CS 221M Final Project Description

Overview

To demonstrate understanding of fundamental concepts and gain appreciation for recent developments in mechanistic interpretability, students will **implement and present results from a published research paper** in the field.

The goal of the final project is *not* to replicate an entire research paper – most interpretability projects have a GitHub repository for this exact purpose. Instead, the goal is to create a **walk-through explanation** of the motivation, methods, and analysis behind a key experiment. Students can present their implementation as either (1) a Jupyter notebook containing interleaved code (implementation) and text (explanation), or (2) a GitHub repository (implementation) accompanied by a one-page report (explanation).

Details

Students are welcome to choose any paper in mechanistic interpretability published within the past five years. Please reach out to the course staff in advance if you're unsure whether a paper is in mechanistic interpretability or isn't a recent publication.

We will **assign each project group a mentor** – a PhD or industry researcher in mechanistic interpretability – who will advise their project development and presentation.

Papers | Any of the readings listed on the course website are excellent choices for the final project. The course staff will create a post on the Ed Forum with an additional list of recommended papers. Topics include: *data attribution*, *causal abstraction*, *unsupervised interpretability*, *representation geometry*, *training dynamics*, *cognitive interpretability*.

Examples | We encourage looking at the NNSight mini paper tutorials for examples of walkthrough implementations; any of these tutorials would receive full credit for the final project.

- Geometry of Truth
- Dual Route Induction
- Demystifying Memorization

Grading

The final project will be graded for **completion** (did you replicate a key figure?) with a focus on **readability** (could a student from the course follow your code and explanation to understand the gist of the paper?). Should students choose to conduct an independent research project in addition to the replication, the project will be graded not on completeness but on **clarity of presentation** (could other students in the course understand the goal of the project and the current results?).

Timeline

The final projects will follow the timeline below.

Week 5: Students choose paper to implement
Students choose their final project group, and submit their choice of topic and paper. We will use this submission to pair students with a project mentor.

Week 6: Staff assigns project mentors
Project mentors will reach out to students and set up expectations for the project.

Week 7: In-class project check-in
We will dedicate a day in class to receive peer feedback on the final project. Students will be expected to have code that recreates a single figure from their chosen paper. During class, students will give each other feedback on their projects.

Week 10: Final project presentations
We will hold a poster presentation where students will present their work. Instructors will walk around to ask questions about the student's projects.

End of week 10: Final project DUE
Submit a one- to two-page report providing a clear and detailed explanation of the paper implementation that other students who took CS 221M would be able to follow and use to learn about the paper you chose.

Questions

Can students work in groups?
Students are **strongly encouraged** to work in **groups of 2-3**. We will assign all groups to a research mentor – however, we cannot guarantee that individual students will have an assigned mentor, should they choose to complete the project independently.

How do I find a group?
We recommend looking for project groups through the Ed discussion forum. The staff can also help pair students if they reach out with specific research interests.