Mayank Goel

+91 789-054-5454 | themayankgoel28@gmail.com | linkedin | github | substack

Education

International Institute of Information Technology - Hyderabad Hyderabad, India
B.Tech in Computer Science, CGPA: 7.96/10 *Jul. 2019 – Apr. 2023*

International Institute of Information Technology - Hyderabad Hyderabad, India
MS in Computational Linguistics *May 2023 – Dec. 2024*

Experience

Researcher Aug. 2025 - Present
Lossfunk *Bangalore, India*

- Exploring new paradigms to approach AI Safety. Primary topics of interest are moral philosophy, game theory, multi-agent systems.
- Ongoing exploration of Moral Machine Dataset by MIT - ran several experiments: treating this as a non-convex optimization problem, trying out several baselines. Primary direction of research on this has been to fit a custom Transformers architecture and run mechanistic interpretability experiments on it such as circuit probing. Ran miscellanous experiments using LLMs on morality before choosing this specific direction.

Research Mentor Jul. 2025 - Present
Algoverse Research *Remote*

- Mentoring teams as part of a research bootcamp, consisting primarily of college students and early-stage engineers. Responsibilities include holding meets, discussing and coming up with the direction, unblocking team members when needed, and acting as liaison to senior mentors and organizers.
- One paper from this has been accepted to the SEA Workshop, NeurIPS '25.
- Executed skills about effective conflict management while maintaining strong epistemic hygiene.

Research Engineer Mar. 2025 - Jul. 2025
IBM Research *Bangalore, India*

- Contributed to IBM's Granite LLM team, focusing on improving data quality for model training through various pre-processing methods, synthetic data generation and ablation studies - implementing several research papers on improving data quality. Became very familiar with the data that were being fed as part of the SFT pipeline. Constructed synthetic data for specific tasks based on literature.
- A primary effort led was to construct a pipeline for data rewriting efficiently scaling inference using vLLM. The focus was on building infra for automated workflow to effectively use GPU resources and handle concurrency issues while handling multiple files across GPUs and nodes with an aim to minimize idle GPU while allowing others to access resources.

Researcher Jan. 2025 - Apr. 2025
AI Safety Camp *Remote*

- Part of a team selected by Prof. Masaharu Mizumoto of JAIST on a model to improve the morality of LLMs.
- Involved with discussions on philosophical grounding of research methodology, primarily around Virtue Ethics as a moral framework and approaches for moral dilemma generation.
- Tried SFT/DPO on human-labeled moral dilemmas on models to test performance on several morality related benchmarks. Owned the evaluation pipeline for this project.

Research Engineer Sep. 2024 – Mar. 2025
Deccan AI *Hyderabad, India*

- Involved with benchmarking of models on various tasks, across modalities, such as tasks on image, speech, code. From searching for the appropriate metrics to explaining them to client-facing teams.
- Worked on automated QC for data generated by human annotators. Included fine-tuning models on curated data as part of a company repository.
- Took several interviews for part-time and full-time interviews for ML role. Effectively managed two interns to help their skill development and contribution to the company.

Research Intern Oct. 2023 – Jul. 2024
ComplexData Lab, MILA AI *Remote*

- Designed experiments using the Debate framework for uncertainty quantification of LLM outputs.
- Worked on experiments to analyze and visualize data from other team projects on uncertainty quantification.
- Work done during this phase was accepted to COLM '24 and NeurIPS SafeGenAI Workshop '24

Teaching Assistant Jul. 2021 – Apr. 2024

IIIT Hyderabad *Hyderabad, India*

- Was involved with designing and evaluating assignments and conducting tutorials for the course Introduction to NLP in Spring '22 and Advanced NLP in Monsoon '23

Research Intern Jun. 2023 – Jul. 2023

Robert Bosch GmbH *Bangalore, India*

- Built upon an existing document hierarchal clustering project, add functionality to use sentence transformers.
- Built a robust pipeline to fine-tune sentence transformers using various methods.
- Developed a novel evaluation metric exclusively tailored to hierarchical clustering, addressing existing limitations with evaluation methods.

Research Associate Mar. 2021 – Feb. 2023

Proxzar AI *Remote*

- Fine-tuned Llama-7B on finance data using LORA and quantization. Did a hyperparameter search.
- Developed a robust e-commerce website crawling pipeline using Apache Nutch and Selenium
- Built a Spacy3 model utilizing Transformersfor a production-grade custom NER model for an e-commerce company, as part of an end-to-end system, with a dataset augmented by 20% using fuzzy matching. Achieved accuracy of 91% on a multi-classsystem

Research Intern May 2021 – Jun. 2021

Trivedi Centre for Political Data, Ashoka University *Remote*

- Built a streamlit dashboard to display visualized data of MPs and Governors in India, which accepts a variety of parameters. Work can be viewed at [link].

Publications

Building Interpretable Models for Moral Decision-Making AAAI '26, Machine Ethics Workshop

Authors: **Mayank Goel**, Aritra Das, Paras Chopra

Automated Specialization of Stateful Agent Systems NeurIPS '25 Scaling Environments for Agents Workshop

Authors: Myan Vu, Harrish Ayyanar, Pang Jiang, Anwiketh Reddy, **Mayank Goel**, Kevin Zhu

Conceptual Limitations of Current AI Safety Approaches and Virtue Ethics as an Alternative Technical AI Safety Conference '25

Authors: Masaharu Mizumoto, Rujuta Karekar, Mads Udengaard, **Mayank Goel**, Daan Henselmans, Nurshafira Noh, Saptadip Saha, Pranshul Bohra

Epistemic Integrity in Large Language Models NeurIPS '24 SafeGenAI Workshop

Authors: Bijean Ghafouri, Shahrad Mohammadzadeh, James Zhou, Pratheeksha Nair, Jacob-Junqi Tian, **Mayank Goel**, Reihaneh Rabbany, Jean-François Godbout, Kellin Pelrine

Web Retrieval Agents for Evidence-Based Misinformation Detection COLM '24

Authors: Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, **Mayank Goel**, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, Kellin Pelrine

Joke Generation Using Masked Language Infilling on Automatically Extracted Templates Humor Research Conference '23

Authors: **Mayank Goel**, Abhijit Manatkar

Towards conversational humor analysis and design Humor Research Conference '21

Authors: Tanishq Chaudhary*, **Mayank Goel***, Radhika Mamidi

Automated Evalauation of Conversational Prompts using Ordinal Ranking for Humor Generation ICON '24
Authors: **Mayank Goel**, Parameswari Krishnamurthy, Radhika Mamidi

Technical Skills

Languages: Python (proficient), Bash, Javascript, C++
Frameworks: PyTorch, HuggingFace Transformers, Streamlit, Selenium, SpaCy

Awards and other Relevant Information

- **Research Award** from IIIT Hyderabad in 2021
- **Top 30** in the **Indian edition of International Olympiad of Linguistics**
- **Participant**, **ARENA 2024, Spring**
- **Participant**, **AI Safety Fundamentals Course, BlueDot Impact**
- **Ex-head** of University **Debate Society** and **Humor Club**
- **Debated competitively in several intra and inter collegiate competitions.**
- **Founder-Moderator** of **Philosophy Friday Discussion Group**
- **Events Head** of **Felicity, IIIT's student fest**.
- **Best Project, Theme Track** in **Natural Language API Hackathon by expert.ai**
- **1st Prize, Equity Track** in **HackUMBC, University of Maryland**
- **Best Web Application** in **HackAtHome, Brown University**
- **3rd Prize, Walmart Track** in **TAMU Datathon, Texas A&M University**
- **1st Prize, Stellantis Track** in **Megathon '22, IIIT Hyderabad**