SAN DIEGO, US • SBALIJA@UCSD.EDU • 858-319-6721

SREE BHARGAVI BALIJA

Machine Learning and Data Science

RESEARCH SUMMARY

Aspiring PhD researcher with a strong foundation in Machine Learning and Data Science, specializing in federated learning, interpretability, and large language models, with a focus on building ethical and responsible AI systems.

EMPLOYMENT HISTORY

SOFTWARE ENGINEER **Mar 2025 - Present**
Akdene Technologies

- Leading the ML team at Akdene Technologies, driving innovation in customer data analytics for next-gen financial planning solutions.
- Specialized in building scalable ML pipelines and deploying predictive models that optimize financial decision-making.

RESEARCHER **Apr 2024 - Feb 2025**
US Meta Research Team

- Pioneered an optimized Hybrid Sparse Attention framework, leveraging sensitivity-based clustering to enhance transformer efficiency while maintaining state-of-the-art performance.
- Achieved a groundbreaking 50% reduction in memory footprint for transformer models by implementing advanced clustering techniques and sparsification, significantly accelerating large-scale NLP tasks.

CONTRIBUTOR **Nov 2024**
ML Commons

- Led the development of multimodal projects, translating video to text
- Pioneered transformer-based models, enhancing automated video understanding and summarization

ARTIFICIAL INTELLIGENCE INTERN **Apr 2024 - Jun 2024**
Radical AI

- Led AI application development using top-notch frameworks like Langchain, OpenAI, Google Gemini.
- Engineered multi-modal interactive elements and AI-powered games, fostering an engaging learning environment

SOFTWARE ENGINEER **Jun 2020 - Aug 2022**
ServiceNow

- Worked on integrating multiple REST APIs 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.

EDUCATION

MASTER OF SCIENCE IN MACHINE LEARNING AND DATA SCIENCE **Jun 2024**
University of California San Diego

BACHELORS OF TECHNOLOGY IN ENGINEERING **Jul 2020**
Indian Institute of Technology Hyderabad

SKILLS

Python, Java, JavaScript, Angular, C++, R, Fortran, Prolog, Perl, Kotlin, Swift, PyTorch, Kubernetes, NLTK, NLP,
Machine Learning, Deep Learning, Federated Learning, SQL, Django, ETL, Azure, GCP, Elasticsearch,
AWS (EC2, S3, VPC, IAM, RDS, Lambda, ECS/EKS, CloudFormation), Terraform, Kubernetes,
Docker and Microservices Architecture, CI/CD Pipelines , Load Balancing and Auto Scaling, Digital Twin Familiarity.

ADDITIONAL INFORMATION

PUBLICATIONS

- Building Communication Efficient Peer-to-Peer Federated LLMs with Blockchain, AAAI, Stanford University
- CPTQuant - A Novel Mixed Precision Quantization Techniques for Large Language Models
- FedNAM+: Executing Interpretability Analysis using Novel Conformal Predictions method, CVPR 2025 (In review)
- FedNAM: Executing Interpretability Analysis in Federated learning Context
- AILUMINATE: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons

AWARDS AND ACHIEVEMENTS

- Skill development incentive program award, ServiceNow 2021
- Academic excellence award, IIT Hyderabad 2018
- Founder of AutoPatch+, presented at MIT AI summit 2025
- Developed Multimodal small LLM of size 125M parameters

RESEARCH PROJECTS

DIFFERENTIAL PRIVACY - MULTIMODAL CLINICAL DATA

- Prof. Praneeth Vepakomma, MIT 2025

- Developed a patient-specific differential privacy pipeline by generating synthetic clinical reports, applying DP-Prompt paraphrasing, and training privacy-preserving embeddings.
- Organized and optimized embeddings per patient to enable scalable privacy analysis across medical datasets.

FOUNDATION MODELS

- Prof. Debashis Sahoo, UCSD 2024

- Introduced a novel fine-tuning method that adapts tensor representations during training to minimize quantization-induced errors, enabling robust performance in resource-constrained environments.
- Utilized quantum annealing to dynamically assign mixed precision levels to tensors, balancing computational efficiency and model accuracy in Large language models.
- 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.