Saathwik Mailapalli

Visakhapatnam,India | saathwik43@gmail.com | saathwikm.me | github.com/Saathwik43 | linkedin.com/in/saathwik-mailapalli

Professional Summary

MLOps-focused Data Science Engineering student with hands-on experience deploying quantized LLMs, building NLP classification systems, and architecting secure containerized AI infrastructure. Strong foundation in Python, Docker, Linux, and scalable model deployment. Experienced in building production-ready ML pipelines and self-hosted AI systems under resource constraints.

Technical Skills

Languages: Python, C++, SQL, JavaScript, HTML/CSS
Machine Learning: Scikit-learn, TensorFlow, PyTorch, NLTK
LLM & AI Systems: Quantization (GGUF), Ollama, OpenAI APIs, mem0
Backend & APIs: Flask (REST APIs), Chrome Extensions
Infrastructure & DevOps: Docker, Linux, Caddy, Tailscale, Git
Databases: MySQL

Projects

Real vs Fake Job Classifier (NLP + Full-Stack Deployment)

- Built NLP-based fraud detection model trained on 17,000+ job listings using TF-IDF and Multinomial Naive Bayes.
- Engineered feature extraction and preprocessing pipeline to optimize classification performance.
- Deployed model via Flask REST API with Chrome Extension frontend for real-time inference.
- Designed low-latency prediction system achieving response times under 10-50 ms.

Self-Hosted AI Lab (MLOps & Infrastructure Engineering)

- Deployed Mistral 7B and Phi LLMs locally on 8GB RAM Linux server using GGUF quantization.
- Reduced memory footprint by approximately 70-75% enabling stable local inference.
- Containerized AI services using Docker for modular and reproducible deployment.
- Configured automated HTTPS using Caddy and secure remote access via Tailscale.
- Integrated models with Open WebUI to enable multi-user browser-based interaction.

J.A.R.V.I.S – Self-Learning AI Agent

- Developed voice-enabled AI agent using OpenAI APIs with long-term contextual memory via mem0.
- Implemented cross-platform workflow automation using n8n MCP server and Spotify API integration.
- Designed real-time voice interaction using LiveKit and serverless backend architecture.
- Optimized conversational latency to under 15-30 ms. (Currently in Progress)

Education

Raghu Engineering College, Visakhapatnam
B.Tech – Data Science Engineering Expected 2026
CGPA: 8.19
Relevant Coursework: Data Structures & Algorithms, OOPS, DBMS, Machine Learning

Sri Chaitanya Junior College, Visakhapatnam
Intermediate (MPC) 2020–2022
Percentage: 95%

Certifications

- Image Processing with MATLAB
- Python and SQL Skill Rack Certification