Neemesh Yadav

Singapore - neemeshyadav137@gmail.com - +65-91747746 - Linkedin - Github - Google Scholar - Website

EDUCATION

Indraprastha Institute of Information Technology, Delhi, India — Jan 2021 — June 2024
B.Tech. in Computer Science and Biosciences — CGPA: 8.14/10

Ryan International School, Mayur Vihar, Delhi — 2019 — 2020
CBSE - Grade XII (Science) — Percentage: 94%

PUBLICATIONS

Revealing Hidden Mechanisms of Cross-Country Content Moderation with Natural Language Processing
Neemesh Yadav*, Jiarui Liu*, Francesco Ortu, Roya Ensafi, Zhijing Jin, Rada Mihalcea
ACL 2025 Findings

Effects of Theory of Mind and Prosocial Beliefs on Steering Human-Aligned Behaviors of LLMs in Ultimatum Games
Neemesh Yadav, Palakorn Achananuparp, Jing Jiang, Ee-Peng Lim
To be submitted for ACL 2026

Are LLMs Good Safety Agents or a Propaganda Engine?
Neemesh Yadav*, Francesco Ortu*, Jiarui Liu, Joeun Yook, Bernhard Schölkopf, Rada Mihalcea, Alberto Cazzaniga Zhijing Jin
Under-work

Inference-Time Selective Debiasing
Gleb Kuzmin, **Neemesh Yadav**, Ivan Smirnov, Timothy Baldwin, Artem Shelmanov
NAACL 2025 Short

QUENCH: Quizzing Benchmark for evaluating the reasoning capabilities of LLMs
Mohammad Aflah Khan*, **Neemesh Yadav***, Sarah Masud, Md. Shad Akhtar
COLING 2025

Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech
Neemesh Yadav*, Sarah Masud*, Vikram Goyal, Md Shad Akhtar, Tanmoy Chakraborty
ACL 2024 Findings

The Duality of Hope: A Critical Examination of Controversial Annotations in HopeEDI
Mohammad Aflah Khan*, **Neemesh Yadav***, Diksha Sethi, Raghav Sahni
ICLR 2024 - Tiny Paper Track

Beyond Negativity: Re-Analysis and Follow-Up Experiments on Hope Speech Detection
Neemesh Yadav*, Mohammad Aflah Khan*, Diksha Sethi, Raghav Sahni

* implies equal contribution

ICLR 2023 - Tiny Paper Track

The Art of Embedding Fusion: Optimizing Hate Speech Detection
Mohammad Aflah Khan*, **Neemesh Yadav***, Mohit Jain, Sanyam Goyal
ICLR 2023 - Tiny Paper Track

EXPERIENCE

Singapore Management University Singapore
Research Engineer January 2025 - Present
Guides: Dr. Ee-Peng LIM & Dr. Palakorn Achananuparp

- Leading efforts in understanding the effects of theory of mind as a social reasoning component in LLMs for controlled economic games and benchmarking their ToM reasoning in real-life social discourse – both to be submitted to ACL 2026.

Max Planck Institute for Intelligent Systems Remote
Research Assistant October 2024 - Present
Guides: Dr. Zhijing Jin & Dr. Bernhard Schölkopf

- Led a project on studying and analyzing the various content moderation policies in different countries across time (ACL Findings 2025).
- Led a project on studying the effects of self-censorship within LLMs and its distinction from content moderation policies.

Laboratory for Computational Social Systems (LCS2) Delhi, India
Undergraduate Student Researcher May 2022 - September 2024
Guides: Dr Md. Shad Akhtar & Dr. Tanmoy Chakraborty

- Led a project on highlighting the existing retrieval and knowledge incorporation methods for generating explanations or stereotypical implications of Hate Speech (ACL Findings 2024).
- Co-authored a project on benchmarking the indic reasoning capabilities of LLMs in quiz-style benchmark (COLING 2025).

LibrAI Remote
Research Collaborator August 2023 - August 2024
Collaborators: Dr. Xudong Han & Dr. Artem Shelmanov

- Collabarted on a project for devising an inference time safety mechanism based on selective debiasing and uncertainty quantification for promoting fairness in models (NAACL 2025 Short).

Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI) Abu Dhabi, UAE
Undergraduate Research Intern May 2023 - June 2023
Advisor: Dr. Preslav Nakov

- Developed a tool to help users profile news media sources on the basis of their factuality and political stance on the fly
- Worked on extending previous work by incorporating other dimensions such as rationale/reasoning to follow work in Explainable AI by generating rationales for the decisions

Fiverr Remote
Freelance Software Developer January 2021 - February 2022

- Creating chatbots, software development, and solving algorithmic puzzles.

PROJECTS

Automatic QnA Generation from YT Videos Course Project

Github, **Report**

Course Instructor: Dr. Rajiv Ratn Shah

- We developed a streamlit webapp to help with the existing problem of loss of attention span during lectures.
- The webapp segments the videos into fixed sections, and generates questions over those segments which are then presented to the user. The goal is to help the user understand the concepts being taught in the video.

dCrypy

Github,

- A simple Python tool that incorporates multiple ciphers, and can be used to decode ciphertexts or encode plaintexts using the appropriate keys.

TLDR-Bot Hackathon Project

Github

- A Twitter bot that summarizes long threads into a single tweet, using the T5 model, for those who want to scope the concept of the thread.

ZoneIn: Mindmap Generator Hackathon Project

Github

- We developed a tool that can generate mindmaps of some query YouTube videos based on their transcription
- This is extremely useful for people who suffer from a relatively low attention span and can also be extended to real-time meetings with appropriate transcribing tools.

Hate Speech Curbing Bot Hackathon Project

Github

- A discord bot that has a hate speech filter and can be used to curb the same in real-time discord threads
- The bot also offers multiple games, such as Hangman and Guess the Word

SKILLS

- **Relevant Coursework:** Natural Language Processing†, Deep Learning†, Reinforcement Learning†, Machine Learning, Information Retrieval, Reinforcement Learning, Linear Algebra, Probability & Statistics, Multivariate Calculus
- **Programming:** Python, Javascript, Arduino
- **Libraries, Frameworks & Other Technologies:** Huggingface, PyTorch, TensorFlow, Keras, W&B, Jupyter Notebooks, Streamlit, Flask, Scikit-Learn, FastAPI, Git, Linux, Bash

AWARDS AND ACHIEVEMENTS

- Received **Microsoft Travel Grant** for **ACL 2024** in Bangkok, Thailand
- Attended **Google Research Week 2024** in Bangalore, India.
- 1^{st} **place (as a solo-player) in 3 consecutive cryptic hunts organized by IEEE IIITD**: Won cryptic hunts, where participants are required to solve questions with ciphers and OSINT.

† Graduate Level Course

- 1^{st} **place** in **The12Rings**: An intense 144-question-long cryptic hunt spanning across 10 months and recording a participation of over 50,000 people.
- **Finalist Anveshan Hackathon**: Designed a Mindmap Generator from YouTube videos using transcripts and open source models.
- **Runner Up Byld + WiT Hackathon**: Designed a Discord Bot that detects toxic messages and alters them to become funny alternatives, thereby removing the hate and introducing humor in the conversation.

ORGANIZER, REVIEWER & VOLUNTEER WORK

- **Organizer**: Research Event Organizing Team Lead for **Esya 2023**
- **Reviewer**: LREC COLING 2024, COLING 2025, ACL 2024, ACL 2025, EMNLP 2025, EACL 2026, ICWSM 2026
- **Volunteer** ICON 2022 held at IIITD: Helped out in the smooth conduct of keynotes, talks, and paper presentations

CO-CURRICULAR ACTIVITIES

- **Core Member Byld**: The Development Club at IIITD
- **Research Mentor** for the **Undergraduate Research Club**: Mentoring students interested in research
- **Research Event Organizing Team Lead** for **Esya 2023**: IIITD's Annual Technical Festival
  - Led the organization of HackCog, which was predominantly focused on solving tasks related to human cognition
  - Mentored the organizing team of PromptCraft, where participants were asked to craft their own prompts and come up with creative but unique solutions to problems using LLMs
  - Mentored the organizing team of MLWars, a kaggle competition, where participants were asked to differentiate between a set of spectrograms of human speech from other noises.
- **CompSoc Secretary** for **IEEE IIITD**: IEEE Student Branch of IIITD
  - Mentoring and organizing events that come under CompSoc, such as cryptic hunts (both online and offline) and workshops.