DHANUSH KUMAR

AI Engineer & Full Stack Developer

Coimbatore, Tamil Nadu | +91 93455 90559 | dhanushkumaramk@gmail.com | Portfolio | LinkedIn
GitHub | Codolio

Professional Summary

Results-driven AI Engineer and Full Stack Developer with hands-on experience building production-grade Large Language Model (LLM) applications, multi-agent systems, and distributed backend architectures. Expertise in end-to-end development, AWS deployment, Retrieval-Augmented Generation (RAG) pipelines, vector database systems, and LLM security. Proven ability to deliver scalable, secure solutions with measurable performance improvements. Seeking AI Engineer or Full Stack roles to drive impactful product development.

Skills

AI & Generative AI: LangChain, LangGraph, RAG, Multi-Agent Systems, MCP, Fine-tuning (LoRA/QLoRA), Prompt & Context Engineering, Guardrails AI, LLM Security, LLM Gateway, LangSmith, MLflow, RAGAS, Hugging Face, SLM, VLM

Frontend: React.js, Next.js, Tailwind CSS, JavaScript, TypeScript

Backend: Node.js, Express.js, FastAPI, Python, REST API Design, WebSockets, SSE

Databases & Vector Stores: PostgreSQL, MongoDB, Redis, Qdrant (vector database), SQL

DevOps & Cloud: Docker, AWS, GitHub Actions, Nginx, CI/CD Pipelines

Other: System Design (HLD & LLD), Microservices Architecture, DSA, Agile Development

Experience

Python AI/ML Intern — Srishti Innovative Computer Systems Pvt Ltd *Jan 2026 — May 2026*

- Applied Python fundamentals and AI/ML concepts across structured real-world project environments during an intensive corporate training program
- Built a supervised classification model using scikit-learn to predict student performance, achieving 87% accuracy after feature engineering and hyperparameter tuning on a 5,000-record dataset
- Developed a text preprocessing and sentiment analysis pipeline using NLTK and Logistic Regression on customer review data with end-to-end experiment tracking via MLflow

Web Developer Intern — CodeSoft *Dec 2024 — Jan 2025*

- Developed scalable backend services and REST API design implementations for production client projects
- Optimized frontend performance and delivered production-ready features on schedule in an agile environment

Projects

ResearchMind — Agentic RAG Research & Document Intelligence *2025 — 2026*

Live | GitHub | Architecture

- Architected a 5-agent LangGraph system with parallel Search Swarm and RAG Vault; dynamic routing across arXiv, PubMed, GitHub, and Tavily — only reports scoring $\geq$7/10 by Critic Agent are delivered
- Built end-to-end Retrieval-Augmented Generation pipeline supporting PDF, Word, YouTube, and URL ingestion using Gemini embeddings into scoped Qdrant vector database collections; sub-80ms retrieval with Groq $\rightarrow$ Gemini failover
- Developed custom MCP server exposing Tavily and Qdrant as standardized tools for unified tool-calling across all 5 agents

- Implemented SSE streaming (7 event types, first token <3.5s), PostgreSQL persistence, Redis caching, jailbreak mitigation, rate limiting, and input sanitization
- Instrumented 5-table PostgreSQL telemetry tracking token count, latency, and quality score per agent run

Stack: *Node.js, TypeScript, LangGraph, LangChain, Qdrant, LiteLLM, Groq, Gemini, React.js, PostgreSQL, Redis, Docker, AWS*

GuardLayer — Open Source Self-Hostable LLM Security Gateway *2026*

Live | GitHub | Architecture

- Designed microservices-based LLM Security Gateway (API Gateway, Input Guard, Output Guard, LLM Proxy, Audit) with each service independently deployable via Docker Compose
- Achieved detection of all known prompt injection and jailbreak patterns with <50ms overhead using Microsoft Presidio and custom classifiers; PII scrubbing applied on every request
- Built output moderation pipeline — toxicity via Hugging Face Detoxify, hallucination checks, format validation — applied before every LLM response reaches the user
- Designed YAML-based zero-config setup with plugin system for custom validators; real-time threat dashboard with per-API-key audit logs and instant key revocation

Stack: *Node.js, TypeScript, Python, FastAPI, Guardrails AI, Presidio, Hugging Face, Redis, PostgreSQL, React.js, Docker, AWS*

Distributed Queue Engine — Production Job Queue System *2024 — 2025*

Live | GitHub | Architecture

- Built a distributed job queue in Node.js and TypeScript with atomic Redis Lua scripts for race-free state transitions (enqueue, move-to-active, fail, stall recovery) across horizontally scaled workers
- Implemented priority queues (high/normal/low), sliding-window rate limiting, heartbeat stall detection, and exponential backoff retry using Redis Sorted Sets
- Built real-time React dashboard with Socket.IO streaming and p50/p95/p99 latency percentile tracking across a full REST API control plane

Stack: *Node.js, TypeScript, Redis, React.js, Socket.IO, Docker, AWS*

Achievements

- Solved **1,000+** problems on LeetCode & GeeksforGeeks — **View on Codolio**
- **Rank #1** College Level Coding Score — GeeksforGeeks
- Open Source Contributor — Appwrite, n8n, Hoppscotch — **2,013 contributions** in the last year across 76+ public repositories
- Published technical blogs on Medium — **medium.com/@dhanushkumaramk**
- Shipped **30+** full-stack projects across AI, web, and infrastructure domains

Certifications

- **AWS Cloud Practitioner** — GeeksforGeeks
- **Prompt Engineering** — Simplilearn
- **React.js** — GeeksforGeeks
- **JavaScript** — GeeksforGeeks

Education

M.Sc. Information Technology *2025 — Present*

Hindusthan College of Arts and Science, Coimbatore

BCA — Bachelor of Computer Applications *2022 — 2025*

Nallamuthu Gounder Mahalingam College | CGPA: 7.92