Mohammed Safi Ur Rahman **Khan**

PhD Student, IIT Madras | AI4Bharat

safikhansoofiyani.github.io Linkedin GitHub safikhan2000@gmail.com Google Scholar

SSB 434, Subramonian Shankar Block, IIT Madras, TN, India

Education

Present – Jul 2024 | **Indian Institute of Technology (IIT), Madras** — **Chennai, India**
Ph.D, Wadhwani School of Data Science and AI
Advisor : Mitesh M. Khapra

Jul 2023 – Aug 2021 | **Indian Institute of Technology (IIT), Madras** — **Chennai, India**
M.Tech, Computer Science & Engineering
CGPA: 9.75/10
Advisors : Pratyush Kumar, Mitesh M. Khapra

Jul 2021 – Jun 2017 | **Osmania Univeristy** — **Hyderabad, India**
B.E, Computer Science & Engineering
CGPA: 8.94/10

Experience

Present – Jul 2024 | **AI4Bharat, IIT Madras** [🌐] — **Chennai, India**
PhD Researcher
Working on building and evaluating Multilingual Large Language Models (LLMs) focusing on Indian languages.

Jul 2023
AI Resident
Worked on building large-scale data infrastructure to create, curate, and clean Indic Language data for training Large Language Models (LLMs).

Jul 2022
Graduate Student Researcher
Worked on narrow-domain adaptation of Automatic Speech Recognition (ASR) systems using Class-Based Language Models.

Jul 2022 – May 2022 | **Predactica** [🌐] — **Remote**
Data Science Intern
Worked on Time Series Forecasting. Added the Time-Series specific data preparation, cleaning, and modeling functionality in the existing ML-Studio product.

Apr 2021 – Jan 2021 | **Cisco Systems** [🌐] — **Bangalore, India**
Worked on integrating Thousand Eyes network monitoring tool in the existing UCM Cloud.

Publications

C=Conference, P=Preprint, *=Equal Contribution

[C.1] **IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages** [🔗] [Code]
Mohammed Safi Ur Rahman Khan*, Priyam Mehta*, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad G, Varun Balan G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra
62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand
🏅 Outstanding Paper **[ACL 2024]**

[P.2] **Finding Blind Spots in Evaluator LLMs with Interpretable Checklists** [🔗] [Code]
Sumanth Doddapaneni*, Mohammed Safi Ur Rahman Khan*, Sshubam Verma and Mitesh M. Khapra
[Under Review]

[P.1] **Airavata: Introducing Hindi Instruction-tuned LLM** [🔗] [Code]
Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan
[Under Review]

Projects

Narrow Domain Adaptation of Automatic Speech Recognition (ASR) Systems [🔗] M.Tech Thesis

- Worked on adapting existing general purpose CTC-based Automatic Speech Recognition (ASR) systems on narrow domains like Finance, Medical, E-Commerce, etc.
- Explored inference-time approaches like constrained decoding using Class-Based Language Models.

IndicPunct - Inverse Text Normalization for Indic Languages [Code]

- Worked on enhancing Weighted Finite State Transducer (WFST) based Inverse Text Normalization (ITN) System for 12 Indic Languages.
- Deployed as a post-processor at various speech systems due to its capability of handling various colloquial forms efficiently.

Indic-numtowords - Lightweight Text Normalization for Indic Languages [Code]

- Built a simple, lightweight Python package for text normalization for 13 Indic Languages.

Application Screening of Crowdfunding projects using A.I [Code] B.E Thesis

- Explored the problem of automatic screening of Crowdfunding project proposals.
- A large case study with extensive Exploratory Data Analysis and evaluation of multiple strategies and models on real-world data.

Voice Prescription [Code] SIH Hackathon

- An Android application designed to help small-scale doctors and clinics automate their writing process of medical prescriptions.
- Instead of manually typing or writing the prescription, the app uses speech recognition, where the doctor only speaks to the app.

Help me Cook [Code]

- An android application designed to help people decide what they want to cook based on their preferences any single day.

Awards and Achievements

Outstanding Paper Award @ ACL 2024 [🌐] Won the Outstanding paper award for the IndicLLMSuite.

MSR Travel Grant, 2024 [🌐] For attending ACL 2024 held in Bangkok, Thailand.

IIT Madras STAR TA Award Star Teaching Assistant award for the year 2023.

GATE CS& IT 2024 Secured All India Rank 113 among 100K candidates.

Cassini Scientist for a Day 2013 National Winner in the Cassini Scientist for a Day 2013 competition organized by NASA.