SAI TEJA SRIVILLIBHUTTURU

(682) 234-3567 | saiteja.srivilli@gmail.com | linkedin.com/in/saitejasrivillibhutturu | saitejasrivilli.github.io

4+ Years Software Engineering | 1+ Year ML/AI Research

EDUCATION

University of Texas at Arlington, Master of Science in Computer Science **Aug 2023 – May 2025**
Andhra University, Bachelor of Technology in Computer Science **Jun 2015 – Apr 2019**

SKILLS

LLM & Training: PyTorch, TensorFlow, Hugging Face, vLLM, LoRA/QLoRA, PEFT, CUDA, Distributed Training (DDP/NCCL), Quantization (GPTQ/AWQ), Speculative Decoding
Evaluation & Agents: RAGAS, BERTScore, HumanEval, Red-Teaming, LangChain, LangGraph, Multi-Agent Pipelines, RAG, Pinecone, Prompt Engineering
CV & Deep Learning: CNN, Supervised Learning, Image Augmentation, CLIP, PyTorch
Cloud & MLOps: AWS (SageMaker, EC2, S3, Lambda, ECR), Docker, Kafka, MLflow, CI/CD
Software Engineering: Python, Java, SQL, Spring Boot, FastAPI, REST APIs, Microservices
Simulation & RL: SUMO/TraCI, DDQN, Actor-Critic, Sionna 6G

EXPERIENCE

DentalScan (ReplyQuick AI LLC) | *Machine Learning Engineer Intern* **Dec 2025 – Present**

- **Designed** CNN-based supervised ML training pipelines on AWS SageMaker, S3, EC2, and Lambda for intra-oral image classification across 6 clinical categories on a 50K+ labeled dataset.
- **Improved** weighted F1 from 0.74 to 0.89 across 6 diagnostic categories by executing 15+ iterative SageMaker training runs with systematic per-category error analysis, image augmentation, and class-balancing strategies.
- **Containerized** inference endpoints via Docker and REST APIs and tracked all experiments using MLflow with version-controlled model registry and automated evaluation gates, compressing release cycles to same-day.
- **Identified** an 18% minority-class recall gap through confusion-matrix error analysis and rebuilt the S3/Lambda augmentation pipeline with weighted sampling and targeted synthetic generation.
- **Scaled** the dataset ingestion pipeline to 50K+ images using AWS ECR, SQS, Step Functions, and DynamoDB for metadata tracking, maintaining reproducible checkpoints and sustaining 0 regression incidents across releases.

University of Texas at Arlington | ***Graduate Research Engineer,*** Arlington, TX **Dec 2025 – Present**

- **Architected** a deep Stackelberg MARL system in SUMO via TraCI, modeling congestion pricing as a two-level leader-follower problem across 3x3, 5x5, and 7x7 urban grids.
- **Implemented** Dueling DDQN and A2C Actor-Critic agents in PyTorch with replay buffers, gradient clipping, and entropy regularization, with toll actions physically modulating SUMO edge efforts per step.
- **Quantified** efficiency-equity tradeoffs using 6 fairness metrics (Gini, Theil, Atkinson, HorizEquity, CV, PoE) across 4 vehicle classes with calibrated VoT weights over 3 seeds with 95% CI.
- **Modeled** a digital twin with Gaussian sensor noise, observation delays, time-varying tolls, and BPR-calibrated travel-time functions, validating robustness across demand variance from 0.05 to 0.5.
- **Benchmarked** DDQN and Actor-Critic against no-toll, marginal-cost, and static-toll baselines, producing 32 statistical figures including convergence curves, fairness radar charts, and sensitivity plots.

University of Texas at Arlington | ***Graduate Research Assistant****, Arlington, TX* **Jun 2025 – Nov 2025**

- **Fine-tuned** LLMs on 3+ domain-specific textbooks using PyTorch, Hugging Face, and PEFT/SFT, cutting domain adaptation time by 40% compared to full fine-tuning baselines.
- **Indexed** 1,000+ research paper chunks in Pinecone with embedding-optimized chunking strategies, achieving 85%+ retrieval relevance via RAGAS at sub-200ms latency.
- **Engineered** a Kafka-based async document ingestion pipeline on AWS EC2 processing 200+ papers/hour with Dockerized FastAPI environments, health checks, and Prometheus latency monitoring.
- **Benchmarked** 5+ LLM configurations using BERTScore and ROUGE metrics, identifying a 3x cost-quality tradeoff gap that reprioritized the research roadmap for a team of 8.

University of Texas at Arlington | ***Graduate Teaching Assistant,*** Arlington, TX **Aug 2024 – May 2025**

- **Architected** an LLM-driven path planning system integrating OpenStreetMap real-time geolocation with Sionna 6G channel simulation, fine-tuning GPT-4o on 10K+ Dijkstra-generated routing samples.
- **Benchmarked** inference latency, route accuracy, and token efficiency across 5+ LLMs on OpenStreetMap-derived routes using ablation studies and A/B testing, reducing Sionna simulation incident recurrence by 3x.
- **Addressed** 31 peer-reviewer concerns in the IEEE OJCOMS revision including outage probability analysis and alpha-mapping sensitivity studies, resulting in acceptance at IEEE ICC 2026.

Tata Consultancy Services | *Software Engineer ,Chennai,India* **Jun 2019 – May 2023**

- **Designed** a microservices-based middleware platform using API Gateway, Circuit Breaker design patterns in Java and Spring Boot, connecting 5+ distributed financial systems via REST/SOAP APIs handling 10K+ transactions.
- **Reduced** manual processing overhead by 40% across 3 operational teams by engineering automated ETL pipelines in Python and Spark processing 50K+ records at 99.8% data integrity.
- **Optimized** high-volume SQL queries and Spark jobs improving throughput by 35%, enforcing TDD with 85%+ JUnit/Mockito coverage and CI/CD via GitHub Actions, cutting defects by 25%.
- **Mentored** 3 junior engineers on Saga, Circuit Breaker, and TDD with JUnit/Mockito, and led sprint planning for a 5-engineer team using Git-based workflows, delivering 2 consecutive milestones 2 weeks ahead of schedule.

PROJECTS

ML / Computer Vision

- **Multi-Object Tracking** (SORT/DeepSORT, MOTA/MOTP, PyTorch, TraCI): Reproduced and extended SORT/DeepSORT tracking pipelines for autonomous vehicle perception, benchmarking MOTA/MOTP improvements across occlusion scenarios with supervised learning baselines.
- **Foundation Model Fine-Tuning** (LLaMA, Mistral, LoRA, QLoRA, PEFT): Ran systematic fine-tuning experiments across LoRA, QLoRA, and full fine-tuning on LLaMA and Mistral, capturing data efficiency and loss convergence tradeoffs.
- **Amazon Hybrid Recommender** (Collaborative Filtering, ETL, NDCG, A/B Testing): Built a hybrid collaborative filtering recommender on Amazon data with ETL pipelines via NDCG, MRR, and A/B testing.

LLM / AI Systems

- **Distributed LLM Pre-Training** (PyTorch DDP, NCCL, CUDA): Built a production-grade training system achieving 3.50x speedup on 4 GPUs at 87.5% parallel efficiency, processing 152K tokens/second with fault-tolerant checkpointing and real-time P95 dashboards.
- **vLLM Throughput Benchmark** (vLLM, ONNX, Speculative Decoding, GPTQ/AWQ): Achieved 18.6x throughput gains over Hugging Face via speculative decoding and GPTQ/AWQ quantization at sub-100ms P95 latency.
- **Multi-Agent Research Assistant** (LangGraph, LangChain, ChromaDB, FastAPI, RAGAS): Implemented a 4-stage multi-agent pipeline (Researcher, Critic, Synthesizer, Evaluator) with RAGAS evaluation, ChromaDB RAG retrieval, FastAPI backend, and Gradio frontend with CI/CD via GitHub Actions.
- **LLM Code-Agent Eval Benchmark** (Groq, Gemini, HumanEval, BERTScore): Developed sandboxed evaluation infrastructure measuring pass@1 and BERTScore across 164 HumanEval tasks and 3 models (Groq, Gemini, GPT) with ROUGE-based output quality scoring and structured failure taxonomy reporting.
- **Red-Teaming and Safety Evaluation** (LLMs, Diffusion Models, Ablation Studies): Conducted structured red-teaming surfacing jailbreaks, hallucinations, and adversarial vulnerabilities across LLMs and diffusion models, producing ablation-backed risk-tiered failure taxonomies and mitigation strategies.
- **TeluguGPT** (GPT, SFT, Low-Resource NLP, Jupyter): Adapted a generative language model on Telugu corpus using SFT for 80M+ speakers, addressing tokenization, script normalization, and cultural context challenges across Linux-based GPU training environments.

PUBLICATIONS

CTMap: Digital Twin-Guided AI Path Planning for Connectivity-Aware Mobility - IEEE ICC 2026 (Accepted), IEEE OJCOMS Under Review

CERTIFICATIONS

Advanced Large Language Model Agents - UC Berkeley EECS, AWS Certified Data Engineer – Associate Microsoft Certified: Fabric Data Engineer Associate, AI Evals for Everyone, Oracle GenAI Professional, Oracle AI Vector Search Certified Professional, Salesforce Agentforce Specialist, Salesforce Certified AI Associate