Mohammed Safi Ur Rahman Khan

+919618326289 • ✉ da24d003@smail.iitm.ac.in
https://github.com/safikhanSoofiyani

Education

Program	Institution	%/CGPA	Year
Ph.D (Data Science and AI)	Indian Institute of Technology, Madras	-	Ongoing
M.Tech. (Computer Science and Engg.)	Indian Institute of Technology, Madras	9.75	2023
B.E. (Computer Science Engineering)	Osmania University, Hyderabad	8.94	2021
Intermediate - IPE	MS Junior College, Hyderabad	94.5%	2017
Xth Std. - SSC	FIITJEE School, Hyderabad	9.3	2015

Projects

1.IndicLLMSuite: IndicLLMSuite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Lang
Guide: Prof. Mitesh M. Khapra

- Developed resources for Indic LLMs in 22 languages, totaling 251B tokens and 74.8M instruction-response pairs, including curated, unverified, and synthetic data.
- Built an open-source pipeline for data curation from websites, PDFs, and videos, ensuring best practices in crawling, cleaning, flagging, and deduplication.
- Implemented instruction-fine tuning by combining Indic datasets, translating/transliterating English datasets, and generating conversations using LLaMa2 and Mixtral models, with a focus on toxicity alignment.
- Paper: https://arxiv.org/abs/2403.06350

2.Finding Blind Spots in Evaluator LLMs with Interpretable Checklists **Mar 2024 - Jul 2024**
Guide: Prof. Mitesh M. Khapra

- Investigated the effectiveness of LLMs as evaluators for text generation tasks, focusing on factual accuracy, instruction following, coherence in long-form writing, and reasoning proficiency.
- Developed the FBI framework to test Evaluator LLMs by introducing 2400 targeted perturbations across 22 categories in LLM-generated answers.
- Found significant shortcomings in Evaluator LLMs, with over 50% failure to detect quality drops, highlighting the need for cautious implementation in practical applications.
- Paper: https://arxiv.org/abs/2406.13439

3.Towards IndicRASA: Building a narrow domain conversational system **Aug 2022 - May 2023**
Guide: Prof. Mitesh M. Khapra *M.Tech Thesis Project*

- Working towards building a functioning "narrow-domain" conversation system for the 22 official Indian Languages.
- Building speech recognition system for Indian languages that is curated to work on narrow domains
- Approaching the problem from scratch – starting with appropriate Indian language conversation data collection and working towards building an efficient conversation model.

4. Indic Transliteration (Telugu Language) **Apr 2022 - May 2022**
Mentor: Prof. Mitesh M. Khapra, Team Size : 2 *Course Project - CS6910*

- Built a fully configurable Seq2Seq Model to do transliteration task for general Indic languages. Demonstrated for Telugu language.
- Utilized the data from Dakshina dataset. Tried out various hyperparameters like different cells (LSTM, RNN, GRU), attention, beam search, etc.

5. Application Screening of Crowdfunding projects using A.I. **Sept 2020 - June 2021**
Guide: Prof. Shyama Chandra Prasad, Team size : 3 *B.E. Project*

- Developed basic ML models that enables easy screening of applications submitted for crowdfunding. Predicts whether the application should be accepted or not.
- Utilized the data given by DonorsChoose.org, did extensive EDA as well as pre-processing.

6. Voice Prescription **Dec 2019 - Jan 2020**
Mentor: Mrs. Bhagyalakshmi, Team Size : 6 *Project, SIH Internal Hackathon*

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

7. Help me COOK **Feb 2019 - Apr 2019**
Mentor: Mrs. K. Shalini, Team Size : 3 *Mini Project*

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

Experience

1. AI Resident **June 2023 - May 2024**
AI4Bharat, IIT Madras

- Worked on creating Data Infrastructure for creating, curating and cleaning Indic Language data for training Large Language Models

2. Data Science Intern **May 2022 - Jul 2022**
Predactica

- Worked on time series data forecasting. Wrote general purpose efficient code to handle time series specific functionality like time series pre-processing, forecasting models (both statistical and deep learning models), validation of forecasted values, etc.
- Helped in integrating this new time series module in the existing ML Studio product.

3. Technical Undergraduate Intern **Jan 2021 - Apr 2021**
Cisco Systems (India) Pvt. Ltd.

- Completed DevNet and CCNA training
- Worked on integrating ThousandEyes network monitoring tool into the existing UCM Cloud

4. C++ Development Intern **Oct 2019 - Dec 2019**
cppsecrets.com

- Worked on advanced C++ libraries (like Boost) and published various articles.

Scholastic Achievements

- Secured All India Rank 113 among 100 thousand candidates in GATE 2021
- National Winner in the Cassini Scientist for a day 2013 competition held by NASA

Technical Skills

- **Programming Languages**: C, C++, Java, Python
- **Libraries and Frameworks**: WandB, PyTorch, Tensorflow, Keras, FairSeq, Sklearn, Numpy, Pandas.
- **Tools**: LaTeX, Android Studio, Packet Tracer, Action Orchestrator, Postman, ThousandEyes

Extra Curricular Activities

- Winner in Quiz competition at Aarunya 2k18 Cultural Fest at MEC.
- Winner in Algorithmiac, Blind Coding events at DhruvaMedha 2k18 Technical Fest at MEC.
- Selected in and successfully completed Technical and Leadership Development Program organized by CETI Foundation.
- Successfully completed the Leadership, Education and Development Camp organized by Direct Opinion Consultancy.