

# Mohammed Safi Ur Rahman Khan

Google PhD Fellow, IIT Madras | AI4Bharat

[🌐 safikhansoofiyani.github.io](https://safikhansoofiyani.github.io) [in LinkedIn](https://www.linkedin.com/in/mohammed-safi-ur-rahman-khan/) [GitHub](https://github.com/safikhansoofiyani) [@ mohammed.safi@dsai.iitm.ac.in](mailto:mohammed.safi@dsai.iitm.ac.in) [Google Scholar](https://scholar.google.com/citations?user=QWzgkxUAAAAJ&hl=en)  
📍 619, New Academic Complex-2, IIT Madras, TN, India

## Education

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

## Experience

Present	<b>AI4Bharat, IIT Madras</b> [🌐]	Chennai, India
Jul 2024	<i>PhD Researcher</i> Working on building and evaluating Multilingual Large Language Models (LLMs) focusing on Indian languages. Recipient of the Google PhD Fellowship 2025.	
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.	
Jan 2026	<b>Sarvam AI</b> [🌐]	Bengaluru, India
May 2025	<i>Research Fellow</i> Worked on building the sovereign AI stack for India. Led the LLM Evaluation Efforts.	
Jul 2022	<b>Predictica</b> [🌐]	Remote
May 2022	<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.	
Apr 2021	<b>Cisco Systems</b> [🌐]	Bangalore, India
Jan 2021	Worked on integrating Thousand Eyes network monitoring tool in the existing UCM Cloud.	

## Select Publications

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

- [C.8] **FairI Tales: Evaluation of Fairness in Indian Contexts with a Focus on Bias and Stereotypes** [🔗] [Code]  
Janki Nawale\*, Mohammed Safi Ur Rahman Khan\*, Janani D, Mansi Gupta, Danish Pruthi and Mitesh M. Khapra  
*63<sup>rd</sup> Annual Meeting of the Association for Computational Linguistics, Vienna, Austria* [ACL 2025]
- [C.7] **Can Vision-Language Models Evaluate Handwritten Math?** [🔗] [Code]  
Oikantik Nath, Hanani Bathina, Mohammed Safi Ur Rahman Khan and Mitesh M. Khapra  
*63<sup>rd</sup> Annual Meeting of the Association for Computational Linguistics, Vienna, Austria* [ACL 2025]
- [C.6] **Towards Building Large Scale Datasets and State-of-the-Art Automatic Speech Translation Systems for 14 Indian Languages** [🔗] [Code]  
Ashwin Sankar\*, Sparsh Jain\*, Nikhil Narasimhan, Devilal Choudhary, Dhairyा Suman, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M. Khapra and Raj Dabre  
*63<sup>rd</sup> Annual Meeting of the Association for Computational Linguistics, Vienna, Austria* [ACL 2025]

- [C.5] **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  
*63<sup>rd</sup> Annual Meeting of the Association for Computational Linguistics, Vienna, Austria* [ACL 2025]
- [C.4] **Pralekha: An Indic Document Alignment Evaluation Benchmark** [🔗] [Code]  
Sanjay Suryanarayanan, Haiyue Song, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, and Raj Dabre  
*14<sup>th</sup> International Joint Conference on Natural Language Processing and the 4<sup>th</sup> Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* [IJCNLP-AACL 2025]
- [C.3] **MILU: A Multi-task Indic Language Understanding Benchmark** [🔗] [Data]  
Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy and Jaydeep Sen  
*2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics* [NAACL 2025]
- [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, Miami, Florida*  
🥇 Outstanding Paper [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, Suriyaprashaad 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]

## 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.

## Awards and Achievements

---

**Google PhD Fellowship** [🔗] Received the Google PhD Fellowship 2025 for Natural Language Processing.

**Outstanding Paper Awards** Won the Outstanding paper awards in ACL-2024 and EMNLP-2024.

**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.