CS 333: Natural Language Processing

Fall 2025

Prof. Carolyn Anderson

Wellesley College

Reminders

- Final project presentations are next Tuesday (2-3 minutes each!)
- **All work due by 12/18 at 4pm**
- Guest speaker this Friday: Jin Zhao (Brandeis)
- My next help hours: Thursday 4-5

Final Project Presentations

- Final project presentations are next Tuesday
- No more than 3 minutes per presentation (2 minutes is fine)
- You should have 1 slide
- Your slide should show an example of your task

Recap

Aligning Models with Human Preferences

Users tend to have preferences about generated text that go beyond its statistical frequency.

Companies also have preferences for how their models respond.

How can we align model behavior with these preferences?

Mark Dredze
@mdredze

Representation learning:
extract attention features and use as input features to another model

Pretraining:
learn good representations via an unlabeled task.

Finetuning:
train some more on in-domain data or separate labeled task

Prompt engineering:
craft prompts that disguise task of interest as a language generation problem.

Fine-tuning = Further training

Fine-tuning:
train some more on in-domain data or separate labeled task

Simple Fine-tuning

Just train on more (labeled) data

Proxy Tuning

Fine-tune a smaller model and use it steer a larger model

Reinforcement Learning

Tune the model with a reward function

Interpretability

Interpretability

Mechanistic interpretability is the study of how models make their decisions.

How can we figure out what is happening inside the model, given that contemporary models have *billions* of parameters?

Common Tools for Interpretability

Logit Lens: **apply the output layer** to an intermediate layer's output in order to understand what prediction the model would make at that point in its computation.

Common Tools for Interpretability

Logit Lens: apply the output layer to an intermediate layer's output in order to understand what prediction the model would make at that point in its computation.

Activation Patching: **replace** a specific set of the model's activations with some other vector to see if it changes the model's prediction.

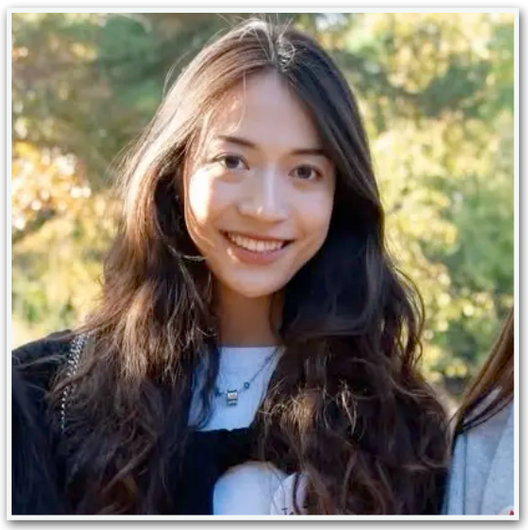
Common Tools for Interpretability

Logit Lens: apply the output layer to an intermediate layer's output in order to understand what prediction the model would make at that point in its computation.

Activation Patching: replace a specific set of the model's activations with some other vector to see if it changes the model's prediction.

Activation Steering: add a vector to the model's activations at a specific point to see if it changes its behavior.

CS333 Alums Working on Interpretability



Jennifer Meng Lu '24

Paths Not Taken: Understanding and Mending the Multilingual Factual Recall Pipeline

Meng Lu* Brown University meng_lu@brown.edu
Ruochen Zhang* Brown University ruochen_zhang@brown.edu
Carsten Eickhoff University of Tübingen carsten.eickhoff@uni-tuebingen.de
Ellie Pavlick Brown University ellie_pavlick@brown.edu

Abstract

Multilingual large language mode
often exhibit factual inconsistencies
derlying pipeline that LLMs empl
involves using the English-centric
call mechanism to process multilingu
and then translating English answers

EMNLP 2025

Pathway to Relevance: How Cross-Encoders Implement a Semantic Variant of BM25

Meng Lu* Brown University meng_lu@brown.edu
Catherine Chen* Brown University catherine_s_chen@brown.edu
Carsten Eickhoff University of Tübingen carsten.eickhoff@uni-tuebingen.de

Abstract

Mechanistic interpretation has greatly contributed to a more detailed understanding of generative language models, enabling significant progress in identifying structures that implement

principles. These models incorporate explicit components for semantic TF and IDF computations, blending neural architectures with established IR heuristics to improve relevance estimation.

However, the advent of transformer-based models revolutionized the field of IR. These models, trained end-to-end on large numbers of query-document pairs (Nguyen et al., 2016; Thakur et al., 2021), excel at extracting context-dependent semantic signals for ranking tasks. By leveraging multi-headed attention and vast parameter spaces

whether a document is relevant to a query. In

EMNLP 2025

Yik Siu Chan '24

CAN WE PREDICT ALIGNMENT BEFORE MODELS FINISH THINKING? TOWARDS MONITORING MISALIGNED REASONING MODELS

Yik Siu Chan* **Zheng-Xin Yo**
Department of Computer Scienc
Brown University

ICML 2025

NeurIPS 2025

SPEAK EASY: Eliciting Harmful Jailbreaks from LLMs with Simple Interactions

Reasoning
long chair
adversaria
for predic
of final re

significant efforts have
ai et al., 2022a;b; Gan-
023). However, these
"jailbreaks" (Jin et al.,
tacks that aim to breach
ce harmful responses,
d by malicious actors

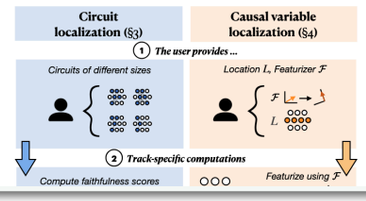
LLMs by non-technical
l insights into how jail-
interactions. Existing
a technical understand-
s (Zou et al., 2023) or
et al., 2023; Mehrotra
er, may not accurately
average user attempts
ses (NPR, 2025).

MIB: A Mechanistic Interpretability Benchmark

Aaron Mueller [*1] **Atticus Geiger** [*2] **Sarah Wiegreffe** [3]
Dana Arad [4] **Iván Arcuschin** [5] **Adam Belfki** [6] **Yik Siu Chan** [7] **Jaden Fiotto-Kaufman** [6] **Tal Haklay** [4]
Michael Hanna [8] **Jing Huang** [9] **Rohan Gupta** [10] **Yaniv Nikankin** [4] **Hadas Orgad** [4] **Nikhil Prakash** [6]
Anja Reusch [4] **Aruna Sankaranarayanan** [11] **Shun Shao** [12] **Alessandro Stolfo** [13] **Martin Tutek** [4] **Amir Zur** [2]
David Bau [6] **Yonatan Belinkov** [4]

Abstract

How can we know whether new mechanistic interpretability methods achieve real improvements? pro- nch- and sely s or causal variables in neural language models. The



ICML 2025

The Fragility of Model Alignment

Alignment is Fragile

Today we're going to use mechanistic interpretability methods to explore a key challenge with contemporary LLMs: model alignment is very fragile.

Our goal is to *jailbreak* a model: trick the model into telling us something that it has been fine-tuned not to reveal.

Activation Patching Demo