Neeraja Kiran Kirtane

Website | LinkedIn | GitHub | Google Scholar | Email: kirtane.neeraja@gmail.com

Research Interests

Robustness and Interpretability of Large Language Models; Hallucination Mitigation; Jailbreaking

Education

University of Illinois Urbana-Champaign 2023 – 2025
Master 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*

Research Experience

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

- Investigating whether language-specific and task-specific directions in multilingual LLMs can be causally isolated and manipulated; computing steering vectors via English-to-target-language activation differences across 9 languages using SAE and raw activation methods on various models for mathematical reasoning.
- Steering experiments across layers show middle layers achieve around 5% performance gains in target languages, demonstrating causal control over language-specific subspaces to improve low-resource language task performance.

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

- Designed and ran large-scale stress tests on GPT-4, Claude, Qwen, and DeepSeek using 500+ linguistically varied math problems, revealing systematic failures where models break under surface-level perturbations (variable swaps, paraphrasing, recontextualization). Numbers in the question and final answer remained the same.
- Building an education-centric benchmark spanning physics, chemistry, economics, sociology, and undergraduate-level quantitative reasoning tasks to evaluate models' real-world mathematical capability.

University of Illinois Urbana-Champaign - Alta Lab 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) for short-answer QA tasks.
- Trained an intervention model that modifies internal representations, improving factuality by up to 34% across Llama, Mistral, Qwen, and Gemma models in Wikipedia, Math, Medical, and STEM domains.

University of Illinois Urbana-Champaign - ConvAI Lab May 2024 – May 2025
Graduate Student Researcher *Advisors – Prof. Hao Peng and Prof. Dilek Hakkani-Tur*

- Investigated how authoritative scientific-language framing can jailbreak LLMs into producing stereotypically biased or harmful outputs.
- Used fabricated abstracts, credible-sounding citations, and famous author and venue names to test persuasion via "authority," which successfully produced biased and harmful outputs across frontier models.

Indian Institute of Technology Madras Jul 2022 – Aug 2023
Post Baccalaureate Fellow *Advisors – Prof. Balaraman Ravindran & Dr. Rajashree Baskaran*

- Scraped data of notable women in STEM, built knowledge graphs and built a graph-to-text generation model to generate a Wikipedia-style biography text. This was done for the project Hidden Voices, aiming to reduce the gender gap on Wikipedia.

Indian Institute of Technology Madras Jan 2022 – Jun 2022
Research Intern *Advisors – Prof. Balaraman Ravindran & Dr. Ashish Tendulkar*

- Proposed and tested implicit algorithmic techniques to handle class imbalance in graph neural networks by using custom loss functions and attention weight tuning in graph attention networks. The methods gave more importance to the weights of the minority nodes than the majority nodes.

Publications (* - equal contribution)

1. Evaluation of Large Language Models' Robustness to Linguistic Variations in Mathematical Reasoning
Preprint *Link*

- Authors: ***Neeraja Kirtane****, *Yuvraj Khanna**, *Peter Relan*

2. FactCheckmate: Preemptively Detecting and Mitigating Hallucinations in LMs
EMNLP 2025 Findings *Link*

- Authors: *Deema Alnuhait**, ***Neeraja Kirtane****, *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** *Link*

- Authors: *Yubin Ge*, **Neeraja Kirtane***, Hao Peng, Dilek Hakkani-Tur*

4. Hidden Voices: Reducing gender data gap, one Wikipedia article at a time
Wikiworkshop 2023 *Link*

- 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
Graphs and Complex structures for Learning and Reasoning (GCLR) workshop at **AAAI 2023** *Link*

- 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*

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

7. Mitigating gender stereotypes in Hindi and Marathi
Gender bias in NLP workshop at **NAACL 2022** *Link*

- Authors: ***Neeraja Kirtane**, Tanvi Anand*

8. Transformer based ensemble for emotion detection
WASSA workshop at **ACL 2022** *GitHub* | *Link*

- 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*

- Authors: ***Neeraja Kirtane**, Tanvi Anand*

Other 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 and 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.

Technical Skills And Relevant Coursework

Programming Languages: Python (expert), C++, Java, C, SQL.
ML/AI Frameworks: PyTorch, TensorFlow, Hugging Face Transformers, DeepSpeed.
Specialized Skills: Mechanistic interpretability (SAEs, activation steering, neuron attribution), RL Techniques (PPO, DPO, GRPO).
Relevant Courses: Advanced NLP, Advanced Topics in Privacy, and Machine Learning, User-centered ML, LLMs Post Pretraining.

Presentations and Talks

- Persuasion techniques for jailbreaking models. **TAMU NLP Group** [slides]
- Do LLMs ”know” internally when they follow instructions? **FLAIR Lab, TAMU**
- FactCheckmate: Preemptively Detecting and Mitigating Hallucinations in LLMs. **UIUC NLP Group** [slides]

Awards and Service

- Outstanding Teaching Assistant, CS 105: Introduction to Computing, UIUC (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.