Praveen Kumar D

+91 8870652361 | darkparadise877@gmail.com | abipravi.in | LinkedIn | GitHub

Summary

AI Systems Engineer with a focus on secure LLM applications, backend systems, and data pipelines. Experienced in building RAG-based systems, identifying security vulnerabilities, and implementing access control and secure architectures in production environments. Combines AI system design with cybersecurity practices to build reliable and secure intelligent systems

Experience

Axia AI Technologies *Oct 2025 – Present*
Software Engineer *Pune, India*

- Built a GPU-accelerated fraud detection system on Pure Storage FlashBlade + Kubernetes, distributing inference workloads across 4 NVIDIA L40S GPUs (2 nodes, 1 GPU/pod); benchmarked FlashBlade throughput under sustained ML load, validating parallel read/write performance for real-time financial transaction pipelines.
- Built a RAG-based system for enterprise document search using hybrid retrieval (vector + keyword), improving document retrieval accuracy to >93% and reducing manual lookup time by 60%.
- Developed a Medical Records Management pipeline using OCR and LLMs to digitize doctor-uploaded patient scans — automatically extracting and structuring prescriptions, diagnoses, and patient summaries, reducing manual data entry by ~75%.
- Developed a full-stack HR platform with automated candidate screening and onboarding workflows, reducing recruiter screening time by 40%.

Ridemap *Apr 2024 – Sep 2025*
Software Engineer Intern *Puducherry, India*

- Built a real-time GeoFence alert system for live bus tracking — supporting dynamic radius configuration per bus and delivering location-triggered push notifications to 20K+ active users with sub-second alert latency.
- Identified and fixed a critical Firebase misconfiguration exposing 20K+ user records by enforcing strict access rules and RBAC.
- Built and maintained real-time backend services using Node.js, Firebase, and WebSockets supporting 20K+ active users with 100ms data synchronization latency.
- Performed security reviews on APIs and authentication flows, identifying vulnerabilities and improving overall system security.

Independent Software Engineer *Jan 2021 – Present*
Self-Employed *Remote*

- **EstuLife** — Built analytics dashboards and a self-serve ad management interface for a student platform, enabling partners to create and track campaigns. *Tech: React.js, Next.js, Analytics APIs*
- **PrintA4** (printa4.in) — Designed and built the full backend for a web-based print management platform handling print job processing, secure payments via Razorpay, and business logic with API security controls and tamper-resistant job queues. *Tech: Flask, MongoDB, Razorpay*
- **POS & Billing Software** — Built a full-stack POS and billing system with real-time inventory tracking, automated stock alerts, and GST-enabled invoice generation. *Tech: React.js, Django, PostgreSQL*
- **ParkWheels** — Built a smart parking system with number plate recognition and real-time slot allocation, handling 1,500+ concurrent requests. *Tech: Next.js, React, MongoDB*

Projects

Keel — Autonomous Developer Identity & Shadow Clone | *Git, Claude Code, Gemini, Slack API, GitHub API* GitHub

- Built a pipeline that extracts developer decision patterns from Git history, PR discussions, and AI coding sessions using structured parsing + embeddings.
- Designed a "developer identity profile" capturing coding preferences, architectural patterns, and past decisions.
- Injected this profile into LLM workflows (Claude, Gemini) to provide persistent context across sessions. Reduced repetitive prompt engineering and improved consistency of generated code across sessions.
- Implemented similarity-based contradiction detection to flag deviations from prior decisions.

RAG Pipeline for LLM Hallucination Reduction | *NVIDIA NeMo, NeMo Agent Toolkit, LangChain, ChromaDB* *Final Year Project*

- Built a multi-stage RAG pipeline using retrieval optimization and structured prompting, achieving >93% retrieval precision and <4.1% hallucination rate.
- Implemented adversarial prompt testing and output validation to mitigate prompt injection, model manipulation, and unsafe outputs.

DOJO 2.0 — Lucas TVS | *Node.js, MongoDB, Socket.io, SecureGen Fingerprint API*

- Built a real-time training platform with biometric authentication and role-based access control, supporting 300+ trainees annually
- Reduced onboarding effort by ~50% through automation of training workflows and user management

Education

Manakula Vinayagar Institute of Technology, Puducherry *Aug 2022 – Jul 2026*
Bachelor of Technology, Artificial Intelligence and Machine Learning *CGPA: 7.9 / 10*

Skills

Languages: Python, JavaScript, TypeScript
AI / Gen AI: NVIDIA NeMo, NeMo Agent Toolkit, LangChain, ChromaDB, NVIDIA RAPIDS, XGBoost, Graph Neural Networks (GNN), RAG pipeline, LangGraph Multi-Agents, Embedding Models, RAGAS Evaluation, Guardrails, Vector Databases, Pydantic Validation
Cybersecurity: Vulnerability Assessment, RBAC, API Security, Auth Flows, Secure SDLC
Backend & Frameworks: Node.js, Express.js, Flask, Django, WebSockets, React.js, Next.js, React Native
DevOps / Infra: Docker, Kubernetes

Achievements

- **Top 75** — TCS HackQuest Season 10 (2026), national-level competitive programming and security challenge.
- **2nd Place** — MVIT Cyber Hacking Challenge (2023): identified vulnerabilities and applied defensive countermeasures under time-bound conditions.
- **3rd Place** — NIT Karaikal Coding & Cybersecurity Contest. Mentored contributors at Hacktoberfest on secure coding practices.