

Neeraja Kiran Kirtane

Website | [LinkedIn](#) | [GitHub](#) | [Google Scholar](#) | Email: kirtane.neeraja@gmail.com

RESEARCH INTERESTS

Robustness and interpretability of large language models; hallucination detection and mitigation; Jailbreaking

EDUCATION

University of Illinois Urbana-Champaign	2023 – 2025
<i>Masters of Science in Computer Science (Thesis Track: Advisor- Prof. Hao Peng) (MSCS)</i>	CGPA: 4.0/4.0
Manipal Institute of Technology, Manipal, India	2018 – 2022
<i>B.Tech in Computer Science and Engineering (Minor: Computational Intelligence)</i>	CGPA: 9.14/10

EXPERIENCE

MathGPT.ai	Jul 2025 – Present
<i>AI/ML Research Engineer</i>	<i>Advisor – Peter Relan</i>
<ul style="list-style-type: none">• Stress-tested reasoning models (GPT-4, Claude, Qwen, DeepSeek) on 700+ linguistically varied math problems, uncovering systematic robustness failures.• Doing RL finetuning using GRPO coupled with curriculum learning, for Small language models (SLMs) for reasoning.	
Texas A&M University	Jul 2025 – Present
<i>Research Collaborator</i>	<i>Advisor – Prof. Kuan-Hao Huang</i>
<ul style="list-style-type: none">• Investigating whether reasoning in multilingual LLMs is language-agnostic using neuron-level interpretability techniques.• Applying activation analysis and attribution methods to identify cross-lingual reasoning patterns.	
University of Illinois Urbana-Champaign	Aug 2023 – May 2025
<i>Graduate Student Researcher</i>	<i>Advisor – Prof. Hao Peng</i>
<ul style="list-style-type: none">• Developed a hidden-state based classifier achieving >70% accuracy in preemptive hallucination detection (i.e. detection even before the generation of hallucinated output).• Designed intervention methods altering model activations, improving factuality upto 34% across 7B–70B LLMs (Llama, Mistral, Qwen, Gemma) in Wiki, Math, Medical, and GK domains.	
University of Illinois Urbana-Champaign	May 2024 – May 2025
<i>Graduate Student Researcher</i>	<i>Advisors – Prof. Hao Peng and Prof. Dilek Hakkani-Tur</i>
<ul style="list-style-type: none">• Explored jailbreaking strategies for LLMs to elicit biased content via persuasive scientific-language prompts with citations.• Demonstrated bias intensity escalation in multi-turn interactions across GPT-4, Claude, and Llama3-405B.	
Indian Institute of Technology Madras	Jul 2022 – Aug 2023
<i>Post Baccalaureate Fellow</i>	<i>Advisors – Prof. Balaraman Ravindran & Dr. Rajashree Baskaran</i>
<ul style="list-style-type: none">• Built knowledge graphs and graph-to-text generation pipelines for Hidden Voices, aiming to reduce the gender gap in Wikipedia biographies.• Fine-tuned GPT-J and GPT-Neo for table-to-text generation with DeepSpeed-based multi-GPU training.	
Indian Institute of Technology Madras	Jan 2022 – Jun 2022
<i>Research Intern</i>	<i>Advisors – Prof. Balaraman Ravindran & Dr. Ashish Tendulkar</i>
<ul style="list-style-type: none">• Proposed implicit algorithmic techniques to handle class imbalance in graph neural networks, using custom loss functions and attention weight tuning in graph attention networks.• Achieved +5% F1 score over then-SOTA on Cora and Citeseer datasets.	

PUBLICATIONS (* - EQUAL CONTRIBUTION)

1. Evaluation of Large Language Models' Robustness to Linguistic Variations in Mathematical Reasoning	
Under review Link	Oct 2025
<ul style="list-style-type: none">• Authors: <i>Neeraja Kirtane*</i>, <i>Yuvraj Khanna*</i>, <i>Peter Relan</i>	
2. FactCheckmate: Preemptively Detecting and Mitigating Hallucinations in LMs	
EMNLP 2025 Findings Link	Oct 2024
<ul style="list-style-type: none">• Authors: <i>Deema Alnuhait*</i>, <i>Neeraja Kirtane*</i>, <i>Muhammad Khalifa</i>, <i>Hao Peng</i>	
3. LLMs are Vulnerable to Malicious Prompts Disguised as Scientific Language	
Workshop on Socially Responsible Language Modelling Research (SoLaR) at COLM 2025 Link	Jan 2025
<ul style="list-style-type: none">• Authors: <i>Yubin Ge*</i>, <i>Neeraja Kirtane*</i>, <i>Hao Peng</i>, <i>Dilek Hakkani-Tur</i>	

4. Hidden Voices: Reducing gender data gap, one Wikipedia article at a time

Wikiworkshop 2023 [Link](#)

May 2023

- Authors: *Neeraja Kirtane, Anuraag Shankar, Chelsi Jain, Ganesh Katrapati, Raji Baskaran, Balaraman Ravindran*

5. ReGrAt: Regularization in graphs using attention mechanism to handle class imbalance

GCLR workshop at AAAI 2023 [Link](#)

Sep 2022

- Authors: *Neeraja Kirtane, Jeshuren Chelladurai, Balaraman Ravindran, Ashish Tendulkar*

6. Efficient Gender Debiasing of Pre-trained Indic Language Models

Deployable-AI workshop at AAAI 2023 [Link](#)

Aug 2022

- Authors: *Neeraja Kirtane, V Manushree, Aditya Kane*

7. Mitigating gender stereotypes in Hindi and Marathi

Gender bias in NLP workshop at NAACL 2022 [Link](#)

May 2022

- Authors: *Neeraja Kirtane, Tanvi Anand*

8. Transformer based ensemble for emotion detection

WASSA workshop at ACL 2022 [GitHub](#) | [Link](#)

Mar 2022

- Authors: *Aditya Kane, Shantanu Patankar, Sahil Khose, Neeraja Kirtane*

9. Occupational Gender Stereotypes in Indian Languages

Widening NLP workshop at EMNLP 2021 [Link](#) | [Video](#) | [Poster](#)

Nov 2021

- Authors: *Neeraja Kirtane, Tanvi Anand*

PROJECTS

Evaluating Mathematical Reasoning Chains [Github](#)

Advisor: Prof. Heng Ji

- Identified limitations of existing Chain-of-Thought (CoT) evaluation methods for math reasoning, which often overlook logical correctness and focus only on numerical accuracy.
- Developed a pretrained metric using Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) to evaluate nine aspects of reasoning chains.
- Achieved an average **8% improvement in correlation scores** over existing baselines for mathematical reasoning tasks.

LLMs for Privacy Policy Analysis

Advisor: Prof. Varun Chandrasekaran

- Applied the *Contextual Integrity* (CI) framework to classify information flows in privacy policies in LLMs.
- Designed a cost-efficient auto-tagging pipeline using LLaMA models and AI-driven data augmentation, achieving a **10% F1 score improvement** over the base model and comparable performance to GPT models.
- Long-term goal: formalize CI into first-order logic for longitudinal policy analysis.

TECHNICAL SKILLS AND RELEVANT COURSEWORK

Programming Languages: Python, C++, Java, C, SQL

ML/AI Frameworks: PyTorch, TensorFlow, NumPy, Hugging Face Transformers

Specialized Skills: Reinforcement Learning (GRPO, PPO, DPO), LLM interpretability (activation analysis, neuron attribution).

Courses: Introduction to Data mining, Advanced NLP, Advanced Topics in Security, Privacy, and Machine Learning, User-centered ML, LLMs Post Pretraining.

TEACHING EXPERIENCE AND EXTRACURRICULAR

- **Teaching Assistant, CS 105: Introduction to Computing** (Fall 2023, Spring 2024, Fall 2024, Spring 2025). Assisted in course delivery, grading, and student mentoring; awarded **Outstanding TA** for Spring 2024.
- Volunteer, **EMNLP 2021** and **NAACL 2022**. Supported conference logistics and session coordination.
- Managing Committee Member, **IEEE Student Branch**. Organized technical events and coordinated student outreach activities.