

# Aatman Vaidya

 [aatmanvaidya.github.io](https://github.com/aatmanvaidya)  [aatmanvaidya@gmail.com](mailto:aatmanvaidya@gmail.com)  [github.com/aatmanvaidya](https://github.com/aatmanvaidya)  [Google Scholar](#)

**Research Interests:** Social Computing, Online Safety, Evaluation, AI Safety, Language Models & their Societal Impact

## Education


**University of Tübingen**, Tübingen, Germany  
Master of Arts in Computational Linguistics

Oct 2025 – Present

**Ahmedabad University**, India  
Bachelor of Technology in Computer Science & Engineering












Aug 2019 – Jul 2023

## Experience

**Tattle Civic Tech**, India 

Aug 2023 – Sep 2025

Research Engineer

- > I help build citizen centric open-source tools and datasets to understand and respond to online harms and abuse. 
- > Led the development of **Feluda**, a configurable engine for multilingual and multimodal content analysis to assist fact-checkers and researchers in combating misinformation. 
- > Contributed to **Uli**, a browser extension that redacts slurs and abusive language, archives problematic content, and counters online gender-based violence experienced by marginalized communities in India.  
- > Contributed to the creation of the **Deepfakes Analysis Unit**, a WhatsApp-based tipline for detecting and responding to AI-generated synthetic media. I built scalable data pipelines to cluster large volumes of audio and video data into social media thematic categories, enabling more efficient analysis by fact-checkers. 
- > Collaborated with **MLCommons** to develop an AI safety benchmark dataset focused on hate speech and sex-related crimes in India, created through a human-centered participatory approach with experts and advocates from marginalized communities   **Dataset** 
- > Contributed to **Viral Spiral**, an adaptive digital card game about sharing news on the Internet. The game reflects the ways and reasons misinformation is shared.  
- > Worked with NGOs in the health and education sectors through **Tech4Dev's AI Cohort** program, implementing technical safety guardrails for AI Chatbot applications. 

## Publications

S=In Submission, C=Conference, W=Workshop, P=Poster/Demo, J=Journal, R=Report

### S.2 Quantifying the Illicit Ecosystem of Betting Apps in India

Aatman Vaidya, Kiran Garimella

In Preparation, to be submitted at ICWSM 2026

### S.1 Modelling the Spread of Toxicity and Exploring its Mitigation on Online Social Networks

Aatman Vaidya, Harsh Bhagat, Seema Nagar, Amit Nanavati

In Submission at ACM WebSci 2026

### P.1 Analysis of Indic Language Capabilities in LLMs


Aatman Vaidya, Tarunima Prabhakar, Denny George, Swair Shah

arXiv:2501.13912

[ ArXiv ]

### R.1 AI Safety Benchmark Dataset for Hate and Sex Related Crimes in Hindi

Mansi Gupta, Srravya Chandhiramowuli, Vamsi Pothuru, Saumya Gupta, Tarunima Prabhaka, Kaustubha K, Aatman Vaidya

Report Submitted to **MLCommons** and work included in the **AILuminate Safety Benchmark Dataset** 

### W.1 The Uli Dataset: An Exercise in Experience Led Annotation of oGBV

Arnav Arora, Maha Jinadoss, Cheshta Arora, Denny George, Brindaalakshmi...Aatman Vaidya, Tarunima Prabhakar

Workshop on Online Abuse and Harms at North American Chapter of the Association for Computational Linguistics

★ Outstanding Paper Award

[ NAACL'24 ]

### C.3 Analysing the Spread of Toxicity on Twitter

Aatman Vaidya, Seema Nagar, Amit Nanavati

ACM International Conference on Data Science and Management of Data

[ CODS-COMAD'24 ]

### C.2 Forecasting the Spread of Toxicity on Twitter

Aatman Vaidya, Seema Nagar, Amit Nanavati

IEEE International Conference on Cognitive Machine Intelligence

[ IEEE CogMI'23 ]







### C.1 Overview of the 2023 ICON Shared Task on Gendered Abuse Detection in Indic Languages

Aatman Vaidya, Arnav Arora, Aditya Joshi, Tarunima Prabhakar









International Conference on Natural Language Processing

[ ICON'23 ]

## Work Mentions and Media Coverage

<b>Center for Democracy &amp; Technology</b> on “Moderating Tamil Content on Social Media” 	June 2025
<b>FOSS United</b> on “Meet The Maintainers, 31 days 31 FOSS Maintainers from India” 	May 2025
<b>World Economic Forum</b> on “Year of elections: Lessons from India’s fight against AI-generated misinformation” 	Aug 2024
<b>NY Times</b> on “A Small Army Combating a Flood of Deepfakes in India’s Election” 	Jun 2024
<b>The Nieman Lab</b> on “Indian journalists are on the frontline in the fight against election deepfakes” 	May 2024
<b>Digital Public Goods</b> on “Safeguarding Information Integrity during Elections with Digital Public Goods” 	Apr 2024

## Talks

<b>Tech4Dev AI Cohort Program</b> on “Technical Implementation of AI Safety Guardrails”  	Bengaluru, Sep 2025
<b>Tech4Dev AI Cohort Program</b> on “Safety by Design in AI for Social Good Applications” 	Remote, Jul 2025
<b>MisinfoCon India’25</b> on “A Look at Open-Source Deepfake Detection”  	Bengaluru, Mar 2025
<b>AI for Global Development by Agency Fund</b> on “Evaluating Indic Language Performance in LLMs” 	Bengaluru, Mar 2025
<b>Ahmedabad University SNA Guest Lecture</b> on “Modelling Hate Speech on a Social Network” 	Ahmedabad, Apr 2024
<b>ACM Winter School on Network Science</b> on “Building NLP classifiers to detect Hate Speech”	Ahmedabad, Dec 2023
<b>DEF India Digital Citizen Summit</b> on “Tools to respond to Online Gender Based Violence” 	Remote, Apr 2024

## Featured Academic Projects and Collaborations





**Financial Scams on Social Media** Jun’25 - Present  
w/ *Kiran Garimella*

- Performed an extensive quantitative analysis of the betting app ecosystem in India through a multi-faceted, mixed-methods approach.
- Analyzed 20K Meta ADs, 10K Instagram Posts and 329K Google Play Store reviews to uncover patterns in promotion, user sentiment, and thematic content.
- We systematically identify and quantify a range of harmful practices, including the use of misleading advertisements, deceptive celebrity endorsements, and illegal deepfakes designed to feign legitimacy.
- Our analysis reveals that, despite platform policies and local laws, thousands of betting app advertisements persist on major social media networks.


**Modelling Spread of Hate Speech** Aug’22 - Sep’25  
w/ *Amit Nanavati, Seema Nagar*

- Developed a network-based model to simulate the spread of toxicity on social media, treating users as transformers who amplify, attenuate, or replicate toxic content. [Full Paper @ **CODS-COMAD’24**]
- Conducted a temporal and behavioral analysis of Twitter (100K users, 17.2M posts), Koo (215K users, 16.9M posts), and Gab (62K users, 20.1M posts) networks, constructing directed graphs of up to 2.3M edges to study how toxicity evolves across time.
- Proposed and evaluated peace-bot interventions to mitigate online toxicity. We saw up to 11.6% reduction in total network toxicity depending on graph structure and bot placement strategy. [In Submission @ **ACM WebSci’26**]

**Tracking Online Gender Based Abuse** Jul’23 - Present  
at/ *Tattle Civic Tech*

- Created a dataset of 1.1M **YouTube comments** to track gender-based violence, leveraging a crowd-sourced Indic **slur list** curated by gender rights researchers and activists. 
- Fine-tuned NLP models (Llama Guard, Qwen Guard) to detect coded and dog-whistle language in Indian contexts by creating a multilingual dataset of 1,500+ Reddit comments. 
- Developed media matching workflows with TMK and PDQ hashing and clustering pipelines for large-scale detection and visualization of harmful or misleading content. 
- Prototyped linguistic methods to identify duplicate or derived slurs and trace their root word origins for improved lexical content moderation. 

**Building AI Safety Guardrails** Jul’25 - Present  
w/ *Tattle Civic Tech and Project Tech4Dev*

- Developed a RESTful content moderation API integrating lexicon-based slur filtering and Llama Guard model for real-time classification of harmful content. 
- Conducted workshops to present a taxonomy of AI Risks and Hazards to NGO’s working in the health and education sector.

## Awards and Honors

- Outstanding Paper Award @ Workshop of Online Harms and Abuse NAACL 2024
- Bosch Future Mobility Challenge | Semi-Finalist Programmed and Engineered an autonomous driving car to navigate a miniature city. One of the 24 teams out of 118, and only team to represent India for the finals in Cluj, Romania.

## Teaching and Academic Service

- Reviewer: ICWSM’26, COLM’25, WOAHL ACL’25, ACL Rolling Review’24
- Shared Task: Gendered Abuse Detection in Indic Languages@ICON’24
- Discrete Mathematics (MAT 101), Ahmedabad University. Teaching Assistant Winter 2022
- Center for Learning and Empowerment (NGO), Volunteer Teacher. Jul 2023 - Apr 2024
  - > Taught 20+ secondary school kids mathematics for an entire academic year, at a tribal village in Jharkhand, India.
- Programming Club, Ahmedabad University. Content Head and Event Organiser July 2020 - May 2022
- Reading Club, Ahmedabad University. Organiser Dec 2021 - May 2022

## Skills

- Languages Python, JavaScript, Elixir C/C++, Kotlin, GraphQL
- Frameworks PyTorch, Tensorflow, Nodejs, OpenCV, networkx, Elasticsearch, Phoenix
- Technologies MySQL, PostgreSQL, Docker, AWS (S3, EKS, EC2), Kubernetes, CI/CD, Android Studio