

Ajeet Singh

AI/ML Engineer | Generative AI · LLMs · RAG · Agentic AI

Bengaluru, India ajeetsingh7632@gmail.com +91 76329 13947 linkedin.com/in/ajeet-singh-66810616a
github.com/Ajeetsingh

SUMMARY

AI/ML Engineer with 4 years of experience designing, building, and deploying production-grade Generative AI, LLM, and machine learning systems. Owns AI systems end to end, from architecture and system design to deployment and optimization: enterprise RAG platforms with hybrid search and re-ranking, multi-agent workflows with LangGraph and tool calling, multimodal document intelligence, and fine-tuned domain-specific LLMs. Strong in scalable backend engineering, including FastAPI microservices, concurrent request handling, caching, database design (PostgreSQL, MongoDB), and cloud deployment on AWS (Bedrock, SageMaker, EC2) with Docker. Led a team of 10 AI engineers, driving architecture decisions and mentoring, and works with product stakeholders to turn business problems into reliable, low-latency AI solutions in production.

TECHNICAL SKILLS

- **Generative AI & Agents:** LLMs, RAG, Agentic AI, Multi-Agent Systems, LangGraph, LangChain, Tool Calling, Conversational Memory, Prompt & Context Engineering, LLM Fine-Tuning, Multimodal LLMs, Guardrails
- **Retrieval & Search:** Hybrid Search (Dense + BM25), Embeddings, FAISS, Semantic Search, Cross-Encoder Re-ranking, Reciprocal Rank Fusion, Query Expansion, Semantic Chunking
- **ML, Deep Learning & Vision:** PyTorch, TensorFlow, Hugging Face Transformers, Sentence Transformers, scikit-learn, Computer Vision, NLP, Regression & Time-Series Forecasting
- **LLMs:** AWS Bedrock, OpenAI, Gemini, Groq, LLaMA, Qwen, Falcon, Ollama
- **Backend & APIs:** Python, FastAPI, REST APIs, Microservices, JWT Auth, RBAC, React, TypeScript
- **Cloud & Deployment:** AWS (Bedrock, SageMaker, EC2, S3), Docker, Docker Compose, Git, Linux, PyTest
- **Data & Databases:** PostgreSQL, MongoDB, SQL, Alembic, Pandas, NumPy

EXPERIENCE

AI/ML Engineer · IDA Automation, Bengaluru

03/2026 – Present

- Design and build production Generative AI document intelligence for U.S. real estate records, covering mortgage, title, ownership, commercial, and residential documents.
- Built intelligent document processing pipelines that extract, validate, and structure complex fields such as legal land descriptions from unstructured property records.
- Fine-tuned LLMs on real estate data to improve domain-specific extraction accuracy over general-purpose models.
- Designed a multimodal computer vision pipeline for assessor parcel maps, combining classical CV techniques with a vision-language model to detect parcel geometry and compass orientation.
- Built RAG pipelines with semantic chunking, embeddings, and vector search for context-aware retrieval across large property document sets.
- Architected FastAPI microservices on AWS (Bedrock, SageMaker, EC2, S3) for document ingestion, AI orchestration, and retrieval, and tuned prompts and context construction to reduce hallucinations in production.

AI/ML Engineer, Team Lead · HeyBobo, Bengaluru

07/2025 – 02/2026

- Led a team of 10 AI engineers building a multi-domain intelligent assistant for health and education, owning architecture decisions, sprint planning, and technical mentoring.
- Designed multi-agent workflows in LangGraph and LangChain with multi-step reasoning, tool calling, conversational memory, and agent orchestration.
- Fine-tuned open-source LLMs (LLaMA, Qwen, Falcon) on health and education datasets to improve response accuracy and task performance.
- Built and fine-tuned image classification and recognition models with TensorFlow and PyTorch.
- Deployed model training and inference workloads on AWS (SageMaker, Bedrock, S3), serving them through FastAPI services in Docker.

Software Engineer · GL Communications, Bengaluru

09/2022 – 02/2025

- Developed Python linear regression models to forecast daily revenue for 50+ products from 100K+ historical sales records spanning 2+ years, producing 7-day forecasts for sales planning.
- Improved forecast accuracy by ~15% (MAE/RMSE) over a seasonal naive baseline (same weekday, previous week) through feature engineering, cross-validation, and model tuning.
- Wrote and optimized complex SQL queries (joins, window functions, aggregations) over large sales datasets for analytics and model training.
- Built Python ETL pipelines with Pandas and NumPy for extraction, cleaning, and feature engineering, feeding ML models and analytics dashboards.
- Developed reusable backend services exposing processed data and model outputs to downstream applications.

PROJECTS**Knowra: Enterprise Knowledge Platform** · github.com/Ajeeetsingh/enterprise_knowledge_assistant 2026 – Present*Permission-aware RAG platform that turns organisational documents into a searchable knowledge base with cited answers.*

- Architected and built a full-stack enterprise RAG platform (FastAPI, React/TypeScript, PostgreSQL) covering ingestion, retrieval, chat, and admin tooling.
- Designed a hybrid retrieval pipeline combining dense (FAISS, Sentence Transformers) and sparse (BM25) search with Reciprocal Rank Fusion, multi-query expansion, metadata-aware rescoring, and cross-encoder re-ranking.
- Built an intent-based query router that classifies each request (document, general, product-help, unsafe), skipping retrieval when not needed and applying guardrails to unsafe queries.
- Enforced fail-closed, role-based document access control before retrieval, so restricted content never reaches the LLM context, alongside JWT auth, RBAC across 4 roles, and rate limiting.
- Built a multi-format ingestion pipeline with structure extraction, semantic chunking, checksum-based duplicate detection, and incremental index updates; answers cite the exact source page with passage highlighting.
- Designed a provider-agnostic LLM layer (Groq, OpenAI, Gemini, Ollama); containerized with Docker Compose (non-root images, health checks, Alembic migrations) and covered by PyTest and frontend test suites.

fitness.ai: AI Workout Planning Backend · github.com/Ajeeetsingh/fitness_ai

HeyBobo · 2025

AI-powered fitness backend that generates personalized standard and athlete-specific workout plans with LLMs.

- Owned the full architecture design and implementation of a production FastAPI backend for LLM-driven personalized workout plan generation, supporting standard and athlete-specific plans.
- Built agentic layers with LangGraph and LangChain, with conversation history and memory management for context-aware plan generation and follow-ups.
- Implemented response caching so repeated requests reuse earlier generations instead of spending tokens on new LLM calls, cutting inference cost and latency.
- Designed and maintained the application and user databases, with authentication and role-based access control (RBAC).
- Engineered the service for production load handling, with flexible LLM configuration (Ollama or custom APIs) and local deployment via Uvicorn.

EDUCATION**Bachelor of Engineering, Computer Science** · Visvesvaraya Technological University

2018 – 2022

LANGUAGES

English (Fluent), Hindi (Native)