Neeraja Kiran Kirtane

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

Research Interests

Detection and mitigation of Hallucinations in LLMs, Jailbreaking LLMs, Interpretability, Factuality.

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

University of Illinois, Urbana-Champaign Aug 2023- Present
Graduate Student Researcher *Advisor – Prof. Hao Peng*

- Working on detecting and mitigating hallucinations in large language models by analyzing **internal activations**.
- Our results showed that the middle layers of the model has information about the misbehaviour which can be altered to produce factual outputs.
- Devised a method to preemptively predict and intervene hallucination of generated text through analysis and consequent modification of the hidden states. Our intervened models produce significantly more factually correct generations than the base models on a wide variety of benchmark datasets.

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

- Working on **jailbreaking LLMs** to elicit stereotypical biased content.
- Investigating various persuasion methods in this context and also exploring using a multi-turn approach to solve the problem.
- Our findings show that feeding scientific summaries of papers along with citations supporting the claims of having benefits of bias leads to effective jailbreaking.

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 at the **Robert Bosch Centre for Data Science and Artificial Intelligence (RBC-DSAI)**.
- Built intelligent tools to aid in adding notable women's biography drafts to Wikipedia, which aims to **reduce the gender gap** in Wikipedia data.
- Worked on building knowledge graphs and doing graph to text generation using large language models. Finetuned models like GPT-J and GPT-Neo for this process.

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 at **RBC-DSAI.**
- 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.
- Additional Links: *Report* | *Slides* | *Github*

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).
- Additional Links: *Github* | *Report*

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.

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.
- Volunteer at **EMNLP 2021**, **NAACL 2022**.
- **Regional Mathematics Olympiad (RMO)** Finalist.
- Outstanding TA award for Spring 2024 semester.