

CHETAN SAI BORRA

979 575 9026 | chetansai2003@tamu.edu | [LinkedIn/chetan-sai-b](https://www.linkedin.com/in/chetan-sai-b) | [github/Chetansai11](https://github.com/Chetansai11)

Education

Texas A&M University, College Station, TX

Aug 2024 - May 2026

Master of Science, Computer Engineering (GPA: 4.0/4.0)

College Station, TX

• **Coursework:** Machine Learning, Parallel Computing, Reinforcement Learning, Computer Vision & Robot Perception

Experience

Texas A&M Utilities & Energy Services | *Machine Learning Engineer*

Feb 2025 - May 2026

- Engineered high-throughput time-series ML pipelines in Python and SQL for continuous energy anomaly detection, cutting fault response latency by 40% across campus utility systems.
- Deployed LSTM models in TensorFlow and Prophet ensemble models for real-time energy telemetry analysis, attaining 95% precision in detecting transient grid instabilities across live production infrastructure.
- Developed end-to-end MLOps pipelines for data validation and Feature Stores in Python, supporting predictive maintenance model deployment and cutting unplanned equipment downtime across utility systems.
- Containerized ML inference services via Docker and Kubernetes, optimizing SQL and Python validation layers to sustain sub-100ms latency and 99.9% uptime for real-time utility fault monitoring.

Research Assistant, Deep Learning

Aug 2023 - May 2024

- Designed and trained PyTorch U-Net and CNN architectures for volumetric MRI segmentation, reaching 90% pixel-level accuracy and 72% mean IoU on multi-class clinical datasets.
- Fine-tuned Vision Transformer models in PyTorch using systematic hyperparameter optimization for model evaluation, including learning-rate scheduling and weight decay, that increased cross-dataset segmentation IoU by 4% and created a reproducible benchmark for future experiments.
- Accelerated PyTorch model training throughput via FP16/BF16 mixed-precision and GPU-optimized preprocessing pipelines in OpenCV and Bash, shrinking per-epoch training time on datasets exceeding 50GB.
- Benchmarked five leading segmentation architectures against Scikit-learn classical statistical baselines, authoring reproducible evaluation protocols that were adopted as the lab standard for clinical AI experimentation

Technical Skills

- Generative AI & Agents:** LangGraph, LangChain, LlamaIndex, RAG, GraphRAG, HuggingFace Transformers, LoRA/QLoRA Fine-Tuning, MCP (Model Context Protocol), NVIDIA NIM
- Machine Learning & Vision:** PyTorch, TensorFlow, XGBoost, Scikit-learn, OpenCV, Diffusion Models, CNNs
- AI & MLOps:** Docker, Kubernetes, Terraform, MLflow, CI/CD, GitHub Actions, Triton Inference Server, Model Quantization
- Data & Cloud:** FAISS, Pinecone, Pandas, Polars, NumPy, AWS (S3, EC2, SageMaker, Bedrock), PostgreSQL, Azure AI, Azure DevOps
- Languages & Core Tools:** Python, C/C++, Java, Bash, SQL, Git, CUDA

Certifications

- Microsoft Certified: Azure AI Apps and Agents Developer Associate

Projects

Autonomous Multi-Agent Network SRE Platform

- Architected a self-healing multi-agent system with LangGraph and coordinated tool-calling to execute Root Cause Analysis (RCA) and anomaly detection across distributed telecom network sites.
- Engineered an RAG-based orchestration layer for autonomous hypothesis testing across 100+ sites, slashing Mean Time to Detection (MTTD) by 35% and lifting signal integrity monitoring precision by 30%.
- Deployed NVIDIA NIM-accelerated LLM inference for real-time conversational analytics over live network telemetry, surfacing multi-layer KPI correlations for Tier-1 operator workflows.
- Launched a Dockerized Gradio interface with Human-in-the-Loop (HITL) feedback loops, introducing operator review cycles cut false-positive escalations in agentic RCA decisions.

Production Grade Data Analysis & MLOps Platform

- Designed a Medallion Architecture on AWS S3 with Polars for high-performance data curation, delivering 5x faster data loading and accelerating real-time feature engineering for fraud detection at scale.
- Trained an XGBoost classifier with Focal Loss to address a 96.5:1 class imbalance, securing 0.874 ROC-AUC and maintaining high recall on minority fraud cases in production-grade validation.
- Established a full CI/CD MLOps pipeline via GitHub Actions and Terraform, wiring in continuous drift detection to gate model quality before every production deployment.
- Scripted data ingestion and validation in Bash and Python, integrating MLOps pipelines with existing CI/CD infrastructure to support continuous model retraining on production data streams.

Multi-Agent AI Lawyer & Audit System

- Engineered an asynchronous multi-agent pipeline with Python AsyncIO to parallelize LLM tool-calling, trimming end-to-end inference latency for RAG-based legal document workflows.
- Shipped a full-stack FastAPI/React application serving agentic AI assistants, applying heuristic-based validation to guarantee consistent fact extraction across legal intake documents.
- Authored a real-time templated brief generator via Model Context Protocol (MCP) for local database integration, eliminating manual documentation steps and cutting intake preparation time.
- Standardized engineering quality by integrating Pytest and Ruff into a CI/CD workflow, eliminating technical debt and raising code maintainability across all modules.

Achievements

- 2nd Place, Aggie Invent for the Planet:** Prototyped a real-time computer vision system on aerial imagery with object detection to monitor agricultural runoff, placing 2nd among 30+ competing teams.
- Best Use of MongoDB, TAMU Datathon:** Developed a RAG-based CLI agent with Gemini Pro and MongoDB.