SHUBHAM UBHE

AI Engineer | Data Scientist — Production Generative AI · Agentic AI · RAG

Sambhajinagar, India • +91 8459296471 • ubheshubham.37@gmail.com • linkedin.com/in/shubhamubhe • research.techntrek.in

SUMMARY

AI Engineer / Data Scientist with 4+ years building production generative AI services in Python that turn enterprise data into measurable business outcomes. Most recently lead engineer on PRISM at STAND 8, an AI recruiting platform taken from zero to production in 6 months. Hands-on ownership of large language model (LLM) workflows — prompt orchestration, retrieval-augmented generation (RAG), structured outputs, tool/function calling — plus agent-based multi-step workflows, evaluation frameworks, SQL data pipelines, and Docker/Kubernetes deployment on AWS.

CORE SKILLS

Generative AI & LLMs: LLM APIs (GPT-4o, Gemini, Claude, Llama-3, Mistral) • advanced prompting & prompt orchestration • retrieval-augmented generation (RAG) • structured outputs • tool & function calling • fine-tuning • hallucination mitigation

Agentic AI: multi-agent, multi-step workflows • agent orchestration & state management • LangChain • LlamaIndex • agent memory

Evaluation & Experimentation: offline test sets / golden datasets • automated regression testing • LLM-as-judge • A/B testing • model evaluation • structured human review

Data Engineering & SQL: Python • SQL • PostgreSQL (JSONB, pgvector) • data pipelines (clean, transform, aggregate) • Spark • AWS S3

Retrieval: hybrid search (BM25 + dense vectors) • Reciprocal Rank Fusion (RRF) • OpenAI/Gemini embeddings • Pinecone • Weaviate

Platform & Observability: FastAPI • Docker • Kubernetes • AWS • GCP • monitoring & observability dashboards • Voice AI (Telnyx, ElevenLabs) • CUDA/GPU optimization (2 yrs academic research)

PROFESSIONAL EXPERIENCE

AI Data Engineer (Innovation Team) — STAND 8 Technology Consulting | Remote (USA) | Nov 2025 – Nov 2026

Lead engineer on PRISM, an AI-powered recruiting platform; owned end-to-end delivery from solution design through deployment, monitoring, and iteration.

- Built agent-based AI workflows with multi-step orchestration and state management: a 4-phase pipeline (ingestion, enrichment, embedding, ranking) coordinating specialized agents for resumes, recruiter notes, and voice screenings — with dependency ordering and batch race-condition controls for safe, controlled execution.
- Engineered retrieval-based augmentation: hybrid BM25 + dense-vector search (OpenAI text-embedding-3-large) fused via Reciprocal Rank Fusion, significantly improving recall on niche recruiter queries.
- Built multi-LLM routing across GPT-4o and Gemini by workload type (extraction, reasoning, summarization); cut per-candidate cost via automated A/B prompt evaluation.
- Authored the structured-output extraction service parsing 60+ fields per candidate into PostgreSQL JSONB with a versioned schema and dynamic field exclusion for token/cost optimization.
- Implemented the evaluation framework — offline golden test sets, automated regression harness, LLM-as-judge scoring, and dashboards for structured human review; experimentation surfaced a 43-point model accuracy gap and invalid machine-generated ground-truth labels.
- Owned the prompt orchestration layer (49 YAML templates across 78 production endpoints); led a full system audit fixing prompt_id collisions and composer config, and authored stakeholder-ready explainers for team leads.
- Shipped voice-AI screening agents (Telnyx + ElevenLabs) with post-call transcript retrieval and structured extraction; was on-call for the AI pipeline with observability dashboards covering token usage, prompt latency, and per-stage failure rates.

AI & Machine Learning Engineer — LTIMindtree | Jan 2022 – Nov 2025

- Architected a multi-agent RAG system (retrieval, reasoning, and output-generation agents), cutting complex query resolution from hours to minutes.
- Led an enterprise LLM platform processing 20M+ tokens/day at 99.5% reliability; reduced LLM API costs 35% via token optimization.
- Built a hybrid-retrieval RAG system for enterprise log analysis (BM25 + dense vectors, 7-prompt router) — 89% relevance on technical queries; 500GB+/day and 30M+ embeddings (LangChain, LlamaIndex, Pinecone, Weaviate, FastAPI).
- Optimized AWS S3 data-pipeline workflows — partitioning strategy, lifecycle policies, multipart parallel I/O, Spark read/write tuning — cutting ingestion latency for Fortune 500 cloud migrations.
- Delivered Llama-3 / Mistral fine-tuning pipelines (65% hallucination reduction, 78% domain-accuracy gain) and Canvas Eureka (50+ enterprise DB migrations, 85% fewer errors via AI-powered validation).

EDUCATION, CERTIFICATIONS & AWARDS

B.Tech — Vishwakarma Institute of Information Technology (2018–2022) | GPA 8.99/10

Certifications: PCAP (Python Institute) • DeepLearning.AI — Building & Deploying LLM Apps • Hugging Face NLP Specialist • Google Cloud GenAI • NVIDIA Transformer-Based NLP • AWS ML Specialty (in progress)

Awards: Hackathon Winner — RAG log-analysis agent (INR 75K) • VishwaCTF Winner (INR 5K) • LTIMindtree Solvathon Finalist (INR 1,600)