ANSHU PRAJAPATI

Delhi, India | +91-8882374849 | anshuuprajapati01@gmail.com
github.com/anshuu-prajapati | linkedin.com/in/anshuu-prajapati | anshu.tech | Hugging Face | Kaggle | LeetCode

SUMMARY

AI Engineer with 2 years of experience (1 in production) building agentic AI, RAG and multimodal systems. Architected multi-agent orchestration on AWS Bedrock and LangGraph scaling to 1,000+ concurrent conversations, shipped as Dockerized FastAPI microservices on Kubernetes with CI/CD, versioned model rollouts and controlled A/B evaluation. Depth in LLM routing and function calling, embedding-based retrieval over vector databases and knowledge graphs, and guardrailed human-in-the-loop agents.

TECHNICAL SKILLS

Generative & Agentic AI: LangChain, LangGraph, Multi-Agent Orchestration, Tool Use & Function Calling, Intent Routing, RAG, Prompt Engineering, Structured (schema-constrained) Output, AWS Bedrock, OpenAI / Azure OpenAI, Groq, Ollama, AI Safety, Guardrails, Human-in-the-Loop Approval

Embeddings & Retrieval: Sentence Transformers, Pinecone, Qdrant, FAISS, Knowledge Graphs, Chunking & Indexing, Semantic Search, Retrieval Evaluation (hit rate, MRR, groundedness), LLM Caching, Token & Cost Optimization

Backend, Cloud & MLOps: FastAPI, Flask, REST, gRPC, WebSockets, Microservices, Docker, Kubernetes, CI/CD, pytest, AWS, Azure, GCP, Controlled A/B Evaluation

Deep Learning & Vision: PyTorch, TensorFlow, Hugging Face Transformers, CNNs, YOLOv8, ByteTrack / BoxMOT, FaceNet, Whisper (speech), OpenCV

AI / ML Fundamentals: Supervised & Unsupervised Learning, Classification, Regression, Clustering, Neural Networks, PCA, Time-Series & Sensor Signals, Feature Engineering, Model Evaluation

Data & Programming: Python (Pandas, NumPy, scikit-learn), TypeScript, JavaScript, Java, C/C++, SQL; PostgreSQL, MySQL, MongoDB, Redis; ETL, OCR (Tesseract), Web Scraping

CS Fundamentals: Data Structures, Algorithms, Problem Solving; Linear Algebra, Probability, Statistics; Git, Agile/Scrum

EXPERIENCE

Maaze Underwriting Pvt Solutions Jan 2026 – Present

AI Engineer — Agentic AI, Applied Research & Data Systems *Delhi, India*

- Architected multi-channel agentic AI on AWS Bedrock and LangGraph — an autonomous Twilio voice-calling agent and a multi-tenant WhatsApp agent — **scaling to 1,000+ concurrent conversations and cutting manual handling time by 45%.**
- Designed an LLM routing orchestrator with intent parsing and function calling across rule engines, ML pipelines and LLM agents, **improving intent-classification accuracy by 28%**; token batching and LLM caching **cut inference cost by 30%.**
- Advanced RAG by coupling Pinecone with multi-tenant Knowledge Graphs for entity-level reasoning, **improving retrieval precision 20% and latency 30%.**
- Built an end-to-end ETL and unstructured data pipeline (documents, audio, video, telemetry) — parsing, OCR, chunking, embedding and indexing into PostgreSQL and vector stores — feeding retrieval and agent reasoning.
- Engineered a real-time AI meeting-assistant agent with streaming speech-to-text, joining Teams, Meet and Zoom calls for **150+ daily users.**
- Shipped all services as Dockerized FastAPI microservices on Kubernetes with CI/CD, versioned model rollouts and controlled A/B evaluation of agent, model and retrieval changes.

DMAC Jan 2025 – Jan 2026

AI Full Stack Developer – Generative AI & Web Applications *India*

- Shipped full-stack GenAI apps on React, Express and MongoDB: BizpilotAi (Gemini business insights, dashboards, PDF reports), QuickGPT (OpenAI/Gemini chat and image generation, Stripe billing), a web scraper and an MGNREGA data dashboard.

PROJECTS

GPS Fleet Voice Agent LangGraph, Bedrock, Twilio | /gps-voice-agent

- Autonomous outbound calling agent that extracts service needs from **Hindi / Hinglish speech** with an LLM — code-switched ASR and intent extraction over real telephony, not clean English benchmarks.

Vehicle Support Chatbot — Production RAG Qdrant, PostgreSQL, Llama 3.3 70B | /Vehicle_support_chatbot

- WhatsApp RAG assistant with Sentence-Transformer embeddings in Qdrant, live vehicle-API tool calls with retry logic, conversation state and Redis caching.
- Instrumented a retrieval evaluation harness (golden question set, hit rate, MRR, groundedness) to benchmark chunking strategies and embedding models against a fixed baseline.

DocuSentinel — Multimodal Document Screening Smart India Hackathon | /DocuSentinel

- Screens passports, visas and IDs: OCR with ICAO MRZ checksum decoding, tamper forensics (error-level analysis, noise, EXIF) and FaceNet identity matching, fused into one explainable 0–100 risk score with a SHA-256 hash-chained audit ledger and pytest suite; a full document set in ~2 s.

Agentic Outreach Automation Claude agents, Streamlit, Gmail API | /Email_Automation

- Four-stage agent pipeline — discover companies, find leads, draft emails, send — with human-in-the-loop approval, spend budgets, rate limits and step caching, so the agent cannot act or spend beyond what is approved.

EDUCATION, CERTIFICATIONS & ACHIEVEMENTS

Indira Gandhi National Open University 2023 – 2026

Bachelor of Science in Computer Science | CGPA: 8.2 / 10 *New Delhi, India*

- **Certifications** — AI/ML Certification (IBM); AI/ML Certification (Cisco)
- **Achievements** — Winner, DS2S International Hackathon; Smart India Hackathon (SIH) finalist team; GDG Program Achievement