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: 4.0/4.0*

Experience

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

- Investigating whether multilingual LLMs encode language-specific and task-specific directions/neurons, and whether these latent subspaces can be causally isolated.
- Steering models at different layers using these directions to test if targeted manipulations improve task performance in low-resource languages.

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).
- Building an education-centric benchmark spanning physics, chemistry, economics, sociology, and undergraduate-level quantitative reasoning tasks to evaluate models' real-world mathematical reliability.

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).
- Developed activation-based interventions that modify internal representations, improving factuality by up to **34%** across Llama, Mistral, Qwen, and Gemma models in Wikipedia, Math, Medical, and General Knowledge settings.

University of Illinois Urbana-Champaign 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 biased or harmful outputs.
- Used fabricated abstracts, credible-sounding citations, and venue names to test persuasion via "authority," showing consistent bias escalation across frontier models.

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. Did multi-GPU training to finetune models.

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. Improved performance on Cora and Citeseer datasets.

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
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*

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 (expert), C++, Java, C, SQL
ML/AI Frameworks: PyTorch, TensorFlow, Hugging Face Transformers, DeepSpeed
Specialized Skills: Mechanistic interpretability (SAEs, activation steering, neuron attribution), RLHF techniques (PPO, DPO, GRPO), Cross-lingual analysis
Relevant Courses: 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 conducting lab sessions; 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.