


Ronak Mehta

<https://ronakrm.github.io>
🎓 Google Scholar  LinkedIn
ronak@coordinal.org

Education

Computer Sciences, PhD

2016 - 2022

University of Wisconsin-Madison

Machine Learning and Computer Vision Research

Thesis: Identifying Feature, Parameter, and Sample Subsets in Machine Learning and Image Analysis

Minor in Statistics

Computer Sciences, MS

2014 - 2016

University of Wisconsin-Madison

Selected Coursework: Statistical Machine Learning, Computational Statistics, Nonconvex Optimization

Computer Engineering, B.S.E.

2010 - 2014

University of Michigan-Ann Arbor

Experience

Coordinal Research. coordinal.org

San Francisco, CA

Co-Founder

Present

- Accelerating the research of safe and aligned AI systems
- Developing tools that accelerate the rate human researchers can make progress on alignment.
- Building automated research systems that can assist in alignment work today.

Machine Learning and Theory Scholars Program

Berkeley, CA

Research Scholar

June 2024 - December 2024

- Worked on theoretical and practical solutions for identifying and accounting for worst-case model behaviors.
- Built out mechanisms for finetuning efficient bounds on model performance based on model internals.
- Applied classical optimization schemes such as Lipschitz optimization and mirror descent to find jailbreaks and identify regions in the input sequence space that may exhibit outlier behaviors.

Orca DB, Inc.

Boston, MA

Member of Technical Staff

September 2023 - September 2024

- Founding scientist and engineer building out core ML business solutions and models enabling direct control and interpretability via memory inspection and editing.
- Worked on memory augmentation for machine learning models ranging from large language models to simpler classifiers and regression models for non-generative use cases.

Redwood Research

Berkeley, CA

REMIX Research Resident

January 2023

- Participated in research program on mechanistic interpretability for large language models.
- Worked on grounding topical mechanistic interpretability methods in theoretical foundations from mainstream machine learning research, connecting ideas in interpretability hypothesis testing to classical probabilistic measures of conditional independence.

Skills

Model Experience: Finetuning local LLMs, CNNs (U-Nets, Flow-based methods), Bayesian Methods, Neural Architecture Search, Mixed Effects Regression, Kernel SVMs

ML/Scientific Tools: Transformers, PyTorch, Tensorflow, Scikit-Learn, Lme4, GGPlot, Pandas/NumPy/SciPy

Programming Languages: Python, R, C++, MATLAB, Julia, HTML/JavaScript