

# Mimansa Jaiswal

Research Scientist

[mimansa.jaiswal@gmail.com](mailto:mimansa.jaiswal@gmail.com) ♦ (734) 747-0283 ♦ [mimansajaiswal.github.io/](https://mimansajaiswal.github.io/) ♦ [twitter.com/MimansaJ](https://twitter.com/MimansaJ) ♦ [www.linkedin.com/in/mimansajaiswal/](https://www.linkedin.com/in/mimansajaiswal/) ♦ <mailto:mimansa@umich.edu>

## INTERESTS

---

I am primarily interested in \*LLM post-training and evaluation\*, focusing on understanding why models fail or behave unexpectedly \_\_(Failure Analysis)\_\_, improving them through techniques such as RLHF/RLVR/RLAIF, DPO/GRPO, and reward modeling \_\_(Post-Training)\_\_, developing automated and synthetic evaluation systems to measure fine-grained capabilities and robustness \_\_(Evaluation & Data Augmentation)\_\_, integrating human knowledge and preferences into model development \_\_(Human-in-the-Loop ML)\_\_, and enhancing interpretability and deployment readiness \_\_(Model Explainability & Deployment)\_\_.

## EXPERIENCE

---

### Meta (MSL Org)

*Research Scientist*

Feb 2025 – Present

- Part of Meta AI's Capabilities team, developed and fine-tuned large language models (LLMs) using advanced post-training techniques including RLHF/RLVR/RLAIF (with and without checklists), and used DPO/GRPO with reward models for improving production models.
- Built end-to-end synthetic data generation pipelines and automated evaluation systems, including rubric-based judges, preference data collection workflows, and quality assessment frameworks for model improvement and hill-climbing optimization.

### Norm AI

*Research Engineer*

Oct 2023 – Oct 2024

- Developed domain-specific retrieval-augmented generation (RAG) systems for legal applications, implementing synthetic data augmentation techniques to expand training datasets, create evaluation datasets and improve model coverage.
- Built agentic LLM simulation frameworks to model human group perception and consensus-building for subjective law interpretation and policy evaluation.

### Allen NLP, Allen Institute for AI

*Research Intern*

Sep 2021 – Dec 2021

- Developed interpretable and compositional evaluation frameworks for GPT-3 and other language models, creating benchmarks for multi-aspect quality assessment.
- Used the designed evaluation methodologies to measure fine-grained model capabilities and failure modes in natural language understanding tasks.

### NLP Team, Facebook AI Research (FAIR)

*Research Intern*

May 2021 – Aug 2021

- Identified and analyzed systematic failure patterns in Natural Language Inference (NLI) through adversarial testing and error analysis.
- Developed correction strategies and data augmentation techniques to improve model robustness on challenging NLI examples performing experimentation on CTRL model.

## Conversation AI, Facebook Research

*Research Intern*

May 2020 – Aug 2020

- Developed interpretable user satisfaction metrics for conversational AI systems by integrating human knowledge and behavioral signals, and trained weakly supervised hierarchical label models.
- Created generalizable evaluation frameworks combining quantitative metrics with qualitative human feedback for dialogue quality assessment.

## **EDUCATION**

---

### University of Michigan

Ann Arbor, MI

*Ph.D. in Computer Science and Engineering*

Sep 2017 – Aug 2023

- **Honors:** Awarded Barbour Fellowship Amongst Over 1k+ Applicants
- Computational Human Analysis and Integration (CHAI) Lab
- \*Advisor\*: Prof. Emily Provost

### University of Michigan

Ann Arbor, MI

*M.S. in Computer Science and Engineering*

Sep 2017 – May 2019

- **Honors:** GPA: 3.87/4.00
- **Coursework:** Natural Language Processing, Advanced Artificial Intelligence, Bayesian Inference

### Institute of Engineering and Technology

Indore, India

*B.E. in Computer Science Engineering*

Jul 2013 – May 2017

- **Honors:** GPA: 3.7/4.00
- **Coursework:** Information Retrieval, Machine Learning, Algorithm Design and Analysis

## **PUBLICATIONS**

---

### CAPSTONE: Capability Assessment Protocol for Systematic Testing Of Natural language models' Expertise

Mimansa Jaiswal

*Status: Research Notes*

### Designing Interfaces for Delivering and Obtaining Generation Explanation Annotations

Mimansa Jaiswal

*Status: Demo and Repository*

### The Future of Open Human Feedback

Shachar Don-Yehiya, Ben Burtenshaw, Ramon Fernandez Astudillo, Cailean Osborne, Mimansa Jaiswal, Tzu-Sheng Kuo, Wenting Zhao, Idan Shenfeld, Andi Peng, Mikhail Yurochkin, Atoosa Kasirzadeh, Yangsibo Huang, Tatsunori Hashimoto, Yacine Jernite, Daniel Vila-Suero, Omri Abend, Jennifer Ding, Sara Hooker, Hannah Rose Kirk, Leshem Choshen

*Status: Submitted*

**Temperature Zero Isn't the Best Choice For Accuracy in Legal Reasoning**

Mimansa Jaiswal, Scott Worland, John Nay

*Status: Submitted*

**Lessons from the Trenches on Reproducible Evaluation of Language Models**

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, ...(and 12 others) Mimansa Jaiswal ...(and 10 others)

*Status: Submitted*

**Sally-Anne False Belief Test for LLMs**

Mimansa Jaiswal

*Status: Workshop Paper*

**Qualitatively Studying Gender Biases in LLMs**

Mimansa Jaiswal

*Status: Workshop Paper*

**Assessing Large Language Models: A Comprehensive Survey and Critical Analysis of Evaluation Metrics and Methodologies**

Mimansa Jaiswal

*Status: Survey Paper*

**From Text to Emotion: Unveiling the Emotion Annotation Capabilities of LLMs**

Minxue Niu, Mimansa Jaiswal, Emily Mower Provost

*Interspeech 2024* , Sep 2024

**Capturing Mismatch between Textual and Acoustic Emotion Expressions for Mood Identification in Bipolar Disorder**

Minxue Niu, Amrit Romana, Mimansa Jaiswal, Melvin McInnis, Emily Mower Provost

*Interspeech 2023* , Aug 2023

**Mind the gap: On the value of silence representations to lexical-based speech emotion recognition**

Matthew Perez, Mimansa Jaiswal, Minxue Niu, Cristina Gorrostieta, Matthew Roddy, Kye Taylor, Reza Lotfian, John Kane, Emily Mower Provost

*Interspeech 2022* , Sep 2022

**Human-Centered Metric Design to Promote Generalizable and Debiased Emotion Recognition**

Mimansa Jaiswal, Emily Mower Provost

*Text as Data (TADA) Conference 2021* , Oct 2021

**Noise Based Augmentation of Emotion Datasets: What's ideal and What Isn't?**

Mimansa Jaiswal, Emily Mower Provost

*ACL-SRW 2020* , Jul 2020

**MuSE: Multimodal Stressed Emotion Dataset**

Mimansa Jaiswal\*, CP Bara, Yuanhang Luo, Rada Mihalcea, Mihai Burzo, Emily Mower Provost

*Conference on Language Resources and Evaluation (LREC) 2020* , May 2020

**Privacy Enhanced Multimodal Neural Representations for Emotion Recognition**

Mimansa Jaiswal, Emily Mower Provost

*AAAI Conference on Artificial Intelligence (AAAI) 2020* , Feb 2020

**Controlling for Confounders in Multimodal Emotion Classification via Adversarial Learning**

Mimansa Jaiswal, Zakaria Aldeneh, Emily Mower Provost

*International Conference on Multimodal Interaction (ICMI) 2019* , Oct 2019

**Identifying Mood Episodes Using Dialogue Features from Clinical Interviews**

Zakaria Aldeneh, Mimansa Jaiswal, Michael Picheny, Melvin McInnis, Emily Mower Provost

*Interspeech 2019* , Sep 2019

**MuSE-ing on the Impact of Utterance Ordering On Crowdsourced Emotion Annotations**

Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, Emily Mower Provost

*IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2019* , May 2019

**The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild**

Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin McInnis, Emily Mower Provost

*Interspeech 2018* , Sep 2018

## HONORS AND AWARDS

---

### Barbour Fellowship

Ann Arbor, MI

Issued by *University of Michigan Rackham Graduate School*

Sep 2022

- Awarded amongst over 1,000+ applicants

### **Graduate Student Instructor for Affective Computing Course**

Ann Arbor, MI

Issued by *University of Michigan*

Jan 2020

### **Student Representative in CSE Faculty Hiring Committee**

Ann Arbor, MI

Issued by *University of Michigan CSE Department*

Sep 2020

### Invited Speaker at PyCon 2016

Singapore

Issued by *PyCon Singapore*

Jun 2016

### Invited Speaker at PyCon 2016

India

Issued by *PyCon India*

Oct 2016

### **National Talent Search Examination Scholarship**

India

Issued by *National Council of Educational Research and Training (NCERT)*

- Top 0.01% in India

## SKILLS

---

**Research Areas:** LLM Post-Training (RLHF/RLVR/RLAIF, DPO/GRPO), Model Evaluation & Benchmarking, Failure Analysis, Synthetic Data Generation, Human-in-the-Loop ML, Retrieval-Augmented Generation (RAG), Natural Language Processing, Multimodal Learning

**Technical Skills:** Python, PyTorch, Transformers, Machine Learning, Deep Learning, Statistical Analysis, Experimental Design

**Professional Activities:** Reviewing: ICDMW, ICMI, ACII, CHI, CSCW, \*ACL, AAAI, NeurIPS (2019-present), Public Speaking, Technical Writing, Mentoring

## LANGUAGES

---

English (Fluent), Hindi (Native)