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
Masters of Science in Computer Science (Thesis Track: Advisor- Prof. Hao Peng) (MSCS) *CGPA: 4.0/4.0*

Manipal Institute of Technology, Manipal, India 2018 – 2022
B.Tech in Computer Science and Engineering (Minor: Computational Intelligence) *CGPA: 9.14/10*

Experience

MathGPT.ai Jul 2025 – Present
AI/ML Research Engineer *Advisor – Peter Relan*

- Stress-tested reasoning models (GPT-4, Claude, Qwen, DeepSeek) on 700+ linguistically varied math problems, uncovering systematic robustness failures.
- Engineering RLHF pipelines using GRPO coupled with curriculum learning, fine-tuning 2–7B models on reasoning datasets.

Texas A&M University Jul 2025 – Present
Research Collaborator *Advisor – Prof. Kuan-Hao Huang*

- 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
Graduate Student Researcher *Advisor – Prof. Hao Peng*

- 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
Graduate Student Researcher *Advisors – Prof. Hao Peng and Prof. Dilek Hakkani-Tur*

- 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
Post Baccalaureate Fellow *Advisors – Prof. Balaraman Ravindran & Dr. Rajashree Baskaran*

- 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
Research Intern *Advisors – Prof. Balaraman Ravindran & Dr. Ashish Tendulkar*

- 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.

Centre for Development of Advanced Computing (CDAC), Pune Jun 2020 – Aug 2020
ML Intern *Advisor – Sanjay Wandhekar*

- Developed a keyword and named-entity extraction system using GloVe embeddings, NLTK, and Gensim (LDA).
- Built and deployed a Flask-based web interface for interactive document analysis.

PUBLICATIONS

1. Evaluation of Large Language Models' Robustness to Linguistic Variations in Mathematical Reasoning
Under review — Oct 2025
- Authors: ***Neeraja Kirtane***, *Yuvraj Khanna, Peter Relan*

2. FactCheckmate: Preemptively Detecting and Mitigating Hallucinations in LMs
EMNLP 2025 Findings *Paper* — Oct 2024
- Authors: ***Neeraja Kirtane***, *Deema Alnuhait, Muhammad Khalifa, Hao Peng*

3. LLMs are Vulnerable to Malicious Prompts Disguised as Scientific Language
Workshop on Socially Responsible Language Modelling Research (SoLaR) at COLM 2025 *Paper* — Jan 2025
- Authors: ***Neeraja Kirtane***, *Yubin Ge, Hao Peng, Dilek Hakkani-Tur*

4. Hidden Voices: Reducing gender data gap, one Wikipedia article at a time
Wikiworkshop 2023 *Paper* — 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** *Paper* — 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** *Paper* — 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** *Paper* — May 2022
- Authors: ***Neeraja Kirtane***, *Tanvi Anand*

8. Transformer based ensemble for emotion detection
WASSA workshop at **ACL 2022** *GitHub* | *Paper* — Mar 2022
- Authors: *Aditya Kane, Shantanu Patankar, Sahil Khose,* ***Neeraja Kirtane***

8. Occupational Gender Stereotypes in Indian Languages
Widening NLP workshop at **EMNLP 2021** *Paper* | *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.