Neeraja Kiran Kirtane

(+1) 217-305-2081
kirtane3@illinois.edu
Website ◇ LinkedIn ◇ GitHub ◇ Google Scholar

Research Interests

Mitigation of Hallucinations in LLMs, Jailbreaking LLMs, Interpretability, Factuality, AI alignment

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*
- Working on testing the robustness of SOTA reasoning models.
- Finetuning small language models using RL-based methods like GRPO to use them for AI tutoring needs.

TAMU Jul 2025 - Present
Research Collaborator *Advisor – Prof. Kuan-Hao Huang*
- Working on understanding how reasoning happens in multilingual LLMs.
- Using various interpretability-based methods to understand if reasoning is language-agnostic or not.

University of Illinois, Urbana-Champaign Aug 2023- May 2025
Graduate Student Researcher *Advisor – Prof. Hao Peng*
- Working on detecting and mitigating hallucinations in large language models by analyzing **internal activations**.
- Independently built a lightweight classifier using hidden states of a model to preemptively detect hallucinations even before the model generates any output. Altered the hidden states using an intervention model to steer the generation towards factual outputs by 34%.
- Validated this across models of various sizes (7B to 70B), various families (Llama, Mistral, Qwen, Gemma) tested across four domains: Wiki, Math, Medical, General Knowledge.

University of Illinois, Urbana-Champaign May 2024- May 2025
Graduate Student Researcher *Advisors – Prof. Hao Peng and Prof. Dilek Hakkani-Tur*
- Working on **jailbreaking LLMs** to elicit stereotypical biased content by using various persuasion methods.
- Our findings show that feeding scientific summaries of papers along with citations supporting the claims of having benefits of bias leads to effective jailbreaking. The intensity of biased statement increases with every turn in the conversation.
- We test this across both large open and closed source models like GPT-4, Claude, Llama3-405B using the StereoSet data.

Indian Institute of Technology Madras, Chennai, India Jul 2022- Aug 2023
Post Baccalaureate Fellow *Advisors – Prof. Balaraman Ravindran & Dr. Rajashree Baskaran*
- Worked on the Project Hidden Voices which aims to **reduce the gender gap** in Wikipedia data by building intelligent tools.
- Worked on parsing relevant data, converting it to knowledge graphs, doing a graph to text generation.
- Finetuned models like GPT-J and GPT-Neo to do text generation on multiple GPUs. Used libraries like DeepSpeed to do model parallelism.

Indian Institute of Technology Madras, Chennai, India Jan 2022- Jun 2022
Research Intern *Advisors – Prof. Balaraman Ravindran & Dr. Ashish Tendulkar*
- Worked on handling class imbalance in Graph neural networks.
- Used **implicit ways** at the algorithmic level to handle this imbalance. Used a custom loss function and tuned the attention weights to focus more on minority nodes.
- Tested this on Cora and Citeseer datasets. Increased the F1 score by 5 percent than the SOTA method then.

Centre for development of advanced computing, CDAC Pune Jun 2020 – Aug 2020
ML Intern *Advisor – Rahul Dangi*
- Extracted keywords and named entities from a document for **better comprehension**.
- Used word embeddings of the GLoVe dataset for the predictions. Major libraries used in Python were NLTK (for text processing), Gensim (to use the LDA algorithm), Flask (to create the front end of the project).
- Deployed this pipeline and created a webpage for better user experience.

PUBLICATIONS

1. FactCheckmate: Preemptively Detecting and Mitigating Hallucinations in LMs
Under review *Paper* Oct 2024
- Authors: ***Neeraja Kirtane***, *Deema Alnuhait, Muhammad Khalifa, Hao Peng*

2. LLMs are Vulnerable to Malicious Prompts Disguised as Scientific Language
Under review *Paper* Jan 2025
- Authors: ***Neeraja Kirtane***, *Yubin Ge, Hao Peng, Dilek Hakkani-Tur*

3. 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*

4. 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*

5. 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*

6. Mitigating gender stereotypes in Hindi and Marathi
Gender bias in NLP workshop at **NAACL 2022** *Paper* May 2022
- Authors: ***Neeraja Kirtane***, *Tanvi Anand*

7. 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*
- Developed a pretrained metric to evaluate the chains generated by LLMs for Math reasoning tasks.
- Used SFT and DPO training for this process.
- The metric evaluated nine different characteristics of the chain.

Labelling privacy policies using LLMs *Advisor: Prof. Varun Chandrasekaran*
- Contextual integrity is a conceptual framework for understanding privacy expectations and their implications developed in the literature on law, public policy, etc.
- Fine-tuned Llama models for the task described above on the limited data that we had and additionally used AI based data expansion methods for augmenting data at low costs.
- Our results obtained from Llama models were comparable to GPT models and showed 10 percent increase in the F1 score than the base model.

TECHNICAL SKILLS AND RELEVANT COURSEWORK

Languages: Python, C++, Java, C, SQL.
Tools and Libraries: PyTorch, NumPy, TensorFlow.
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

- **TA for CS 105**: Introduction to Computing for Fall 2023, Spring 2024, Fall 2024, Spring 2025. Outstanding TA award for Spring 2024 semester.
- Volunteer at **EMNLP 2021**, **NAACL 2022**.
- Managing committee member at the IEEE student branch of the university.
- Mentored undergrads to get started with research.