

# Mohammed Safi Ur Rahman Khan




PhD Student, IIT Madras | AI4Bharat

 safikhansoofiyanigithub.io  LinkedIn  GitHub  safikhan2000@gmail.com  Google Scholar  
 619, New Academic Complex-2, IIT Madras, TN, India

## Education




Present Jul 2024	<b>Indian Institute of Technology (IIT), Madras</b> Ph.D, Wadhvani School of Data Science and AI Advisor : <a href="#">Mitesh M. Khapra</a>	Chennai, India
Jul 2023 Aug 2021	<b>Indian Institute of Technology (IIT), Madras</b> M.Tech, Computer Science & Engineering CGPA: 9.75/10 Advisors : <a href="#">Pratyush Kumar</a> , <a href="#">Mitesh M. Khapra</a>	Chennai, India
Jul 2021 Jun 2017	<b>Osmania Univeristy</b> B.E, Computer Science & Engineering CGPA: 8.94/10	Hyderabad, India

## Experience

Present Jul 2024	<b>AI4Bharat, IIT Madras</b>  <i>PhD Researcher</i> Working on building and evaluating Multilingual Large Language Models (LLMs) focusing on Indian languages.	Chennai, India
Jul 2023	<i>AI Resident</i> Worked on building large-scale data infrastructure to create, curate, and clean Indic Language data for training Large Language Models (LLMs).	
Jul 2022	<i>Graduate Student Researcher</i> Worked on narrow-domain adaptation of Automatic Speech Recognition (ASR) systems using Class-Based Language Models.	
Jul 2022 May 2022	<b>Predactica</b>  <i>Data Science Intern</i> Worked on Time Series Forecasting. Added the Time-Series specific data preparation, cleaning, and modeling functionality in the existing ML-Studio product.	Remote
Apr 2021 Jan 2021	<b>Cisco Systems</b>  Worked on integrating Thousand Eyes network monitoring tool in the existing UCM Cloud.	Bangalore, India

## Publications

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

- [C.2] **Finding Blind Spots in Evaluator LLMs with Interpretable Checklists**  [Code]  
Sumanth Doddapaneni\*, [Mohammed Safi Ur Rahman Khan\\*](#), Sshubam Verma and Mitesh M. Khapra  
*The 2024 Conference on Empirical Methods in Natural Language Processing* [EMNLP 2024]
- [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  
*62<sup>nd</sup> Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand*  
 Outstanding Paper [ACL 2024]

- [P.4] **BhasaAnuvaad: A Speech Translation Dataset for 14 Indian Languages** [🔗] [Code]  
Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan,  
Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M. Khapra and Raj Dabre  
[Work In Progress]
- [P.3] **MILU: A Multi-task Indic Language Understanding Benchmark** [🔗] [Data]  
Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishawajeet Kumar, Rudra Murthy and Jaydeep Sen  
[Under Review]
- [P.2] **Cross-Lingual Auto Evaluation for Assessing Multilingual LLMs** [🔗] [Code]  
Sumanth Doddapaneni\*, Mohammed Safi Ur Rahman Khan\*, Dilip Venkatesh, Raj Dabre, Anoop Kunchukuttan, 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

## 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.

## Awards and Achievements

---

- Outstanding Paper Award @ ACL 2024** [🏆] Won the Outstanding paper award for the IndicLLMSuite.
- Google Travel Grant, 2024** [🏆] For attending EMNLP 2024 held in Miami, Florida.
- Microsoft Research 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.