Mechanistic Interpretability Resource Guide

Getting Started

Mechanistic Interpretability Quickstart Guide [Neel Nanda, 2023]
Concrete Steps to Get Started in Transformer Mechanistic Interpretability [Neel Nanda, 2022]
A Comprehensive Mechanistic Interpretability Explainer and Glossary [Neel Nanda, 2022]

Explanatory Papers

Zoom In: An Introduction to Circuits (Olah, et.al. 2020)
Curve Detectors (Cammarata, et.al. 2020)
Toy Models of Superposition (Elhage, et.al. 2022)

Research Papers

List of Mechanistic Interpretability Papers [Neel Nanda]
Multimodal Neurons in Artificial Neural Networks (Goh, et.al. 2021)
Towards Monosemanticity: Decomposing Language Models with Dictionary Learning (Bricken, et.al. 2023)
Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet (Templeton and Conerly, et.al. 2024)
Scaling and Evaluating Sparse Autoencoders (Gao, et.al. 2024)

Code

OpenAI Sparse Autoencoders GitHub (sparse autoencoders trained on gpt2-small)
Transformer Interpretability short course

Videos

Induction Head Circuits [lecture by Chris Olah]
Mechanistic Interpretability & Mathematics [lecture by Neel Nanda]

Future Looking

Interpretability Dreams (Olah, et.al. 2023)
200 Concrete Open Problems in Mechanistic Interpretability [Neel Nanda, 2022]
Open Questions (from Toy Models of Superposition paper)

Current Themes in MI Research (article, 2022)