

SAI TEJA SRIVILLIBHUTTURU

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

TECHNICAL SKILLS

Languages: Python, Java, JavaScript, SQL

ML & Deep Learning: PyTorch, TensorFlow, Hugging Face Transformers, vLLM, CUDA, ONNX Runtime, LoRA, RAG, CNNs

LLMs & Agents: OpenAI API, LLM Fine-Tuning, BERT, CLIP, Quantization (GPTQ, AWQ), LangGraph, ReAct, RAGAS, BERTScore

Full-Stack & APIs: React.js, Next.js, Flask, FastAPI, Spring Boot, RESTful/SOAP APIs, Kafka, Microservices

Data & Databases: PostgreSQL, Pinecone, ChromaDB

Cloud & DevOps: AWS (EC2, S3, Lambda, SageMaker, ECR), Docker, GitHub Actions, CI/CD, Git, Linux

EXPERIENCE

DentalScan (ReplyQuickAI LLC) Remote

Machine Learning Engineer Intern Dec 2025 – Present

- Architected end-to-end ML training pipelines on AWS (S3, EC2, Lambda, SageMaker) for CNN-based intra-oral image classification across 6 clinical categories on a 50K+ labeled dataset, implementing continuous retraining with dentist-corrected feedback loops
- Elevated model performance from F1 0.74 to 0.89 by containerizing inference endpoints with Docker and REST APIs, conducting systematic per-category error analysis across 15+ experiment iterations with version-controlled tracking
- Engineered automated dataset ingestion, augmentation, and class-balancing pipelines scaling from 50K to 100K+ images with clinical-grade validation checkpoints and dataset versioning
- Collaborated with backend engineers to deploy and maintain production ML endpoints, implementing experiment tracking, model registry, and automated evaluation gates for clinical validation before production rollout

University of Texas at Arlington Arlington, TX

Graduate Research Assistant — TopGPT (Full-Stack RAG Platform) Jun 2025 – Present

- Fine-tuned large language models on 3+ domain-specific textbooks using PyTorch and designed a RAG pipeline indexing 1,000+ research paper chunks in Pinecone, achieving 85%+ retrieval relevance with sub-200ms query latency
- Deployed scalable research infrastructure on AWS EC2 with Kafka-based async document ingestion (200+ papers/hour), Docker-containerized environments, health checks, and request-level latency monitoring
- Developed the full-stack web interface (React.js) integrating the RAG pipeline with interactive search and exploration UI for real-time querying across textbook and paper-level knowledge bases

Graduate Teaching Assistant — CTMap (6G Network Routing) Aug 2024 – May 2025

- Proposed CTMap, a novel digital-twin + LLM framework for connectivity-aware mmWave routing: integrated OpenStreetMap real-time location data with Sionna 6G channel simulation, fine-tuned LLMs on 10K+ Dijkstra-generated optimal paths to predict signal-aware routes
- Reduced inference latency by 60% and improved routing accuracy by 25% over baseline by optimizing LLM fine-tuning on graph-generated path datasets using PyTorch for real-time network simulation
- Built evaluation and benchmarking infrastructure measuring inference latency, route accuracy, and token efficiency across 5+ LLM architectures, establishing reproducible comparison baselines
- Co-authored peer-reviewed research published on [arXiv](https://arxiv.org), presenting the CTMap framework for LLM-driven signal-aware path optimization in next-generation 6G networks

Tata Consultancy Services Chennai, India

System Engineer (promoted from Assistant System Engineer) Jun 2019 – May 2023

- Designed microservices-based middleware using API Gateway and Circuit Breaker patterns, connecting 5+ distributed financial systems via RESTful/SOAP APIs handling 10K+ daily transactions, reducing manual processing by 40%
- Modernized on-prem systems into cloud-ready services with automated ETL pipelines for 50K+ records at 99.8% integrity; owned on-call operations sustaining 99%+ availability across quarterly close periods
- Drove engineering quality through code reviews, CI pipelines, and testing suites (JUnit/Mockito, 85%+ coverage), reducing post-release defects by 25%
- Mentored 3 junior engineers on microservices best practices and led sprint planning for middleware integration workstreams within the 5-engineer team

PROJECTS

vLLM Throughput Benchmark | Python, vLLM, Hugging Face, CUDA github.com/saitejasrivilli/vllm-throughput-benchmark

- Benchmarked vLLM vs Hugging Face inference under concurrent load, achieving 18.6× throughput gains via dynamic token batching and KV-cache optimization with sub-100ms P95 latency and 3.6× lower cost per million tokens

LLM Code Agent Eval Benchmark | Python, Groq, Gemini, HumanEval github.com/saitejasrivilli/code-agent-eval-benchmark

- Built evaluation infrastructure for LLM coding agents across 164 HumanEval tasks and 3 models, measuring pass@1, execution accuracy, and failure taxonomy with sandboxed execution and automated test harness

Multi-Agent Research Assistant | LangGraph, ChromaDB, FastAPI, RAGAS, Docker github.com/saitejasrivilli/multi-agent-research-mcp

- Designed a multi-agent pipeline using LangGraph (Researcher → Critic → Synthesizer → Evaluator) with RAGAS evaluation, ChromaDB RAG, FastAPI backend, and CI/CD via GitHub Actions

Multi-Strategy AI Agent System | Python, Groq LLM, Tavily, ChromaDB, Gradio github.com/saitejasrivilli/ai-agent

- Built production-ready AI agent with 4 reasoning strategies (CoT, ToT, ReAct, Multi-Agent), LLM-based auto-classifier routing, web search, vector memory, rate limiting, and real-time streaming

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