Shaibin K B

AI Engineer — Machine Learning — GenAI — Computer Vision
Email: shaibinkb16@gmail.com
GitHub: github.com/shaibinkb16

Mobile: +91-8075885690
Location: Kochi, Kerala, India
LinkedIn: linkedin.com/in/shaibinkb

Professional Summary

AI Engineer specializing in Machine Learning, Large Language Models (LLMs), Generative AI, and Computer Vision with 8+ months production experience. Expert in building end-to-end AI systems using Python, PyTorch, LangChain, LangGraph, RAG, and deep learning architectures including CNNs, ResNet, YOLO, CLIP, and BLIP. Proven track record architecting multi-agent AI platforms, voice AI solutions, and computer vision pipelines. Strong expertise in LLM/SLM fine-tuning, MCP, A2A orchestration, and real-time voice AI. Delivered 5+ production AI projects achieving 95%+ accuracy.

Technical Skills

LLMs & GenAI: GPT-4o, Claude, LLaMA, Mistral, LLM Fine-tuning, SLM Fine-tuning, Prompt Engineering, Few-Shot/Zero-Shot Learning
Computer Vision: CNN, ResNet, EfficientNet, YOLOv8, CLIP, BLIP, Object Detection, Image Embeddings, Semantic Image Search, OpenCV
AI Frameworks: PyTorch, TensorFlow, Hugging Face Transformers, LangChain, LangGraph, LlamaIndex, OpenAI API, torchvision
AI Agents & RAG: Multi-Agent Systems, A2A Communication, MCP, ReAct, AutoGen, CrewAI, RAG, Vector Embeddings, ChromaDB, FAISS, Pinecone
Deep Learning: Graph Neural Networks, Transfer Learning, Model Training, Model Deployment, Neural Networks
NLP: Natural Language Processing, Text Classification, Sentiment Analysis, Text Generation, Semantic Search
Voice AI: LiveKit, WebRTC, Whisper, OpenAI TTS, ElevenLabs, Real-time Audio Streaming
Data Engineering: Pandas, NumPy, Scikit-learn, Spark, Data Preprocessing, Feature Engineering
Backend & Cloud: FastAPI, REST API, AWS (S3, EC2, Lambda, Bedrock), Azure AI Foundry, Docker, Git, Linux, Nginx
Databases: PostgreSQL, MySQL, MongoDB, Redis, Neo4j, SQL, NoSQL, Graph Databases
Frontend: React, Node.js, JavaScript, TypeScript, HTML5, CSS3, Streamlit
Observability & Monitoring: LangWatch, LLM Tracing, AI Pipeline Monitoring, Performance Evaluation

Experience

Junior Software Engineer Feb 2026 – Present
Alignminds Technology *Kochi, Kerala, India*
- Developing backend services for a US client Hollywood payroll management system using FastAPI, handling complex payroll rules, union compliance, and multi-role crew payment workflows
- Designing and managing PostgreSQL database schemas for payroll processing, timecard tracking, and production accounting for large-scale film and TV productions
- Building RESTful APIs for payroll data ingestion, report generation, and integration with third-party production accounting tools

Junior AI Engineer May 2025 – Feb 2026
Art Technology and Software *Kochi, Kerala, India*
- Architected production Multi-Agent AI Voice Platform using LiveKit, WebRTC, MCP, and LangGraph, processing 1000+ daily voice conversations with 98% success rate and sub-500ms latency
- Built 4 AI agents (IntroAgent, HRAgent, ITSupportAgent, LMSAgent) with RAG-powered vector search using Amazon Bedrock, FAISS, and AWS S3 across 10,000+ enterprise documents
- Built CNN-based person retrieval system using CLIP, ResNet, and BLIP for cross-modal search with 91% retrieval accuracy across large image datasets
- Implemented voice AI pipelines using Whisper (STT), GPT-4o (NLU), and OpenAI TTS achieving 94% transcription accuracy for Indian English accents

Education

Master of Computer Applications (MCA), CGPA: 8.78/10 2023 – 2025
St. Joseph's College of Engineering and Technology, Palai *Kerala, India*
Bachelor of Science in Computer Science, CGPA: 7.48/10 2020 – 2023
Sahyajyothi Arts and Science College, Kumily *Kerala, India*

Projects

Multi-Agent AI Voice Platform – Real-time voice AI system using LiveKit, WebRTC, MCP handling 1000+ daily conversations. 4 specialized agents with seamless handoff. RAG using Amazon Bedrock, FAISS, LangGraph (10,000+ docs). Sub-500ms latency, 98% uptime. *Tech: Python, LiveKit, LangChain, LangGraph, FastMCP, GPT-4o, Whisper, FAISS, AWS, Next.js, PostgreSQL*
CNN-Based Person Retrieval System – Deep learning system to retrieve persons from large image datasets using natural language descriptions. CLIP for vision-language embeddings, ResNet-50 for feature extraction, YOLOv8 for detection, BLIP for captioning, FAISS for sub-second search across 50,000+ images with 91% top-5 accuracy. *Tech: Python, PyTorch, CLIP, BLIP, ResNet-50, YOLOv8, FAISS, FastAPI, OpenCV, React*

Certifications

Artificial Intelligence on Microsoft Azure – Microsoft (Coursera, 2024) – Azure AI, Azure ML, Cognitive Services
AWS Cloud Technical Essentials – AWS (Coursera, 2024) – S3, EC2, Lambda, RDS, IAM, VPC
Joy of Computing using Python – NPTEL (2023) – Python, Data Structures, Algorithms