

ZAYED MAKANDAR

📍 Karnataka, IND 📞 +91 7795698903 ✉️ zayedmakandar03@gmail.com
🌐 zayedmakandar.com 🔗 [LinkedIn](#) 🐙 [GitHub](#)

SUMMARY

AI/ML Engineer with 1+ year of experience building GenAI, LLM and computer vision systems: RAG assistants, vision-language document understanding and real-time video analysis. Focused on making models fast on real hardware (vLLM, ONNX, OpenVINO, INT8) and shipping them as FastAPI services.

TECHNICAL SKILLS

Languages & Tools: Python, SQL, C++, Bash, PostgreSQL, Git, Linux, Docker, Jupyter
Generative AI & LLMs: LLMs, Retrieval-Augmented Generation (RAG), Prompt Engineering, Embeddings, Vector DBs (ChromaDB, FAISS), Hugging Face, VLMs, Fine-Tuning (LoRA/QLoRA)
Agentic AI: AI Agents, Tool/Function Calling, Multi-Agent Orchestration, LangChain, LangGraph, Model Context Protocol (MCP), LLM Evaluation (RAGAS)
Machine Learning & Deep Learning: ML Algorithms, Deep Learning (Neural Network Architectures, Transformers), PyTorch, scikit-learn, NumPy, Pandas, Feature Engineering, Model Evaluation
Computer Vision: OpenCV, MediaPipe, Object Detection (YOLO), Image & Video Processing, OCR
Model Optimisation: Quantisation, ONNX Runtime, OpenVINO, TensorRT, vLLM, llama.cpp, CUDA
MLOps & Cloud: FastAPI, REST APIs, Microservices, CI/CD (GitHub Actions), MLflow, Azure

WORK EXPERIENCE

Abhyudyaya Techno Solutions
AI/ML Engineer

Jun 2025 – Present
Karnataka, IND

- Built a CPU-only document-vision pipeline with self-hosted vision-language models (llama.cpp, OpenVINO INT4) for scanned vendor quotations; image-resolution capping cut visual tokens 67% and latency $3.47\times$.
- Added validation that recovers truncated LLM output and rejects extractions whose arithmetic doesn't reconcile, before data reaches purchase orders.
- Built a multi-tenant FastAPI inference platform serving self-hosted speech-to-text, LLM chat and TTS to AR/VR institutes, with GPU-aware queuing and GPU-to-CPU fallback.
- Designed and taught an AI/ML curriculum covering classical ML, deep learning, NLP and computer vision, with hands-on lab kits, for AI/ML trainers and students.

PROJECTS

DocIntel: Document Intelligence API | *FastAPI, vLLM, Transformers, Docker*

[GitHub](#)

- Built a GPU microservice converting PDFs, images and Office files into Markdown/JSON in 90+ languages with a 5B vision-language OCR model and a switchable Transformers/vLLM backend.
- Benchmarked and tuned serving (batching, prefix caching, shared-GPU VRAM budgeting), cutting a 10-page document from 95s to 13.3s ($7\times$ faster).

Edge-AI Health Monitor | *PyTorch, ONNX Runtime, OpenCV, MediaPipe*

[GitHub](#)

- Designed a multi-pathway PyTorch model (1D/2D-CNN, LSTM, self-attention, fusion layer) feeding 5 risk modules; quantised all 10 models to INT8 ONNX for fully offline CPU inference.
- Built a real-time webcam vision module (MediaPipe FaceMesh, InsightFace, OpenCV) estimating contactless heart rate via rPPG plus blink rate, stress and pallor, fused with IoT vital-sign sensors.

VoiceDesk: AI Front-Desk Assistant | *faster-whisper, ChromaDB, Ollama, FastAPI*

[GitHub](#)

- Deployed a self-hosted voice assistant kiosk: local speech-to-text, RAG over an editable ChromaDB knowledge base, a local LLM and streamed speech replies.
- Improved recognition in a noisy lobby with a wake-word fuzzy matcher and correction layer built from production logs, plus automatic offline TTS fallback.

EDUCATION

Government Engineering College, Gangavathi
B.Tech in Computer Science

2021 – 2025
Karnataka, IND