USC Viterbi
School of Engineering

Course ID and Title: CSCI 699: Trustworthy Large Foundation Models: on Bias, Privacy and Safety Issues
Units: 4
Term—Day—Time: [Spring 2025] — [3hrs once per week]

Location: TBD

Instructor: Jieyu Zhao
Office: PHE 332
Office Hours: TBD
Contact Info: jieyuz@usc.edu I will reply to the email in 48 hours. Please make sure to include "CSCI 699" in the email subject.

Teaching Assistant: TBD
Office: TBD
Office Hours: TBD
Contact Info: TBD

Catalogue Description
Improving trustworthiness in large AI models, especially focuses on large pre-trained models, such as large language models (LLMs) and vision-and-language models (VLMs).

Course Description
Although there have been impressive advancements in large pretrained models (large language models or multimodal models), several studies reported that those system contain social biases. Even worse, the models run the risk of further amplifying the stereotypes and causing harms to people. As AI models, especially the large pretrained models continues to advance and be integrated into various domains such as healthcare, finance, marketing, and social media, it raises important ethical concerns that need to be addressed. In this course, students will critically examine the ethical implications of AI systems, including issues related to bias, fairness, privacy, transparency, accountability, and social impact. Through discussions, case studies, and guest lectures, students will explore the ethical challenges associated with AI models and develop a deep understanding of the ethical considerations that arise when designing, implementing, and deploying AI applications.

Learning Objectives
Students will get a broad understanding about possible issues in current large pretrained models and how current research has tried to alleviate those issues. This class will equip students with the ability to read and write critical reviews about research papers. At the same time, they will learn how to conduct research related to AI fairness, interpretability and robustness.

Prerequisite(s): N/A
Co-Requisite(s): N/A
Concurrent Enrollment: N/A
Recommended Preparation:

- Familiarity with natural language processing and/or machine learning. Ideal pre/co-requisites are CSCI 544 (Applied Natural Language Processing) or CSCI 567 (Machine Learning).
- Programming skills. We will mainly use python with PyTorch, but you can use any other libraries for your final project.

Course Notes

Grading type: Letter of Credit/No-Credit. Lecture slides will be posted online after the class.

Technological Proficiency and Hardware/Software Required

For the course project, access to computational resources (e.g., GPU) is highly recommended.

Required Readings and Supplementary Materials

All reading materials will be posted on the course website at the beginning of the course.

Supplementary materials

The following courses are relevant:

- UW: Linguistics 575: Ethics in NLP: http://faculty.washington.edu/ebender/2017_575/
- Berkeley: CS 294: Fairness in Machine Learning: https://fairmlclass.github.io/
- CMU: Computational Ethics for NLP: http://demo.clab.cs.cmu.edu/ethical_nlp2019/

Optional Readings and Supplementary Materials

N/A

Description of Assignments and How They Will Be Assessed

1. **Course Project (60%)**

Each student needs to individually finish one research project related to the class topics. There should be a "deliverable" result out of the project, meaning that your project should be self-complete and reproducible (scientifically correct. A typical successful project could be: 1) a novel and sound solution to an interesting research problem, 2) correct and meaningful comparisons among baselines and existing approaches, 3) applying existing techniques to a new application. We will not penalize negative results, as long as your proposed approach is thoroughly explored and justified. Overall, the project should showcase the student's ability to think critically, conduct rigorous research, and apply the concepts learned in the course to address a relevant problem in the field of AI ethics.
Students should use the standard *ACL paper submission template to finish their writing report regarding the course project.

- **Project proposal (10%)**
  Students are expected to finish a 2-page long project proposal by Week 5. The proposal should articulate the research question, justify the significance of the research, and provide evidence of the student's knowledge and understanding of the research literature. A timeline for the project will be highly recommended to be included in the proposal.
- **Midterm progress report (10%)**
  By Week 10, the students should finish a ~3-page progress report. The report should provide a clear statement about the research goal (could be different from the original one), a concise overview of the work completed so far, including any challenges encountered and solutions implemented and a report of some initial results.
- **Final presentation (20%)**
  During the last two week, the students will make a 30-minute presentation about their project. It should include the research goal, the motivation, related work, their methodology and results. There will be a 5-minute QA session for other students to ask questions.

- **Final project report (20%)**
  Students will write a final project report to describe the details about their research. The report should follow the NLP conference paper format, including the abstract, introduction, related work, result demonstration and discussion section. If the result is negative, it won't be penalized but the students should highlight their analysis about what could be the possible reasons. The report should be in total 8 pages (excluding the reference).

2. Paper Presentation (30%)

- Paper presentation will help students to develop the skills to give research talk to others.
- Each student will present 2 papers to the class. The student will prepare the slides for the paper and lead the discussion.
- Each week, there will be another student signed up as the feedback provider (reviewer). The presenter should finish the rehearsal of their talk with the reviewer. The reviewer will finish a feedback form at least 2 days before the class.
- Reviewer will provide the feedback to the presenter. The instructor will grade the presentation. Grading rubrics: correctness of the content (40%), clarity (20%), discussion (20%), slides & presentation skills (20%).

Participation (10%)

Students are expected to attend the class and get involved in paper discussion. This includes asking questions about the presentations or express their opinions on the topics.

Grading Breakdown

- Grading policy:
  - 60% Course Project
  - 30% Paper Presentation
  - 10% Attendance and discussion participation

Assignment	**% of Grade**
Participation	10
Paper presentation	30
Project proposal	10
Project midterm progress report	10
Project final presentation	20
Project final report	20
TOTAL	100

Grading Scale

Course final grades will be determined using the following scale

A	95-100
A-	90-94
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
D+	67-69
D	63-66
D-	60-62
F	59 and below

Assignment Submission Policy

Assignment will be submitted to google drive by 11:59 pm on the due date.

Course-Specific Policies

Students will have in total **4 late days** to use for the project proposal and progress report (no late days for the final report). The grace period will be used in integer amounts. Additional late days will result in a deduction of 10% of the grade on the corresponding assignment per day.

Attendance

No portion of the grade may be awarded for class attendance, but non-attendance can be the basis for lowering the grade. Exception can be made for valid reasons, such as conference travel or physician's notes.

Academic Integrity for this Class

It is extremely important to stick to the academic integrity during this class. All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s). A violation on the academic integrity will cause you fail this course and more consequences as described under " **Academic Integrity**" Section.

Please ask the instructor [and/or TA(s)] if you are unsure about what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Class Recordings and Course Content Distribution: You may not record this class without the express permission of the instructor and all other students in the class. Distribution of any notes, recordings, exams, or other materials from a university class or lectures — other than for individual or class group study — is prohibited without the express permission of the instructor; violations will be considered an intentional act to facilitate or enable academic dishonesty and reported to the university.

Use of Generative AI in this Course

Generative AI is permitted but limited as follows: In this course, you are permitted to use artificial intelligence (AI)-powered programs to help you improve your writing the course project reports, but you cannot use AI models to directly generate anything for you. In addition:

- You should also be aware that AI text generation tools may present incorrect information, biased responses, and incomplete analyses; thus, their answers may not meet the standards of this course.
- To adhere to our university values, *you must cite any AI-generated material (e.g., text, images, and other content) included or referenced in your work and provide the prompts used to generate the content*. Using an AI tool to generate content without proper attribution will be treated as plagiarism and reported to the Office of Academic Integrity.

Please review the instructions in each assignment for more details on how and when to use AI Generators for your submissions.

Course Evaluations

Course evaluation occurs at the end of the semester university-wide. We will do mid-semester evaluation and final course evaluation in class.

Course Schedule

	Topics/Daily Activities	Readings/Preparation	Deliverables
Week 1	Course introduction; paper candidate list discussion; review about how to do research presentation	Presentation signup	By W2
Week 2	Philosophical Foundations		
Week 3	AI Fairness Overview		
Week 4	Foundations about Large Pretrained Models (LLM, VLM)		
Week 5	Bias in LLM & VLM		Project proposal due by Friday 11:59 pm
Week 6	Privacy and Security in LLMs		
Week 7	Civility & Toxicity		
Week 8	Misinformation & Manipulation		
Week 9	Distributional Robustness in Large Pretrained Models		
Week 10	AI and Society		Project midterm progress report due by Friday 11:59pm
Week 11	Model Explanation		
Week 12	AI for social good		
Week 13	Human-Centered AI		
Week 14	Final project presentation		
Week 15	Final project presentation		
FINAL	Final report		Wednesday, May 7 2025

Academic Integrity

The University of Southern California is foremost a learning community committed to fostering successful scholars and researchers dedicated to the pursuit of knowledge and the transmission of ideas. Academic misconduct — which includes any act of dishonesty in the production or submission of academic work (either in draft or final form) — is in contrast to the university's mission to educate students through a broad array of academic, professional, and extracurricular programs.

This course will follow the expectations for academic integrity as stated in the USC Student Handbook. All students are expected to submit assignments that are their own original work and prepared specifically for this course and section in this academic term. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s). Students suspected of engaging in academic misconduct will be reported to the Office of Academic Integrity.

Other violations of academic misconduct include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

Academic dishonesty has a far-reaching impact and is considered a serious offense against the university. Violations will result in a grade penalty, such as a failing grade on the assignment or in the course, and disciplinary action from the university itself, such as suspension or even expulsion.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment or what information requires citation and/or attribution.

Course Content Distribution and Synchronous Session Recordings Policies

USC has policies that prohibit recording and distribution of any synchronous and asynchronous course content outside of the learning environment.

Recording a university class without the express permission of the instructor and announcement to the class, or unless conducted pursuant to an Office of Student Accessibility Services (OSAS) accommodation. Recording can inhibit free discussion in the future, and thus infringe on the academic freedom of other students as well as the instructor. (Living our Unifying Values: The USC Student Handbook, page 13).

Distribution or use of notes, recordings, exams, or other intellectual property, based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study. This includes but is not limited to providing materials for distribution by services publishing course materials. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relation to the class, whether obtained in class, via email, on the internet, or via any other media. Distributing course material without the instructor's permission will be presumed to be an intentional act to facilitate or enable academic dishonestly and is strictly prohibited. (Living our Unifying Values: The USC Student Handbook, page 13).

Statement on University Academic and Support Systems

Students and Disability Accommodations:
USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Student Financial Aid and Satisfactory Academic Progress:
To be eligible for certain kinds of financial aid, students are required to maintain Satisfactory Academic Progress (SAP) toward their degree objectives. Visit the Financial Aid Office webpage for undergraduate- and graduate-level SAP eligibility requirements and the appeals process.

Support Systems:
Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline consists of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-2500
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health-promoting habits and routines that enhance quality of life and academic performance.