

Anish Bhat

AI Engineer | Generative AI & LLM Engineering | RAG Systems | Agentic AI

+91 6362253632 | bhata744@gmail.com | linkedin.com/in/anishbhat00 | github.com/Cobot69 | anishbhat.in | Bengaluru, Karnataka, India

Professional Summary

AI/GenAI Engineer specializing in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and multi-agent AI systems, with hands-on experience designing, evaluating, and deploying production-grade AI applications across cloud and edge environments. Skilled in building hybrid RAG pipelines with vector and knowledge-graph retrieval, fine-tuning transformer-based models, orchestrating multi-agent workflows, and developing REST APIs with FastAPI. Experienced across the full ML lifecycle – data processing, model quantization, evaluation, containerized deployment, and CI/CD. Author of 3 peer-reviewed papers on model optimization and multimodal document intelligence.

Technical Skills

- **Generative AI & LLMs:** Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Hybrid Retrieval, Prompt Engineering, Embeddings, Vector Search, AI Agents / Agentic AI, Multi-Agent Systems, Function Calling, Model Evaluation, Conversational AI, Voice AI
- **Machine Learning & Deep Learning:** Machine Learning, Deep Learning, Neural Networks, Transformers, NLP, Computer Vision, Fine-Tuning, Model Quantization, Feature Engineering
- **Frameworks & Libraries:** PyTorch, TensorFlow, Hugging Face Transformers, LangChain, LlamaIndex, OpenCV, NumPy, Pandas
- **Backend & APIs:** FastAPI, Flask, REST APIs, Async Python, Next.js, JWT Authentication, Webhooks, Testing & Debugging
- **Databases & Vector Stores:** PostgreSQL, MongoDB, FAISS, Pinecone (Vector DB), Firebase, Supabase
- **Cloud, DevOps & MLOps:** Google Cloud Platform (GCP), AWS, Docker, Kubernetes, CI/CD, Git, GitHub, Vercel, TensorFlow Lite, Edge AI Deployment
- **Languages:** Python, JavaScript, SQL, HTML/CSS
- **Automation & AI Platforms:** n8n, Vapi, Twilio, ElevenLabs, WhatsApp Automation
- **Engineering Fundamentals:** Data Structures & Algorithms, System Design, OOP, DBMS, Linux, Debugging, Root Cause Analysis, Clean Code Principles

Experience

AI Engineer & Automation Developer

Dec 2024 – Jun 2026

Independent Freelance

Remote

- Built an enterprise-style Hybrid Retrieval-Augmented Generation platform using LlamaIndex, FAISS, and Groq-hosted LLMs on a custom enterprise dataset, combining dense vector search, sentence-window retrieval, and knowledge-graph retrieval to improve contextual grounding and reduce hallucinations.
- Implemented cross-encoder reranking and multi-source ingestion (PDF, CSV, and website data via BeautifulSoup) with Hugging Face embeddings; containerized the system with Docker and built a Streamlit interface delivering transparent, citation-backed responses, with full version control maintained on GitHub.
- Architected Svarra, a multi-agent AI voice automation platform with 4 specialized agents (FastAPI, Next.js) for inbound/outbound calling, lead qualification, and appointment booking, using event-driven webhook orchestration (n8n, Supabase, Twilio, Vapi, ElevenLabs); built a real-time analytics dashboard for call cost and conversion tracking.
- Built and deployed a full-stack e-commerce platform for a clothing brand using Medusa.js with a customized backend, product catalog, and order-management workflows; owned architecture, integration, and deployment end-to-end. Led Instagram-based marketing, growing monthly reach to 200K+ views in 3 months and contributing to an estimated 30% month-over-month revenue increase.
- Delivered AI-assisted digital strategy, ad campaigns, and WhatsApp automation for additional small businesses, improving customer engagement and order handling.

Embedded AI Intern

Feb 2025 – Apr 2025

Decibels Lab

Bengaluru, India

- Trained and evaluated YOLOv11 object detection models on KITTI datasets using statistical validation metrics, achieving a 65% reduction in model size through structured pruning and INT8 quantization.
- Engineered preprocessing and feature-extraction pipelines for autonomous-driving datasets, boosting real-time detection throughput by 20% across Raspberry Pi and Jetson Nano edge devices.
- Structured a scalable backend ML pipeline integrating real-time object detection, lane detection, and sensor fusion at

approximately 210ms end-to-end latency, resolving synchronization and inference bottlenecks.

- Containerized embedded AI services with Docker and exposed REST APIs for telemetry and monitoring, maintaining structured GitHub branching and deployment workflows.

Selected Projects

Agentic RAG Medical Data Extractor

2025

- Built a Retrieval-Augmented Generation pipeline to automate extraction from large unstructured medical PDFs, using chunking strategies, vector-database indexing, and LLM-driven retrieval for structured summarization – replacing a slow manual review process.
- **Tech Stack:** Gemini API, LangChain, Pinecone, Docker, Firebase, GCP APIs, n8n. **Impact:** Delivered a modular, scalable backend supporting automated ingestion, embedding generation, and semantic retrieval with rate-limiting and context orchestration.

YOLO-OptiMob: Edge AI Optimization Pipeline

Published, SPIN 2025

- Compressed an oversized YOLO11n model for mobile deployment by applying 40% L1 unstructured pruning and INT8 quantization, converting to TensorFlow Lite and deploying it within an Android application.
- **Impact:** Reduced model size from 11.4MB to 2.5MB (78% reduction) while preserving accuracy for Android and low-power edge deployment.

Embedded ADAS Prototype

Published, CIS 2025

- Built an embedded Advanced Driver-Assistance System combining object detection, lane detection, ultrasonic sensing, and motor control for obstacle-aware navigation on constrained, cloud-free hardware; trained YOLOv12n on a custom dataset of 3,000+ images across 30+ classes and designed the supporting circuit and hardware for the prototype.
- **Tech Stack:** Raspberry Pi, quantized YOLOv12n, ONNX, Flask, sensor fusion. **Impact:** Converted the model to ONNX for optimized on-device deployment; achieved 6 FPS real-time inference at 210ms latency; built a Flask telemetry dashboard for live monitoring.

LongDocAI: Quantized Multimodal PDF Summarization & QA

Published, ICCIS 2025

- Built a pipeline integrating Donut (OCR-free parsing), mBART, and FLAN-T5 for accurate summarization and QA over long unstructured PDFs, fine-tuning mBART on 200+ research papers and FLAN-T5 on 20K+ QA pairs, and designing a multi-transformer pipeline for coordinated end-to-end processing.
- **Tech Stack:** Transformers, Hugging Face, PyTorch, LLM.int8 quantization. **Impact:** 15% improvement in ROUGE-L, 28% in METEOR, 85.3% QA accuracy, 70% model-size reduction, 40% lower inference latency; deployed the final model on Hugging Face.

Education

B.E. Computer Science (Artificial Intelligence), KLE Technological University

2022 – 2026

Hubballi, Karnataka

Pre-University Education (PCMB), Expert PU College

2020 – 2022

Mangalore, Karnataka

Certifications & Honors

Certifications: NVIDIA – Getting Started with AI on Jetson Nano (2026) | Neo4j Certified Professional: Graph Data Science (2026) | Neo4j: Building Knowledge Graphs with LLMs (2026)

Honors: 1st Place, College Mechatronics Competition, 130+ teams (semi-autonomous delivery robot) | Hackathon Finalist & Team Lead, HackKarnataka | Winner, Smart India Hackathon (Internal Round)