SATHISH KUMAR

+91 93441 82389 ◇ Chennai, India
mailbox230498@gmail.com ◇ linkedin.com/in/sathish-llm ◇ github.com/sathish-llm

OBJECTIVE

Machine Learning Engineer with 3+ years of experience in NLP, LLM fine-tuning, and end-to-end AI system design. Seeking impactful roles in LLM product development, inference optimization, and scalable AI infrastructure.

EDUCATION

Bachelor of Engineering (Computer Science) 2016 – 2020
Agni College of Technology, Chennai, India

SKILLS

Languages	Python, Java, HTML/CSS, FastAPI
NLP/LLMs	Transformers, BERT, GPT, LLaMA 3.1, RAG, NER, Topic Modeling, Summarization
Frameworks	PyTorch, TensorFlow, Hugging Face, Scikit-learn, SpaCy, Langflow
Infra	Docker, Tmux, Flask, SDK Packaging, Model Quantization
Databases	MongoDB, MySQL, Neo4j, Qdrant
Tools	OpenCV, InsightFace, Ultralytics, Pandas, NumPy, Seaborn, Matplotlib

EXPERIENCE

Data Scientist May 2024 – Present
IAT Solutions, Chennai, India

- Built and deployed deep learning models for structured and unstructured datasets.
- Used advanced EDA and statistical techniques to improve model explainability and accuracy.
- Developed dashboards and visualizations to communicate insights using Matplotlib and Seaborn.
- Conducted hypothesis testing and A/B experiments to validate ML-driven business decisions.
- Built text classification models and regression pipelines using Scikit-learn and XGBoost.

Machine Learning Engineer Feb 2021 – Apr 2024
Custologix Solutions India Pvt Ltd, Bengaluru, India

- Fine-tuned large language models (BERT, T5, LLaMA 3.1) for summarization, translation, and NER using LoRA, PEFT, and QLoRA.
- Built multilingual translation pipelines supporting 10+ languages with over 95% accuracy.
- Reduced model inference time by 40% via quantization and knowledge distillation techniques.
- Designed and deployed multiple SDKs for internal NLP tools (NER, classification, topic modeling).
- Integrated audio processing modules with transcription, speaker diarization, and real-time translation using Whisper and Wav2Vec2 models.
- Developed and optimized RAG-based chatbot with long-term memory and vector storage using Qdrant and LangChain.
- Engineered a scalable OCR pipeline capable of extracting handwritten and printed text with ¿95% accuracy, leveraging custom-trained CRNN models and PaddleOCR.
- Implemented a face recognition system with reverse lookup by storing embeddings in Qdrant; clustered results using K-Means for efficient video identity resolution.
- Collaborated with front-end and back-end engineers to deploy models using Docker, Flask, and FastAPI.

- Reduced GPU memory consumption by 50% via mixed-precision training and model pruning strategies.
- Documented reusable ML modules for the internal team, improving onboarding efficiency and model reusability.
- Conducted extensive model evaluations using custom benchmarks, confusion matrices, and F1/Recall metrics.

PROJECTS

LLM-Based SDK Toolkit. Packaged multiple NLP features including keyword search, sentiment analysis, template matching, and NER into a single Python SDK using Hugging Face and Scikit-learn. Enabled plug-and-play installation via wheel package.

Multilingual Audio NLP Pipeline. Designed an end-to-end audio pipeline for noise reduction, transcription, translation, and timestamping. Improved transcription accuracy by 25% and reduced noise interference by 30%.

Violence Detection using YOLO. Fine-tuned YOLOv8n for real-time violence pose detection, achieving 95%+ accuracy. Integrated alert system based on early movement patterns to preempt escalation.

Face Recognition and Reverse Lookup. Built a scalable face retrieval system using ArcFace embeddings stored in Qdrant. Clustered video frames with K-Means and reverse-matched identities via cosine similarity.

Custom Object Detection System. Fine-tuned YOLOv8n for detecting custom objects and vehicle number plates. Achieved 98% detection accuracy and stored data with timestamps using PaddleOCR and FastAPI.

LLM Fine-Tuning for Domain QA. Fine-tuned open-source LLaMA 3.1 models on domain-specific QA datasets using LoRA and PEFT techniques. Achieved 88%+ accuracy on custom evaluation benchmarks and integrated into a retrieval-augmented QA pipeline.

Memory-Augmented RAG Chatbot. Designed a production-grade chatbot with long-term memory and document grounding using LangChain, Qdrant, and OpenAI embeddings. Enabled persistent context handling across multiple user sessions with real-time PDF ingestion.

Few-Shot Intent Classification with Prompt Engineering. Built a zero/few-shot classifier using GPT-4-style prompts for multilingual customer service tickets. Achieved 91% precision on 10-class multiclass classification without fine-tuning.

EXTRA-CURRICULAR ACTIVITIES

- Technical mentor for students in ML and LLM-based research projects.
- Active contributor to open-source Hugging Face transformers fine-tuning repositories.

LANGUAGES

- Tamil – Native
- English – Fluent