Kuldeep Solanki

Gujarat, India | 7878790833 | kuldeepsolanki039@gmail.com | LinkedIn | GitHub

Summary

AI/ML Engineer with M.Sc. in Artificial Intelligence & Machine Learning, experienced in LLM fine-tuning, multilingual speech systems, and computer vision. Skilled in edge deployment and quantization with Hugging Face, Ollama, YOLO, and MMDetection. Currently building GenAI pipelines and real-time AI solutions for healthcare.

Education

Gujarat University — Gujarat, India
M.Sc. in Artificial Intelligence & Machine Learning — *2023 – 2025*

Gujarat University — Gujarat, India
B.Sc. in Zoology — *2021 – 2022*

Kamdhenu University — Gujarat, India
Diploma in Animal Husbandry — *2017 – 2020*

Experience

Associate AI/ML Engineer — Aug 2025 – Present
Artem HealthTech Pvt. Ltd. — *Ahmedabad, India*

- Contributing to development and deployment of GenAI pipelines for healthcare applications.
- Optimizing LLMs and speech models for production use in clinical and real-time environments.
- Developing and fine-tuning object detection models for medical imaging tasks using YOLO and MMDetection.
- Building lab report integration systems with OCR and NLP pipelines, enabling automated extraction of structured data.
- Collaborating with cross-functional teams to integrate AI modules into scalable medical software solutions.

AI/ML Intern — Feb 2025 – Jul 2025
Artem HealthTech Pvt. Ltd. — *Ahmedabad, India*

- Completed **22 end-to-end projects** in NLP, Computer Vision, Speech, and Edge ML, gaining full-stack AI development experience.
- Built production-ready NLP modules (translation, NLP-to-SQL, entity extraction) achieving $>$95% accuracy on EMR datasets.
- Designed and fine-tuned lightweight detection models on 30K+ samples, reaching $>$93% precision and pixel-level overlays for diagnostics.
- Researched and optimized mini-LLMs (1-bit quantization, ONNX export) and deployed pipelines on Raspberry Pi with 30% faster inference.
- Implemented low-latency multilingual STT + summarization with Whisper + LLaMA for healthcare transcription automation.
- Delivered production-ready prototypes with **Gradio UIs**, ensuring real-time testing, usability, and deployment readiness.

Technical Skills

Languages & Frameworks: Python, SQL, C++, TensorFlow, PyTorch, FastAPI, LangChain

Libraries & Tools: Hugging Face, Scikit-learn, Pandas, NumPy, Matplotlib, OpenCV, Docker, Git, ONNX

Models & Platforms: LLaMA, Mistral, XLM-R, Faster-Whisper, MMDetection, YOLO

Deployment Tools: Ollama, Hugging Face, Docker, Raspberry Pi

Projects

YOLO Deployment Pipeline 2025
Lightweight Detection Models for Edge Devices *Python, YOLO, ONNX, C++, Raspberry Pi*

- Trained YOLO-based detectors on 30K+ samples, achieving 93% precision in medical imaging.
- Converted PyTorch models to ONNX/C++ and deployed on Raspberry Pi with 30% faster inference.

Object Detection 2025
Medical Imaging with MMDetection *Python, MMDetection*

- Developed detection models with MMDetection for infected regions in chest X-rays.
- Achieved pixel-level overlays for diagnostics and integration into healthcare tools.

NLP-to-SQL Translator 2025
Schema-Aware Query Generator *Python, LLaMA, FastAPI, Gradio*

- Engineered NLP-to-SQL translator with schema awareness for dynamic database querying.
- Deployed with FastAPI backend and Gradio UI for interactive usability.

Speech-to-Text + Summary Generator 2025
Multilingual STT + GenAI Summary *Python, Whisper, LLaMA, Gradio*

- Developed real-time multilingual STT + summarization pipeline using Whisper + LLaMA.
- Optimized inference via pruning/vectorization and deployed with Gradio UI.

OCR Lab Automation 2025
Vision-based Text Extraction *Python, EasyOCR, Tesseract, Regex*

- Built OCR pipeline to extract structured data from lab test reports.
- Enhanced accuracy with regex post-processing and schema mapping for EMRs.

Text Translation Pipelines 2025
Multilingual NLP Models *Python, Hugging Face, MarianMT, Bloom*

- Developed multilingual translation pipelines with MarianMT, Bloom, and mT5 for healthcare datasets.

Other Projects 2023 – 2024
Brief Highlights

- Sentiment Analysis — Classified product reviews with 90% accuracy.
- Speech Emotion Recognition — Trained emotion detection model (85% accuracy).
- Student Attendance System — Automated real-time facial recognition with OpenCV.

Ongoing Projects 2025 – Ongoing
Research and Prototypes

- Persona Mimic AI — Prototyped mini-LLM for cross-device persona emulation.
- Error Finder AI — Built multilingual code error detector using CodeLLaMA + Phi.

Certifications

Data Analysis with Python — IBM (Coursera)

Transformers for NLP — Hugging Face (self-paced)