CS 221M Mechanistic Interpretability Syllabus

Course description

What is the internal structure of modern neural networks and how can we study it? This course provides a broad and deep introduction to interpretability, the subfield of machine learning concerned with understanding precisely how models process information and why they produce the outputs they do. We will cover topics such as probing, steering, causal abstraction, and sparse autoencoders, with a particular emphasis on causal methods and large language models. The course will include guest lectures from leading interpretability labs across academia and industry.

Course information

- Time: Monday, Wednesday 2:30pm-3:50pm
- Location: McMurtry Art Building, Oshman
- Office hours: By appointment - email course staff to schedule

Schedule

Please see the course website for an up-to-date schedule:
cs221m.stanford.edu

Grading

- 25% Participation
  Students are expected to attend all lectures and engage with the course materials; please notify staff if you are unable to attend lectures in advance. Students who regularly attend lectures and inform staff about absences will receive full participation credit.
- 75% Final project
  Students will present a paper implementation; see description below.

Final project

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 figure of the paper. 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).

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?). We will pair each group of students with a mentor – a PhD or industry researcher in mechanistic interpretability – who will advise their project development and presentation.

Policies

Can students work in groups?
Students are encouraged to work in groups of 2-3.

What if I miss a lecture?
Please inform the staff in advance of missing a lecture. For notebook-guided lectures, students should follow the exercises and submit the completed notebook on Canvas; for guest lectures, students will write a 200-word response to a research paper by the speaker.

Support

Students who may need academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE). Professional staff will evaluate the request with required documentation, recommend reasonable accommodations, and prepare an Accommodation Letter for faculty dated in the current quarter in which the request is being made. Students should contact the OAE as soon as possible since timely notice is needed to coordinate accommodations. The OAE is located at 563 Salvatierra Walk (phone: 723-1066, URL: http://oae.stanford.edu).

Stanford is committed to ensuring that all courses are financially accessible to its students. If you require assistance with the cost of course textbooks, supplies, materials and/or fees, you can contact the First Generation and/or Low-Income Student Success Center) to learn about the FLIbrary and other resources they have available for support.

Stanford offers several tutoring and coaching services:

- Academic Skills Coaching
- Tutoring program
- Hume Center for Writing and Speaking