CSC 5991: Trustworthy AI for Large Language Models and Vision-Language Models (Winter 2026)

Wayne Artificial Intelligence

Course Description

This course explores the foundations and frontiers of trustworthy AI in the era of large language models (LLMs) and vision-language models (VLMs). Students gain technical grounding in transformer architecture, pretraining, prompting, and multimodal learning. Building on this, the course examines principles and practices for ensuring trustworthiness, including safety alignment, interpretability, fairness, robustness, privacy, provenance, and regulatory compliance. Through lectures, hands-on labs, and team projects, students will critically analyze risks, implement mitigation techniques, and design responsible AI systems for real-world applications.

Prerequisites: CSC 2200 with a minimum grade of C or graduate standing

Course Instructor

Dongxiao Zhu

- Professor of Computer Science

Research interests: Trustworthy Artificial Intelligence, Adversarial Machine Learning, AI in Science, Health and Mobility

Graduate Student Instructors

Ujunwa Mgboh
PhD candidate

Rafi Sultan
PhD candidate

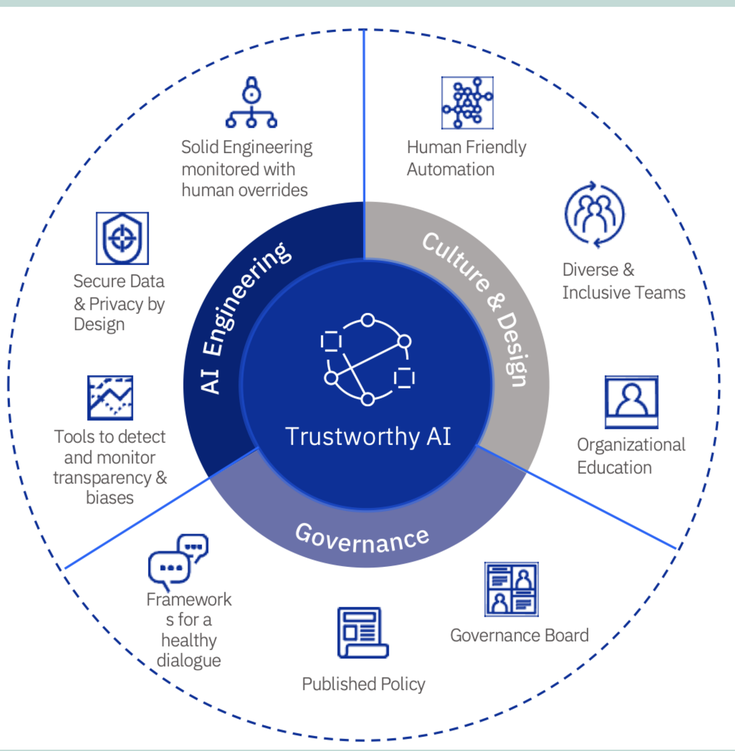
Xiangyu Zhou
PhD candidate

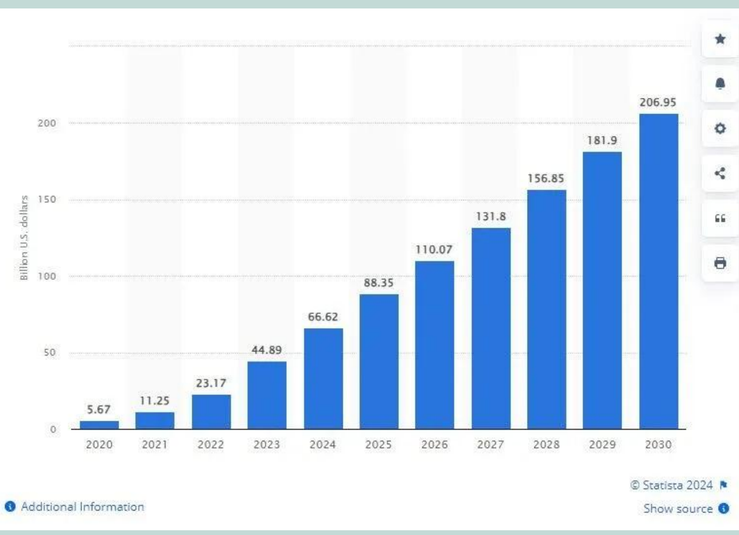
Amin Roshani
PhD candidate

Saleh Zare Zade
PhD candidate

Hui Zhu
PhD candidate

Trustworthy Generative AI (GenAI) Use Cases





Course Learning Objectives

- ❑ **Explain the architectural foundations of LLMs and VLMs.**
- ❑ **Apply prompting, fine-tuning, and multimodal techniques to build and adapt foundation models for downstream tasks.**
- ❑ **Evaluate trustworthiness dimensions, safety, robustness, interpretability, fairness, and privacy, in LLMs and VLMs using established benchmarks.**
- ❑ **Implement alignment and safety approaches such as RLHF, Constitutional AI, red-teaming, interpretability tools, and unlearning methods.**
- ❑ **Collaborate in team projects to design, evaluate, and document trustworthy AI applications with both technical and societal considerations.**

CONTACT US

+1 (313)-577-3104

dzhu@wayne.edu

https://engineering.wayne.edu/profile/ct4442

Course Schedule

Week 1	Introduction: Deep Learning Foundations for LLMs and VLMs **Labs:** Ways to dissect alignment issues in LLMs
Week 2	Pretraining & Fine-tuning Strategies Labs: Supervised Fine Tuning with small LMs
Week 3	Prompting for LLM Inference: Few-Shot & Chain-of-Thought Labs: Implementing In-Context & Reasoning Prompts with small LLMs
Week 4	Vision-Language Models (VLMs) Labs: Image classification using CLIP model
Week 5	Large Vision-Language Models (LVLMs) & Hallucinations in LVLMs Labs: Hands-on experiments on LLaVA, Learning how to do multi-modal inputs, Ways to fine-tune
Week 6	Trustworthy AI, risk, and alignment Labs: Ways to dissect alignment issues in LLMs
Week 7	Safety Evaluations & Red-Teaming for LLMs Labs: Implementing Red-Team Attacks, Defenses & Evaluation for small LLMs
Week 8	Unlearning & Hazardous Knowledge Mitigation Labs: Implementing Targeted and Untargeted LLM Unlearning Methods.
Week 9	Grounded Large Vision-Language Models (LVLMs) (Fine-Grained Application) Labs: Hands-on running SAM, using LLaVA+SAM to learn how to generate grounded conversation
Week 10	Agentic AI Safety Labs: Safety in a small agent AI
Weeks 11 -14	Project Presentation