tests and written work, of all class meetings. No right or privilege exists that permits a student to be absent from any class meetings except for excused absences for authorized University activities or religious observances required by the student's faith. If a student misses three consecutive class meetings, or misses more classes than the course instructor deems advisable, the course instructor may report the facts to the student's academic dean. (See details at `https://catalog.unc.edu/policies-procedures/attendance-grading-examination/#text`)

Honor Code

`http://instrument.unc.edu/`

Accessibility

`https://ars.unc.edu/`

Counseling

`https://caps.unc.edu/`

Title IX

Any student who is impacted by discrimination, harassment, interpersonal (relationship) violence, sexual violence, sexual exploitation, or stalking is encouraged to seek resources on campus or in the community. Please contact the Director of Title IX Compliance (Adrienne Allison – Adrienne.allison@unc.edu), Report and Response Coordinators in the Equal Opportunity and Compliance Office (reportandresponse@unc.edu), Counseling and Psychological Services (confidential), or the Gender Violence Services Coordinators (gvsc@unc.edu; confidential) to discuss your specific needs. Additional resources are available at safe.unc.edu.