Sree Bhargavi Balija

☎ 858-319-6721 ✉ sbalija@ucsd.edu linkedin.com github.com portfolio Location: San Diego, CA

Education

University of California San Diego **June 2024**
Master of Science in Machine learning and Data science ***CGPA: 3.54/4.0***

Indian institute of Technology Hyderabad **July 2020**
Bachelors of technology in Engineering ***CGPA: 9.1/10***

Selected Publications

- Building Communication Efficient Asynchronous Peer-to-Peer Federated LLMs with Blockchain, **AAAI, Stanford Uni**
- FedNAM: Performing interpretability analysis in federated learning context, **ICLR 2025** (In review)
- CPTQuant - A Novel Mixed Precision Quantization Techniques for Large Language Models
- FedNAM+: Executing Interpretability Analysis using Novel Conformal Predictions method **CVPR 2025** (In review)

Job Experiences

Sensitivity based clustered Multi Head Attention, Meta Research (Collaboration) | *LLM'S* **May 2024 - Ongoing**

- Currently collaborating with the **US Meta research team** on improving the CHAI paper by integrating Hybrid sparse attention mechanisms and Targeted fine-tuning with quantization of cluster heads.
- Achieved significant memory reduction of 70% in transformer models by implementing hybrid sparse attention mechanisms and advanced clustering techniques, optimizing large-scale NLP tasks..

Artifical Intelligence Intern, Radical AI | *LLM'S, Python, Gen AI* **April 2024 - June 2024**

- Developed and deployed AI applications utilizing leading frameworks such as Langchain, OpenAI, **Google Gemini**.
- Implementing **multi-modal** interactive elements and AI-powered games to create an dynamic learning environment

Software Engineer, ServiceNow | *Java, Js, Angular, Eclipse, Github* **June 2020 – August 2022**

- Worked on integrating multiple rest api's with ITSM workflows for adding capabilities like **Citrix cloud virtual systems access**, Requested item flow to the **Virtual bot** and developed the **NLU** models for **Conversational AI**
- Designed and developed the **Dashboard** which provides a **prebuilt analytics** for 8 metrics like customer satisfaction score, cost savings etc to demonstrate the **actual business value** achieved through the **top ServiceNow products**.
- Implemented **Java API** for periodic and user triggered compaction, job cancellation and managing compaction statistics

Academic Projects & Research Experience

Federated fine tuning of Heterogeneous Large Language Models | *Pytorch, Flower, Python* [code] **Dec 2023**

- Developed an **NLP** bot by adapting the LLaMA (Language Model for Many Applications) model for medical inquiries. This process entailed a systematic method that included grasping the model's architecture, readying the dataset for fine-tuning, and ultimately deploying the bot on a website using **Django**.
- This framework addresses the **privacy, data scarcity issues** and specifically applicable for **NLP tasks**.

Novel mixed precision quantization technique for Large Language Models | [code] **Sep 2023 - April 2024**

- Developed three novel mixed precision quantization techniques (CMPQ, PMPQ, TDMPQ) for LLM's like Gemini, Llama2 which out performs the **SOTA techniques** in terms of compression ratio by 4X times
- Developed new client pruning method using **conformal predictions** which selects the most efficient clients for high global model performance.

Development of Binarized State-Space Models for Efficient Deep Learning (Mamba Project) | *Python* [code]**Ongoing**

- Engineered the Mamba framework to introduce binarization in state-space models, reducing memory usage by over 70% and improving inference speed by 2.5x. Optimized binarization across projection, convolutional, and state-space matrices, achieving accuracy within 1% of full-precision models for tasks like PIQA and BoolQ.

Technical Skills

Languages: C/C++, Python, R, Java, Javascript, Angular, Fortran, Prolog, Perl, JavaScript, Kotlin, Swift
Frameworks: Pytorch, TFRS, ETL, Kubernetes, NLTK, **RAGTool**, Stanford NLP, Go, Tableau, Cors, MLib
Databases: Elasticsearch, GCP, Firebase, **Microsoft SQL Server**, MySQL, Hive, Azure Databricks, Snowflake
Data Science: LightFM, Django, Classical ML, DL, NLP, Explainable AI, Federated learning, CV

Accolades/ Online Certifications

- Achieved Skill development incentive program award, **ServiceNow** **2021**
- **Teaching Assistant**, Introductory courses in physics and chemistry departments website management **2018**
- **Academic excellence award**, IIT Hyderabad **2018**
- Silver medal, International Master **Mathematics Olympiad** **2013**
- UCSD ECE **Summer research internship** scholar, UCSD **2023**
- Building LLMs Efficiently, **NeurIPS 2024 Challenge** **Ongoing**