COMP 640 003: Graduate Seminar in Machine Learning

Course Information

Instructor: Hanjie Chen

Semester: Fall 2025

Location: KCK 105

Time: Thursday 4:00 PM - 5:15 PM

Email: hanjie@rice.edu

Office Hours: Thursday 3:00 PM - 4:00 PM, Duncan Hall 2081

Course Description

As Natural Language Processing (NLP) continues to advance, the complexity of models has grown exponentially, leading to powerful but often opaque “black-box” systems, like ChatGPT. While these models excel in various tasks, such as commonsense reasoning, machine translation, sentiment analysis, and question answering, their lack of interpretability poses significant challenges in critical applications like healthcare, legal systems, and finance, where understanding the decision-making process is paramount.

The goal of this seminar is to familiarize students with the emerging challenges in NLP and the advancements made in Explainable NLP. This seminar offers an in-depth exploration of the methods and techniques developed to interpret and explain the inner workings and decision-making processes of NLP models. It is designed to bridge the gap between high-performing models and the need for transparency and accountability in real-world applications. Students will be introduced to a wide range of interpretation and explanation techniques, from traditional rule-based approaches to cutting-edge methods involving prompting techniques and transformer understanding.

Key topics include but are not limited to: Introduction of NLP and Language Models, Interpretable and Rationalized Models, Feature Attribution Explanation (perturbation-based, gradient-based, attention-based), Multi-Level Explanation (phrase-level, hierarchical, concept-based), Data Attribution, Natural Language Explanation, Prompting-Based Techniques for Explainability, Human-Centered Explanation, Mechanistic Interpretability, and Explanation Evaluation. Students will gain cutting-edge knowledge through paper readings and hands-on experience through projects, where they will implement various interpretation methods in different NLP tasks.

Throughout this seminar, students will engage with theoretical concepts, practical implementations, and case studies that highlight the importance of interpretability in various NLP applications. The seminar will also address the ethical implications of deploying NLP models, emphasizing the role of interpretability in ensuring transparency, reducing bias, and fostering trust in AI systems. By the end of this seminar, students will not only understand how to apply interpretation techniques to NLP models but will also be prepared to contribute to the development of transparent, ethical, and reliable AI systems in their future studies and careers.

Course Objectives and Learning Outcomes

- **Understand the Importance of Model Interpretability**. Students will understand the critical role of interpretability in NLP and AI, particularly in sensitive applications such as healthcare, finance, and law. They will also recognize the ethical and social implications of deploying black-box models in real-world scenarios.
- **Explore a Wide Range of Interpretation and Explanation Techniques**. Students will learn and implement various interpretation and explanation methods, developed for a wide range of NLP models, from traditional neural language models to recent large language models (LLMs).
- **Critically Assess the Quality of Model Interpretations.** Students will evaluate the validity and reliability of model explanations using established metrics. They will also design and conduct user studies to measure human understanding and trust in model-generated explanations.
- **Apply Interpretability Techniques to Real-World NLP Applications.** Students will implement interpretation methods in practical scenarios, addressing specific challenges such as fact-checking and error analysis. They will develop solutions for enhancing the interpretability of complex models (e.g., LLMs) in real-world applications such as healthcare and finance.

Course Format

- **Introduction Session.** The instructor will give the first lecture, providing an overview of the course, including an introduction to NLP, language models, and the importance of model interpretability and explainability.
- **Weekly Topics and Readings.** Each week we will focus on a specific topic in Explainable NLP. Students will be assigned 2 papers related to the weekly topic to read before class. They are expected to formulate 2-3 discussion questions related to each paper or the topic of the week.
- **Student Presentations.** In each class, students will give one short presentation (15-20 minutes) on one of the assigned papers. Each student will be responsible for 2-3 presentations on different weeks.

- **Discussion Sessions.** After each presentation, there will be a discussion session (15-20 minutes) focused on the presented paper and the broader topic. Discussions will address both the technical aspects of NLP and the social implications of interpretability.

This seminar is offered for 1 credit hour by default. Students interested in a 3-credit option can undertake a course project, which requires additional work. For those enrolled in the 3-credit option, there will be an additional class session dedicated to **final project presentations**.

Grading Policy

- **1-Credit Option**
  - Paper reading and discussion questions: 11 * 3% = 33%
  - Paper presentation: 2 * 30% (or 3 * 20%) = 60%
  - Attendance and active participation in paper discussions: 7%
- **3-Credit Option**
  - Paper reading and discussion questions: 11 * 1% = 11%
  - Paper presentation: 2 * 30% (or 3 * 20%) = 60%
  - Attendance and active participation in paper discussions: 7%
  - Final project: 22%
- **Rubrics**
  - Paper Reading. Read the two assigned papers each week and submit 2-3 discussion questions related to each paper to Canvas before the class.
  - Paper Presentations. Each student is required to give 2-3 presentations on assigned papers, depending on the total number of students in the class. Presentations will be evaluated based on clarity, understanding of the paper, and the ability to engage the class in discussion.
  - Attendance and Active Participation. Students must attend at least 10 classes during the semester. Occasional absences are permitted, but students should notify the instructor in advance.
  - Course Project and Final Presentation (3-Credit Hour Option). For students enrolled in the 3-credit hour option, completing a class project is required. The project involves applying interpretability techniques to a chosen NLP task. It will be assessed based on the thoroughness of research, quality of implementation, and effectiveness of the final presentation. Students are encouraged to contact the instructor at the beginning of the course to discuss their project ideas. Additionally, students should schedule regular appointments with the instructor or attend office hours throughout the semester to provide updates on their progress.

The letter grade will be assigned based on the points accumulated:

Grade	A+	A	A-	B+	B	B-	
Points	98-100	94-97	90-93	87-89	84-86	80-83	
Grade	C+	C	C-	D+	D	D-	F

Points	77-79	74-76	70-73	67-69	64-66	60-63	<60

Prerequisites

Students are expected to have completed at least one machine learning course and possess basic knowledge of NLP, including fundamental concepts and common tasks (such as sentiment analysis and question answering). For students opting for the 3-credit hour option, which includes hands-on projects, prior programming experience, preferably in Python, and familiarity with libraries such as NumPy, pandas, and scikit-learn, are required.

Course Policies

The members of our community at Rice come from many different backgrounds and views. Our goal is to ensure that everyone feels safe, respected, and empowered to be their best selves. We kindly ask our students to treat each other with care and respect.

Rice Honor Code

All students are expected to adhere to the standards of the Rice Honor Code, which you agreed to uphold upon matriculation. For detailed information on the Honor Code, including its administration and the procedures for addressing alleged violations, please refer to the Honor System Handbook available at http://honor.rice.edu/honor-system-handbook/. This handbook outlines the University's expectations for academic integrity, the procedures for resolving any alleged violations, and the rights and responsibilities of students and faculty throughout the process.

Disability Resource Center

If you have a documented disability or other condition that may affect academic performance you should: 1) make sure this documentation is on file with the Disability Resource Center (Allen Center, Room 111 / adarice@rice.edu / x5841) to determine the accommodations you need; and 2) talk with me to discuss your accommodation needs in the first two weeks of class.

Course Schedule

Week	Date	Topic
1	08/28/2025	Course Logistics
2	09/04/2025	Introduction to Classic Explanation Methods
3	09/11/2025	Feature Attribution Explanation: Attention-based Methods

4	09/18/2025	Multi-Level Explanation
5	09/25/2025	Interpretable and Rationalized Models
6	10/02/2025	Data Attribution
7	10/09/2025	Natural Language Explanation
8	10/16/2025	Prompting-Based Techniques for Explainability
9	10/23/2025	Human-Centered Explanation
10	10/30/2025	Mechanistic Interpretability: Neurons, Circuits, Concepts
11	11/06/2025	Mechanistic Interpretability: Probing, Patching
12	11/13/2025	Explanation Evaluation
13	11/21/2025	Explanation Utility
14	11/27/2025	THANKSGIVING RECESS
15	12/04/2025	Final Project Presentation