RAHUL RACHHOYA

India | rahulrachhoya0@gmail.com | +91-8386067676 | Portfolio: github.com/rahulrachhoya

PROFESSIONAL SUMMARY

AI Engineer with 3+ years of experience building production-grade AI systems using Large Language Models, Generative AI, RAG, and Voice AI. Expertise in Python, fine-tuning, prompt engineering, and deploying scalable AI workflows on AWS with observability and cost optimization. Strong track record of reducing errors, improving accuracy, and delivering measurable business impact through data-driven monitoring.

TECHNICAL SKILLS

LLMs & AI: Claude (Anthropic SDK), GPT-4 (OpenAI API), LLaMA, Gemini, Mistral, Hugging Face, Fine-tuning (LoRA, PEFT), Model Evaluation, Prompt Caching, Function Calling

AI Frameworks: LangChain, LlamaIndex, LangGraph, CrewAI, AutoGen, LangSmith, Streamlit, Gradio

RAG & Vector DBs: FAISS, Pinecone, ChromaDB, Weaviate, pgvector, Qdrant, Hybrid Search, Reranking (Cohere), Query Expansion

Cloud & MLOps: AWS (Bedrock, Lambda, EC2, S3, SageMaker), Docker, CI/CD, MLflow, Weights & Biases, Prometheus, Grafana

Programming: Python, FastAPI, Flask, Django, REST APIs, Microservices, pytest, Unit/Integration Testing

AI Production: Cost Optimization, A/B Testing, Performance Benchmarking, Error Tracking, Latency Optimization, Hallucination Reduction

PROFESSIONAL EXPERIENCE

AI Engineer – Voice AI & Conversational Systems

Careers360, India | Jan 2026 – Present

• Architected voice-based AI career counseling agent using Claude/Bedrock, Sarvam AI (STT/TTS), PostgreSQL+pgvector; RAG pipeline reduces hallucinations and grounds responses in structured data

• Implemented LangSmith observability tracking latency, cost, quality across 10K+ sessions; optimized costs by 40% through prompt caching and A/B testing ($0.40→$0.24/session)

• Built multi-agent workflow with function calling (intent classifier→retriever→counselor→roadmap); improved accuracy by 28%; deployed on AWS with 99.7% uptime

AI Engineer

Crystaltech Services, India | Aug 2024 – Dec 2025

• Built RAG pipelines (FAISS, Pinecone) reducing hallucinations by 35%; fine-tuned Mistral 7B using LoRA achieving 92% accuracy (vs 78% base model) with 40% faster inference

• Implemented hybrid search (BM25+vector) with Cohere reranking improving retrieval 85%→91%; set up Weights & Biases for monitoring 50+ LLM iterations

System Engineer (AI Context)

STL Digital Limited, Pune, India | Jun 2022 – Jul 2024

• Implemented RAG-based systems improving accuracy by 40%; built AI-assisted validation reducing errors by 60%; deployed Streamlit dashboard to AWS with 500+ daily users; implemented pytest suite with 85%+ coverage

KEY PROJECTS

Multi-Agent Document Intelligence System (LangGraph, Function Calling) – 4-agent system (extraction, validation, summarization, orchestrator) processing 500+ documents with 94% accuracy; 75% time reduction (30min→7min)

LLM Observability & Cost Optimization (LangSmith, W&B, Prometheus) – End-to-end monitoring dashboard; A/B tested GPT-4 vs Claude achieving 60% cost savings; prompt caching reduced costs 25%; latency improved 41% (3.2s→1.9s)

Fine-Tuned Domain LLM (Mistral 7B, LoRA, PEFT) – Fine-tuned on 5K examples improving accuracy 78%→92%; 40% latency reduction; training cost $50 vs $2K full retraining; 1,000+ daily inferences

Voice AI Career Counselor (Claude/Bedrock, Sarvam AI, pgvector) – End-to-end voice agent with emotion detection, real-time streaming; 10K+ sessions, 88% satisfaction, 1.8s avg response time

EDUCATION

Master of Computer Applications (MCA) | University of Hyderabad, India | 2019–2022 | Focus: Artificial Intelligence, Data Science, Machine Learning

CERTIFICATIONS & PUBLICATIONS

AWS Certified ML – Specialty (In Progress) • Deep Learning Specialization (Coursera) • Technical articles: 'Building Production Voice AI with Claude' (5K+ views), 'RAG at Scale', 'Reducing LLM Hallucinations by 35%' • Open source contributions to LangChain, LlamaIndex