

Davinder Singh

+91 95188 64166 — developxdave@gmail.com — github.com/Davinder1436 — davinder.live — NIT Jalandhar, Jalandhar, Punjab, India

Education

Dr. B R Ambedkar National Institute of Technology Jalandhar

B.Tech in Electronics and Communication

Aug 2023 – Present

Experience

Intern - AI and Software Engineer Intern at Annam.ai

June 2025 – Present

- Designed and built a multi-layer AI engine in **FastAPI** for agriculture, supporting 12 Indian languages with prompt classification, intelligent caching, and dynamic context generation.
- Engineered a **scalable inference pipeline** using fine-tuned open-source LLMs, GPU autoscaling, and queued request handling for high-throughput workloads.
- Integrated **MCP services** (live mandi crop prices, weather APIs) with RAG over user/application/admin data to deliver personalized responses.
- Deployed production-ready stack with Redis (sessions + cache), S3 (long chats), and automated MLOps pipelines for CI/CD, testing, and model versioning.

Intern - AI and Backend Developer at Level SuperMind

Jan 2025 – Feb 2025

- Built and deployed an AI chatbot for spiritual guidance, puja bookings, and personalized horoscopes; served 10k+ user queries.
- Developed scalable RESTful backends using Python (Flask), MongoDB, and Gunicorn; improved API latency by 40%.
- Integrated Vedic astrology datasets and fine-tuned NLP models to enhance chatbot response accuracy.

Projects

BarqScoot - Freelance

Feb 2025 – Mar 2025

- Engineered a backend for an IoT e-scooter platform managing 200+ scooters and 1,000+ daily rides with a dynamic, scalable architecture with 7 scalable golang microservices.
- Designed data persistence using PostgreSQL with Redis as a hot/cold store for real-time scooter state management and fast lookups.
- Implemented high-throughput IoT data ingestion via AWS IoT Gateway and Kafka, ensuring reliable processing of scooter telemetry.
- Deployed on AWS VPC for initial stage and later migrated to Kubernetes on EKS for autoscaling.

Innrev - Freelance

Mar 2025 – May 2025

- Engineered a GenAI pipeline for performance analysis of 96k+ US hotels; delivered insights via a subscription-based SaaS product with RBAC.
- Developed a hotel data dashboard with AI-powered search, smart filters, and optimized rendering—3x faster viewport performance.
- Fulfilled end-to-end ownership across full stack, DevOps, and GenAI workflows; implemented CI/CD, cloud infra, and model deployment.

Property Management System & Booking Engine - Freelance

Aug 2025

- Built a full-stack PMS and booking engine (React, TypeScript, Prisma, PostgreSQL) with industry-standard and course-based booking features.
- Implemented secure transactions and microservices for invoicing/mailing.
- Utilized Redis for caching and asset-level transaction locking to prevent incomplete or conflicting bookings.
- Deployed on AWS (EC2, RDS) using Docker and Jenkins CI/CD, ensuring reliability and scalability.

Personal Intuitive Projects

- Dave-DFS - A decentralized P2P file-sharing system in Go with content-addressable storage, secure encryption, efficient broadcasting, and reliable communication
- Dave-Discord - A discord clone using nextjs with text, audio and video communication with a proper hierarchical channel management

Skills

Languages	Golang, JavaScript, TypeScript, Python, C/C++, Bash
Tools	ReactJS, NextJS, NodeJS, ExpressJS, FastAPI, GraphQL, gRPC, Go-std, Fiber, WebSockets
Databases	PostgreSQL, MongoDB, Redis, MySQL, Qdrant, Cassandra
DevOps	Docker, Kubernetes, Kafka, AWS, Linux, Jenkins
GenAI	CUDA, VLLM, Model Pipelining, Vector Stores, Model Inferecing