CSCE 581: Introduction to Trusted AI

Lectures 27-28-29:
Class Project, Grad Paper Presentation

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

22, 24 AND 29 APRIL, 2025

Carolinian Creed: "I will practice personal and academic integrity."

Credits: **Copyrights of all material reused acknowledged**

Organization of Lectures 27, 28, 29

- Introduction Section
  - Recap from Week 13 (Lectures 25 and 26)
  - Announcements and News
- Main Section
  - L27: Project presentation
  - L28: Project report writing time – no class
  - L29: (Graduate) paper presentation
- Concluding Section
  - About next week – Final/ May 6 - Lectures 30
  - Ask me anything

Recap from Week 13 (Lectures 25, 26)

- We looked at
  - L25: Human-AI Collaboration, Chatbots
  - L26: Emerging AI Trust Landscape - Standards, Privacy

Announcements

- Marks of both quizzes posted
- Marks of project (Check points 1 and 2 posted)

Announcement: Change to Student Assessment

A	=	[920-1000]
B+	=	[870-919]
B	=	[820-869]
C+	=	[770-819]
C	=	[720-769]
D+	=	[670-719]
D	=	[600-669]
F	=	[0-599]

Tests	Undergrad	Grad
Course Project – report, in-class presentation	600	600
Quiz – 2 quizzes	200	200
Final Exam	200	100
Additional Final Exam – Paper summary, in-class presentation		100
Total	1000 points	1000 points

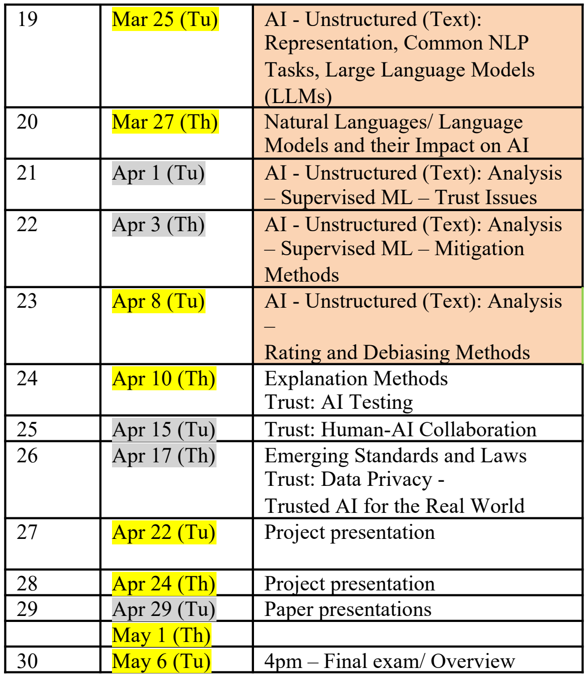
Change: 4 quizzes to 2; no best of 3

Project Discussion

Project Status and Timeline

- Office Hours: 3-4pm (M), 10-11am (Th)
- Finish project presentations by Apr 22
- Project presentations
  - Apr 22 (Tu) Project presentation
  - Apr 24 (Th) Project report writing
- Project delivered

  Apr 29 (Tu) Project in Github



19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods
24	Apr 10 (Th)	Explanation Methods Trust: AI Testing
25	Apr 15 (Tu)	Trust: Human-AI Collaboration
26	Apr 17 (Th)	Emerging Standards and Laws Trust: Data Privacy - Trusted AI for the Real World
27	Apr 22 (Tu)	Project presentation
28	Apr 24 (Th)	Project presentation
29	Apr 29 (Tu)	Paper presentations
	May 1 (Th)	
30	May 6 (Tu)	4pm – Final exam/ Overview

Course Project

- **Framework**
  1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
  2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
  3. (AI Method) Think of what the solution will do to help the primary user
     1. Solution => ML task (e.g. classification), recommendation, text summarization, …
     2. Use a foundation model (e.g., LLM-based) solution as the baseline
  4. (Data) Explore the data for a solution to work
  5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will works? (e.g., 20% reduction in patient deaths)
  6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
  7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Project Discussion: What to Focus on ?

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
  - Due to Users
    - Diverse demographics
    - Diverse abilities
    - Multiple human languages
  - Or other impacts
- What one does to mitigate trust issue

Rubric for Evaluation of Course Project

Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

<Project Title> - <Your Name>

Format for Interim Presentation on April 22, 2025

Project Context

1. Title:
2. Key idea: (2-3 lines)
3. Who will care when done:
4. Data need:
5. Methods:
6. Evaluation:
7. Users:
8. Trust issue:

- Test Case – demonstrate working
  - E.g., <input, output, correct output – if different, trust observation>

1 min context, 2 min demo, 1 min expts, 1 min Q/A

<Project Title> - <Your Name>

Format for Interim Presentation on April 22, 2025

Demonstrate effectiveness/ efficiency

- Metrics (F1, running time, …)
- Empirical results
- Comparison with a LLM (why your method over a general alternative)

Conclusion

- Experience
- Q/A

1 min context, 2 min demo, 1 min expts, 1 min Q/A

Project Report

- Due by Tuesday, April 29, 2025
- Will contain:
  - Project context
  - Demonstration, including trust aspect. Potentially link to a video.
  - Experimental results: effectiveness, efficiency dimensions
  - **Related work (what most relevant prior work is out there)**
  - **Discussion: experience and how it may be extended**
  - Conclusion

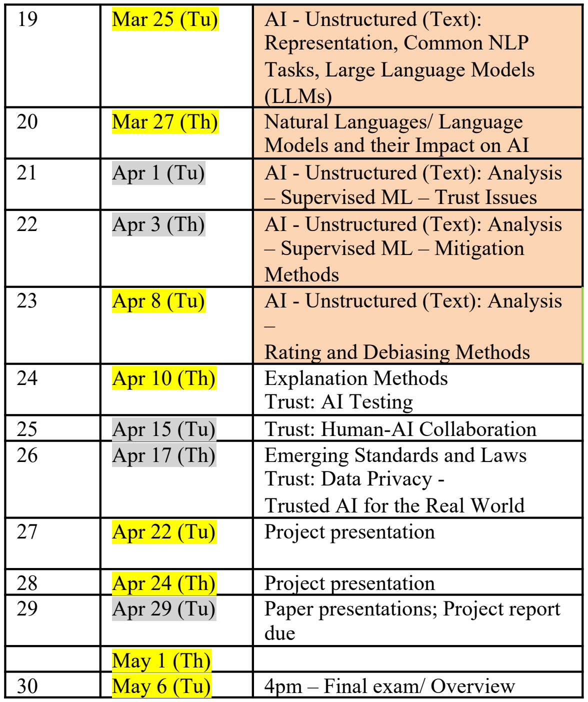
Concluding Section

Week 14 (L25 and 26): Concluding Comments

- We looked at project presentations
- Complete project report

About Next Week – Final Exam

Final Exam



19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods
24	Apr 10 (Th)	Explanation Methods Trust: AI Testing
25	Apr 15 (Tu)	Trust: Human-AI Collaboration
26	Apr 17 (Th)	Emerging Standards and Laws Trust: Data Privacy - Trusted AI for the Real World
27	Apr 22 (Tu)	Project presentation
28	Apr 24 (Th)	Project presentation
29	Apr 29 (Tu)	Paper presentations; Project report due
	May 1 (Th)	
30	May 6 (Tu)	4pm – Final exam/ Overview