Abhishek Sureddy

☎ +1(413)-210-9556 • ✉ asureddy@umass.edu • in abhishek-sureddy • ⌂ home

Education

University of Massachusetts Amherst **Sep 2023 - May 2025**
Master of Science in Computer Science | CGPA: 4.0/4.0 *Amherst, MA*
Relevant Courses: Probability Theory, Algorithms in Data Science, Reinforcement Learning, Advanced NLP, Systems for DS, Computer Vision

Indian Institute of Technology Madras **Aug 2016 - Jun 2021**
B.Tech.(Hons) in Mechanical, M.Tech. in Data Science | CGPA: 9.38 / 10 *Chennai, India*
Relevant Courses: Probability, Statistics and Stochastic Processes, Multivariate Data Analysis , Applied Time Series Analysis, Machine Learning, Deep Learning, NLP, Multi-Armed Bandits, Design & Analysis of Algorithms

Skills

- **Programming:** Python, C/C++ | Familiar: Scala, Java, C#, q, R, sql, JS, HTML, Typescript
- **Software/ Frameworks:** Angular, Flask, FastApi, Git | Familiar: AWS, GCP, Docker, Matlab
- **Tools/ Libraries:** PyTorch, Transformers, Tensorflow, Scikit Learn, Pyspark, Kafka, LangChain, Open-cv, Nltk

Professional Experience

Morgan Stanley **Jun 2024 – Aug 2024**
Summer Associate Quantitative Finance *NewYork, USA*

- Part of Securities Lending desk of Prime Brokerage division.
- Developed a Q-based inventory management tool to monitor and analyze 11 Billion USD, in retail lendable inventory, enabling traders to assess PnL impact from lending to hedge funds.
- Developed analytics tools and metrics to assist traders and sales teams gain insights, boosting revenue by 5%.

Meta AI, FAIR Labs **Feb 2024 – Present**
Graduate Student Researcher [*arXiv*] *Remote, USA*

- Innovated an evaluation method to measure geographical biases in text-to-image models like Stable Diffusion and DALLE, enhancing precision and diversity with decoupled representations from Segment-Anything.
- Developed PatchViT, a novel technique for selecting relevant image patches, boosting feature extraction in Vision Transformer (ViT) models, outperforming standard ViT and CNN methods.
- Analyzed metric trends across regions, improving understanding of model performance under diverse prompts, and applied CLIP Zero Shot classification to real and generated datasets to assess regional biases.
- Awarded "Outstanding Paper" at the Trustworthy Multi-modal Foundation Models Workshop, ICML 2024.

Morgan Stanley **Jul 2021 - Jul 2023**
Full-time: Quant & ML Associate, Macrodatastrats - Interest Rate Strategies *Mumbai, India*

- Implemented a Retrieval Augmented Generation (RAG) based Q&A chatbot using LLMs and a vector DB to answer natural language questions on 200+ complex datasets, saving the data team 10+ hours weekly.
- Enhanced LLM response accuracy by 10% through Chain of Thought & few shot Prompting techniques.
- Built end-to-end tool to approximate joint probability surface of exchange rates given by pricing models using Neural Networks, reduced model runtime by 80%, achieved an MAE of 0.02%.
- Built a Scala and Python-based library to construct time series data of portfolio and Backtest various ML and trading strategies. Currently adapted by 50+ researchers and quants.
- Main Contributor in developing end-to-end framework for FRTB, regulatory requirement calculation of Interest Rate instruments like Swaps, Swaptions and FX options.
- Built an end-to-end web-based tool in Python and Angular to monitor, track, and alert on thousands of failed quality checks on datasets on a weekly basis. Currently used by 50+ starts.

Honeywell **Jun 2020 - Jul 2020**
Summer Internship in Software Engineering *Hyderabad, India*

- Worked on developing an end-to-end application to track and predict the position of various aircraft
- **Phase1:** Connected to iridium services mailbox, for receiving real-time position reports for flights as an email service. Decrypted and decoded the attachments to get the aircraft's position reports in java.
- **Phase2:** Ingested position reports to a Kafka topic and aggregated using Kafka streams. Finally dumped into a Database, created APIs are provided to access the position reports.
- **Phase3:** Developed a position prediction algorithm to predict the aircraft's position. Used Regression model on previous positions, used SGD to update the algorithm in real-time. Changed the pull-based mechanism (APIs) to a push-based mechanism using web sockets.
- **Management Role:** As a Scrum master, lead a team of 9 interns, Followed the Agile framework, split the team into 5 modules, and organized Daily stand-ups, and weekly integrated Demos.

GyanData **May 2019 - Jul 2019**
Summer Internship in Machine Learning *Chennai, India*

- **Phase1:** Developed a novel ML model to predict cricket scores at any point of the match. Achieved an MAE (maximum absolute error) of 32 runs in 25th over prediction and bias of 4 runs. Prediction is an ensemble of an analytical and Kernel ridge regression model.
- **Phase2:** Developed an optimization toolbox (including UI) that uses a stochastic sampling method. Implemented Genetic Algorithms, and multi-objective evolutionary algorithms for returning Pareto front solutions and capturing multi-optimal solutions in a single go. This toolbox was used for hyperparameter tuning of various Machine Learning Models.

Edar Labs **Dec 2018 - Jan 2019**
Winter Internship in Software Development *Chennai, India*

- Worked on Ideating and Developing the early version of an Augmented Reality application for visualizing educational concepts for primary and secondary school students.
- The application is currently being used by 10+ schools for teaching concepts using Augmented reality techniques.

Projects

LLM Alignment Towards Safety and Helpfulness **Feb 2024 – Jun 2024**
Guide: Dr.Mohit Iyyer *Umass Amherst*

- Increased the alignment and safety of LLaMA-2-7B by 40% using SFT, RAFT, DPO and distillation techniques.
- Employed PEFT methods like LoRA and QLoRA to fine-tune LLMs with $<$ 0.5% of total parameters.
- Implemented novel evaluation tasks like LLM as a Judge with sub-claim recall to evaluate model alignment.

Deep RL Algorithms Implementation **Oct 2023 – Dec 2023**
Guide: Dr.Bruno Castro da Silva *Umass Amherst*

- Implemented Reinforce with baseline, Semi-Gradient N-step SARSA, and Deep Q-Learning algorithms.
- Incorporated neural networks for policy and value functions. Conducted comprehensive evaluations of these algorithms on Cartpole, Acrobot, and custom Autonomous toy car environments, performing in-depth analysis.

Leapp.ai: Customized Learning Plan Generation using AI **Jun 2023 - Oct 2023**
Product link: https://leapp.ai *personal project*

- Created Leapp, a comprehensive web tool using advanced tech to facilitate personalized user learning plans.
- Implemented features like content streaming, collaborative sharing, and exploring public learning plans.
- Used prompt engineering to boost ChatGPT's output quality, elevating user satisfaction and engagement.
- Attained impressive user adoption: 10,000+ users, 1000+ learning plans, within 2 months of initial launch.
- Consistently gathered user feedback, iterated features for enhanced experience, and fueled growth.
- **skills:** Quart (Asyncio version of Flask, python), Angular, ChatGPT, AWS (DynamoDb, LightSail, Cognito)

Knibble.ai: Question Answering chatbot on custom Knowledge base **Mar 2023 - Jul 2023**
Product link: https://knibble.ai *personal project*

- Developed Knibble, an innovative generative AI-based web-tool to create a chatbot on custom knowledge base.

- Enabled seamless processing of diverse document formats (PDFs, URLs, text files) into a vector DB.
- Utilized Langchain for efficient text analysis, resulting in optimized chatbot performance and course content.
- Implemented embeddable chatbot for websites, Notion pages and web-crawler, thus enhancing user engagement.
- Achieved 1000+ users in 2 months, gaining recognition as top 1% tool for App Sumo Select Class of 2023.
- **skills:** Langchain (python), pinecone vector DB, python

Network Topology Reconstruction **Mar 2020 - Jul 2021**
Guide: Dr Nirav P Bhatt, Master's thesis *IIT Madras, India*

- Used text mining techniques to identify the entities and determine the relation between them.
- Generated a Pre-train dataset consisting of **20 million sentences** by parsing 2M PubMed abstracts.
- Finetuned **BioBert** and **BERT** on the Generated data and used it For NER and Relation Extraction tasks on various datasets like NCBI, GAD, DDI, BioRelEx, BioInfer.
- Ideated and developed the pipeline to generate a knowledge graph from the given text articles.

Machine Translation **Jun 2020 - Jul 2020**
Guide: Prof. C Chandra Sekhar, Course: Deep Learning *IIT Madras*

- Built a transformer-based encoder-decoder model for English to Hindi translation from scratch.
- Used Glove embeddings as word vectors for English and trained embeddings for Hindi.
- Achieved BLEU-1 score of 0.38 and BLEU-4 score of 0.06 using transformer model.
- Compared it with an LSTM-based encoder-decoder model with an attention mechanism.

Image Captioning with Attention **Mar 2020 - Jun 2020**
Guide: Prof. C Chandra Sekhar, Course: Deep Learning *IIT Madras*

- Developed an Encoder-Decoder-based Image captioning system on the flicker8k dataset from scratch
- Encoder is VGG features, Decoder is Single Hidden layer LSTM based RNN.
- The input to the Decoder is the feature from the VGG + previous hidden layer output of RNN
- Achieved BLEU-1 score of 0.54 and BLEU-3 score of 0.25.
- Used Glove pre-trained word embeddings in the embedding layer to reduce the training time.

Realtime movie rating prediction **Jan 2020 - May 2020**
Guide: Prof. Balaraman Ravindran, Course: Big Data Laboratory *IIT Madras*

- Devised pipeline to predict movie ratings on a real-time basis, by training on YELP (8M reviews) on GCP cluster.
- Achieved 70% accuracy, using only review text. (vector-space + NB classifier) in pyspark.
- Used the trained model to do real-time predictions on data streamed through a Kafka topic.

Fuzzy time-series modelling **Oct 2019 - Nov 2019**
Guide: Prof. Arun K Tangirala, Course: Applied Time Series Analysis *IIT Madras*

- Implemented a forecasting algorithm to predict the Relative Humidity using a fuzzy time-series model. The time-series data is from an Automated Weather Station at Sriharikota from May 15 - Jul 07, 2009, with hourly frequency.
- Data set comprises measurements of many meteorological variables ,viz., air temperature, wind speed
- Achieved an MAE of 2.5% when forecasted for the subsequent 4 days, i.e. 96 points.
- Compared it with linear SARIMA model using air temperature as an exogenous variable.

Awards and Achievements

- Secured All India Rank of 1732 in JEE Advanced 2016 out of 200 thousand candidates (Top 0.8%).
- Secured All India Rank of 784 in JEE Mains 2016 out of 1.3 million students (Top 0.06%).
- Selected provisionally for KVPY fellowship award among 100 thousand applicants (Top 1%).
- Secured Rank of 10 in Andhra Pradesh State Mathematical Olympiad 2012 among 100k applicants (Top 0.01%).