# Ananya Malik

Email: ananyamalik1999@gmail.com
LinkedIn: https://www.linkedin.com/in/ananyamalik/ Mobile: +1 470 753 7363
Github: https://ananyamalik.github.io/ Boston, MA


**Research Interests:** As a CS PhD student, I am broadly interested in aligning AI and large language models (LLMs) to promote equity. My current work explores AI control methods and the impact of personalisation on AI safety. Previously, I investigated culturally informed safety frameworks for online communities, with an emphasis on mitigating misinformation and hate speech.

## Education

- **Northeastern University** — Boston, USA
  PhD - Computer Science; GPA: 3.93 — Sept 2024 - Present
  **Advisor:** Mai ElSherief
  **Key Words:** AI Safety; CoT Reasoning; Culturally-Aligned AI; Human-centered NLP; NLP; ML; Fairness ;Online Communities; Evaluations of AI; Few-shot / Zero-shot Learning; Model Evaluation; Benchmarking; Causal Interventions; Persona Modeling; Prompt Engineering; Mechanistic Interpretability; Model Cognition

- **Georgia Institute of Technology** — Atlanta, USA
  Masters of Science - Computer Science- ML specialisation; GPA: 4.0 — Aug 2021 - Dec 2022
  **Advisor:** Srijan Kumar
  **Courses:** Machine Learning, Deep Learning, Web Search, Data Management with ML, Deep Learning for Text, Data Visualisation

- **Dwarkadas J Sanghvi College of Engineering** — Mumbai, IN
  Bachelors of Engineering - Computer Engineering; GPA: 9.79 — Aug 2017 - Jun 2021
  **Courses:** Analysis of Algorithms, Machine Learning, Artificial Intelligence Soft Computing, Big Data, Software Development, OS, HMI

## Experience

- **MARS: Mentorship for Alignment Research Students** — Cambridge, UK/Remote
  Research Fellow — July 2025 - Present
  - **Mentors**: : Geodesic Research
  - **Research Project**: Investigating the feasibility of different AI control strategies in LLMs through Chain of Thought Injections

- **Amazon** — Seattle, USA
  Software Development Engineer and Intern — Aug 2023 - Sept 2024; May 2022-Aug 2023
  - **Build Data Debugger**: To identify and highlight causally redundant attributes in the data pipeline
  - **Compare Text Based Data for similarity**: Developed an algorithm to compare the attributes in a JSON data by the text values using embedding that lead to a reduction in false positives by 17 %
  - **Created Tool**: Implemented the backend on AWS Lambda, pulling data from S3 buckets and frontend on React to visualise the lineage of the data by type of data loss and attribute

- **SimPPL** — Remote
  Research Mentor
  - **Research**: Led a project to understand the impact of dialects and external identities on hate speech classification. Paper presented orally at NeurIPS SafeGenAI Workshop.
  - **Mentor**: Mentored students to collect and analyse the the impact of online communities on real world situations using YouTube analysis

- **CLAWS Lab, Georgia Tech** — Atlanta, USA
  Research Engineer — Aug 2022 - Aug 2023
  - **Misinformation Platform**: Worked on building a platform to identify and interact with tweets flagged as misinformation, assisted data collection and creating the database
  - **Counter Misinformation**: Working on finding the motivation, classifying counter misinformation on Twitter using LLMs, HCI

## Publications

- **Paper**: **Ananya Malik**, Nazanin Sabri, Melissa Karnzae, Mai ElSherief *"Are LLMs Empathetic to All? Investigating the Influence of Multi-Demographic Personas on a Model's Empathy"* Accepted to EMNLP 2025 Findings
- **Paper**: **Malik, Ananya**; Sharma, Kartik; Ng Lynette Hui Xian; Bhatt, Shaily "Who Speaks Matters: Analysing the Influence of the Speaker's Ethnicity on Hate Classification." Accepted to EMNLP 2025 Findings; NeurIPS SafeGenAI 2024 (Oral Presentation)
- **Paper**: Bangzhao Shu, **Ananya Malik**, Mai ElSherief "Angry Bots: Emotional Bias in Language Model Predictions" Under Submission at Nature
- **Paper**: Nazanin Sabri, **Ananya Malik**, Banghao Shu, Jason Jeffrey Snyder, Laurie Kramer, Mai ElSherief "He gets to be the fun parent": Understanding and Supporting Burnt-Out Mothers in Online Communities" Under Submission at CSCW 2026
- **Paper**: **Malik, Ananya**. "Evaluating Large Language Models through Gender and Racial Stereotypes." arXiv preprint arXiv:2311.14788 (2023).

- **Article**: Intent to Hate: A study on the intent behind hate generated on Twitter against the Asian Community during the COVID-19 pandemic.
- **Article**: Of Multimodal Comprehension and Ignorance.
- **Paper**: Successive Image Generation from a Single Sentence, Amogh Parab, **Ananya Malik**, Arish Damania, Arnav Parekhji, Pranit Bari, ITM Web Conf. 40 03017 (2021), DOI: 10.1051/itmconf/20214003017.
- **Chapter**: **A.Malik**, Y. Javeri, M. Shah, R. Mangrulkar, 'Impact Analysis of Covid 19 News Headlines on Global Economy', Cyber-Physical Systems for COVID-19, Elsevier.
- **Paper**: **Malik A**. Survey paper on applications of generative adversarial networks in the field of social media. Int J Comput Appl (IJCA). 2020;175(20):13–18. doi:10.5120/ijca2020920728

## Academic Service

- **Graduate TA: CS 5200: Database Management Systems** — Boston, USA
  Prof Martin Schedlbauer: Held Office Hours, Grading of Assignments and Papers — May 2026 - Present
- **Graduate TA: CS 4100: Foundations of AI** — Boston, USA
  Prof Chris Amato: Taught about Advancements in AI, Held office hours, graded papers — Jan 2025 - April 2025
- **Research Mentor: SimPPL** — Boston, USA
  Mentoring undergrad students to do research to understand social media landscapes in politics — Jan 2024 - Present
- **Graduate TA: CS 3600 Intro to AI** — Atlanta, USA
  Prof James Rehg, Prof Mark Riedl; Held office hours, recitations, review sessions,graded papers — Jan 2022 - Dec 2022
- **DJ Unicode** — Mumbai, IN
  Leading a 130+ member development team working on full-stack projects for colleges,non-profits — Jul 2018 - May 2021
- **TA: Machine Learning Summer School** — Mumbai, IN
  TA for Machine Learning School, held recitations and office hours — Jul 2018 - May 2021

## Skills Summary

- **Languages**: Python, C, C++, JavaScript, SQL, JAVA
- **Frameworks**: Scikit, NLTK, SpaCy, TensorFlow, Keras, Django, Flask, NodeJS, REST API
- **Tools and Areas**: Prompting, In Context Learning, PyTorch / TensorFlow, Hugging Face Transformers, LongChain, Inspect AI, Flask, Django, ReactJS, GIT, PostgreSQL, MySQL, SQLite, AWS Lambda, AWS Athena, Linux, Web, Windows, Machine Learning, Deep Learning, Computer Vision, Natural Language Processing, GANs