Sree Bhargavi Balija
Software Engineer

CA • sbalija@ucsd.edu • **(858) 319-6721**• Google Scholar • Linkedin • Github - Sreebhargavibalijaa • Portfolio

SUMMARY

- Machine Learning Researcher & Engineer with 3+ years of experience in scalable ML systems, federated learning, and generative AI, backed by 8 published papers (including work collaborated with **MIT Media Lab**) and ongoing research with **Stanford University's** Prof. Ludwig on **Terminal bench** Project.
- Built RAG systems and fine-tuned GPT-4/LLaMA-2 using LangChain, Hugging Face, and Pinecone/Weaviate. Automated workflows with multi-agent frameworks, cutting manual effort by 40%+ via document AI and autonomous agents.
- Spearheaded **quantitative modeling** (XGBoost, LightGBM, Prophet) and privacy-preserving AI via federated learning (TensorFlow **Federated**, Flower). Deployed event-driven microservices (**Kafka, Pulsar**) for real-time fraud detection and risk analytics, improving prediction accuracy by 25% while ensuring SOC-2/GDPR compliance.
- Designed end-to-end MLOps pipelines with CI/CD (**GitHub Actions, ArgoCD**), model monitoring (Evidently, Prometheus), and feature stores (**Feast**). Enhanced latency-critical APIs (Python/Java) using FastAPI, **Spring Boot**, and gRPC, achieving 99.9% uptime for systems serving 100K+ users with **Elasticsearch/Cassandra** for low-latency analytics.

PROFESSIONAL EXPERIENCE

Software Engineering Intern **USA**
Fujitsu Research of America **Aug 2025 – Current**

- Optimizing GPU utilization in the ACB project by reducing idle time through workload balancing and scheduling strategies.
- Implementing multi-server configurations to enhance parallel processing efficiency and minimize resource underutilization.
- Exploring integration of agentic AI technologies into the Kojuchi platform to enable autonomous decision-making and workflow orchestration.

Machine Learning Researcher (Fellowship) **USA**
MIT + Radius Collaboration, USA **July 2025**

- Designed and prototyped a **Radius Testnet trust and discovery layer** for agent marketplaces, integrating **smart contract–based reputation** and ranking mechanisms.
- Implemented **MCP server monetization workflows**, including payments, invoicing, and protected deployments with EVMAuth and TypeScript SDK adapters.
- Advanced research on **agentic AI ecosystems**, focusing on decentralized discovery, trust, and microtransaction-based interoperability for autonomous agents.

Machine Learning Engineer **USA**
Akdene Technologies **Mar 2025 – July 25**

- Developed an AI-powered financial risk engine using TensorFlow, XGBoost, and PyTorch Geometric, deploying real-time fraud detection and credit scoring for 100K+ customers in <3 months
- Built an Agentic AI chatbot using GPT-4o, RAG, and AutoGen on Django, cutting support workload by 120+ hrs/month and slashing response time by 65% via real-time semantic caching and multi-agent automation.
- Deployed on serverless AWS with auto-scaling Kubernetes, achieving 99.9% uptime while reducing API costs by 30% through LoRA fine-tuning and Redis Vector Search.
- Enhanced fraud detection with PyTorch and model interpretability techniques, identifying 1,200+ high-risk transactions in the first weeks of deployment.
- Formulated a model monitoring pipeline using Conformal Predictions, ensuring 75% stability in model accuracy across financial datasets.

Software Engineer **India**
ServiceNow **Jun 2020 – Aug 2022**

- Deployed a scalable backend microservice using Java, Spring Boot, and REST APIs within 6 months, improving data processing speed by 3x for enterprise clients.
- Implemented a secure authentication system using OAuth2, API Development, and system integration, reducing unauthorized access incidents by 60% in the first year.
- Accelerated cloud-based deployment on GCP, reducing infrastructure costs by 25% while maintaining 79% uptime.

Advanced Transformer Optimization for CHAI Model

Lead Researcher – US Meta Research Collaborator **Apr 2024**

- Pioneered a cutting-edge hybrid sparse attention mechanism combined with precision fine-tuning across three domain-specific datasets, enhancing task accuracy by 15% while slashing inference time by 50%.
- Engineered ultra-efficient transformer pipelines with revolutionary memory optimization, reducing GPU memory consumption by over 4GB per training cycle and enabling seamless deployment on resource-constrained systems.
- Innovated AI-driven clustering techniques to streamline token embeddings, reducing pre-processing overhead by 40% and dramatically accelerating end-to-end model training.

RESEARCH PROJECTS

Differential Privacy of Multimodal Clinical Data - Prof. Praneeth Vepakomma **April 25 – July 25**

- 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.

EDUCATION

Master of Science in Machine Learning and Data Science
University of California, San Diego, CA ***June 2024***

Bachelor's in Engineering
IIT Hyderabad, India ***June 2020***

TECHNICAL SKILLS

- **Programming Languages** Python (Expert), Java, JavaScript, C++, R, Fortran, Prolog, Perl, Kotlin, Swift
- **Machine Learning & AI** Machine Learning, Deep Learning, NLP, Federated Learning, Transformers, Multimodal AI, Model Interpretability, Conformal Predictions
- **AI/LLM Tools** NLTK, Langchain, OpenAI, Google Gemini
- **ML Libraries & Frameworks** Scikit-learn, TensorFlow, Keras, XGBoost, PyTorch, Django, Langchain
- **Data Analysis & Visualization** Pandas, NumPy, Matplotlib, Seaborn, Dashboards, Pre-built Analytics Metrics
- **Cloud & DevOps** Azure, GCP, Kubernetes, Git, CI/CD, Technical Documentation
- **Databases & Data Tools** SQL, Elasticsearch, ETL
- **API & Backend Development** REST APIs, API Development, Java APIs, Microservices, System Integration
- **Development Practices** Agile, System Design, Testing, Model Monitoring, MLOps

PUBLICATIONS

- Building Communication Efficient Peer-to-Peer Federated LLMs with Blockchain, **AAAI, Stanford University** (link)
- Explore-Then-Commit: The Optimal Strategy for Scientific Breakthrough Discovery, **AAAI 2025**
- Fortifying the Agentic Web: A Unified Zero-Trust Architecture Against Logic-layer Threats (link)
- AILUMINATE: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons (link)
- An Index-based Federated Registry Architecture for Secure, Capability-Aware Agent Discovery: Leveraging NANDA and ANS (link)
- CPTQuant - A Novel Mixed Precision Quantization Techniques for Large Language Models (link)
- Decoding Federated Learning: The FedNAM+ Conformal Revolution (link)
- The Trust Fabric: Decentralized Interoperability and Economic Coordination for the Agentic Web (Link)

AWARDS AND ACHIEVEMENTS

- Selected as Research Fellow for **MIT NANDA Radius Fellowship program 2025**
- **Skill development incentive program** award, **ServiceNow 2021**
- **Academic excellence award**, IIT Hyderabad 2018
- Selected as a peer reviewer for **NeurIPS 2025**
- Project accepted to the 2025 **NSF AI-SDM** Workshop on Human-AI Complementarity for Decision Making
- Co-Founded **AutoPatch+,** a next-gen AI code validation platform, and pioneered hallucination detection tech for generative coding, showcased at MIT AI Summit 2025 - AutoPatch+
- Developed Multimodal small LLM of size 125M parameters - PixPrompt
- Awarded **a Research Fellowship** with EleutherAI Summer of Open AI Research.
- Granted **a Full Scholarship and Research Fellowship** with Algoverse Open AI Research 2025.
- Chosen **as a Full Scholarship Recipient** for the PEARC25 Student Program 2025.
- Accumulated **28+ citations across 7 research publications** on Google Scholar