VINAYAK K V

AI / ML ENGINEER

+91 9037252890 | vinayakkvjob@gmail.com | linkedin.com/in/vinayak-kv-ds | github.com/vinayak533 | vinayakkv.com

PROFESSIONAL SUMMARY

AI/ML Engineer specializing in agentic LLM systems, low-latency RAG architectures, and scalable machine learning infrastructure. Architected multi-agent FastAPI pipelines cutting audit processing latency by 80%, with proven execution across full-lifecycle model training, vector retrieval, and cloud deployments. Co-author of an open-access clinical ML/LLM research preprint and ranked in the top 2% globally on LeetCode (Weekly Contest 487).

TECHNICAL SKILLS

Languages: Python, TypeScript, SQL, C++, Java

AI & Agent Systems: Multi-Agent Systems, LangGraph, LangChain, RAG, Tool Calling, LLM Routing & Fallbacks, Guardrails

Machine Learning & Retrieval: PyTorch, Scikit-learn, XGBoost, pgvector, Hybrid Retrieval (FTS5), Feature Engineering, MLflow

Backend & Systems: FastAPI, REST APIs, WebSockets, Next.js, Microservices, E2B Sandboxes, Linux

Cloud & Databases: PostgreSQL, Supabase, SQLite, Docker, AWS (EC2, S3), CI/CD (GitHub Actions)

EXPERIENCE

AMnova Technologies **June 2026 – Present**
AI/ML Engineer *Kochi, Kerala*

- Reduced manual store-audit turnaround by 80% during merchant onboarding by architecting an autonomous multi-agent pipeline using Python, LangChain, and FastAPI with deterministic fallback routing.
- Engineered a low-latency real-time analytics backend using FastAPI, Supabase, and PostgreSQL, eliminating query bottlenecks and scaling throughput for concurrent merchant data streams.
- Accelerated release velocity and system reliability by containerizing AI microservices with Docker and implementing zero-downtime CI/CD pipelines to AWS EC2 via GitHub Actions.

Amdox Technologies **April 2026 – May 2026**
Data Science Intern *Bengaluru, Karnataka*

- Enhanced retail demand forecasting and churn detection accuracy by engineering end-to-end predictive pipelines with XGBoost and Scikit-learn, incorporating domain-specific feature engineering.
- Minimized inference latency for high-throughput streaming by packaging trained model artifacts into asynchronous, production-ready FastAPI REST endpoints.
- Streamlined model lifecycle tracking and stakeholder visibility by instrumenting systematic experiment tracking in MLflow and deploying interactive Streamlit performance dashboards.

Cognifyz Technologies **Aug 2025 – Sep 2025**
Data Science Intern *Remote*

- Achieved 95% classification accuracy training and evaluating credit approval models on 280,000+ records using Scikit-learn.
- Improved model reliability on highly skewed data through advanced feature engineering and data cleaning with Pandas and NumPy.
- Mitigated class imbalance by tuning hyperparameters and validating performance across F1-score, precision, recall, and ROC-AUC.

PROJECTS

Loom-AI: Multi-Agent AI Workspace | github.com/vinayak533/loom.ai
Python, FastAPI, LangGraph, Next.js, TypeScript, PostgreSQL, pgvector, E2B

- Cut median Time to First Token (TTFT) by 44% (8.9s → 5.0s, p95 stabilized to 5.1s) under dense 80k-token/4-tool workloads by parallelizing read-only tool calls, caching conversation graphs, and decoupling persistence from the streaming event loop.
- Architected a zero-downtime model orchestration engine across 8 LLMs and 3 providers, implementing an automated task router with two-tier fallback circuits to eliminate session drops caused by provider 5xx outages and rate limits.
- Engineered a secure, stateful multi-agent execution runtime using LangGraph, PostgreSQL graph checkpointing, and isolated E2B micro-VM sandboxes, enabling deterministic trace replays, human-in-the-loop breakpoints, and sub-second context retrieval via pgvector.

LazyHire: AI Career Workspace | github.com/vinayak533/lazyhire
Next.js 16, React 19, TypeScript, SQLite (FTS5), Drizzle ORM, Transformers.js, Playwright

- Engineered an automated aggregation and deduplication pipeline across 9 regional job platforms using source-specific adapters and canonical entity normalization, eliminating duplicate postings and dead application links through URL verification.
- Cut LLM operational overhead and hallucination rates by deploying an on-device hybrid search engine (SQLite FTS5 + local all-MiniLM-L6-v2 embeddings via Transformers.js) with strict relevance gating, fallback routing to DeepSeek/Groq, and automated guardrails rejecting unverified claims.
- Architected a high-availability ranking and caching layer in SQLite with a 45-minute TTL, independent scraper deadlines, and partial-result fallbacks, keeping search response latencies low during upstream source throttling or downtime.

EDUCATION

Mahatma Gandhi University **2022 – 2025**
Bachelor of Computer Applications (BCA) *Kerala, India*

Avodha Centre for IT Education **Sep 2025 – Mar 2026**
Advanced Data Science and AI (7-Month Program) *Accredited by NSDC & Skill India*

ACHIEVEMENTS & CERTIFICATIONS

- **LeetCode Weekly Contest 487** (Jan 2026): Ranked 108 of 6,770 — top 2% globally.
- **HackerRank Orchestrate Hackathon** (May 2026): Built a terminal-based autonomous support-triage agent integrating HackerRank, Claude, and Visa APIs.
- **Research Publication — Preprint** (Zenodo, May 2026): Co-authored PredictLab, a multi-disease clinical risk-prediction platform combining ML models with LLM-based explainability.
- AI Fluency — Framework & Foundations Certification, Anthropic (2026).