Harsh Aathavale

+91 9372311512 · aathavaleharsh@gmail.com
Mumbai, India · LinkedIn · GitHub

aiharsh.com

Senior Software Engineer with 2+ years of experience building scalable distributed systems, microservices, and production ML platforms across AWS, Azure, and GCP. Full-stack expertise (React/Node.js/FastAPI) with cloud-native infrastructure (Kubernetes, Docker, Terraform) and event-driven systems (Kafka, RabbitMQ). Differentiated by deep AI/ML engineering — production model-serving pipelines, RAG systems, and GenAI applications at scale.

TECHNICAL SKILLS

Languages: Python, Java, JavaScript, TypeScript, SQL | **Backend:** FastAPI, Spring Boot, Node.js, gRPC, REST APIs, Nginx | **Frontend:** React, Next.js, HTML/CSS | **Cloud & DevOps:** AWS, Azure, GCP, AKS, Docker, Kubernetes, Terraform, GitHub Actions, CI/CD | **Data & Messaging:** PostgreSQL, MongoDB, Redis, Kafka, RabbitMQ | **AI/ML:** PyTorch, Hugging Face, LangGraph, RAG, CUDA, Numba | **Observability:** Grafana, Prometheus, ELK Stack | **Tools:** Git, Linux, Docker Compose, Jira, Agile/Scrum

EXPERIENCE

Senior AI Engineer · Fulcrum Digital *June 2025 – May 2026*

- Architected production ML serving infrastructure on AKS with Kubernetes, processing 50K+ inference requests daily with sub-200ms latency
- Built full-stack GenAI apps (React + FastAPI) with LangGraph agentic workflows and RAG pipelines, reducing manual document processing by 70%
- Designed event-driven microservices using Kafka and RabbitMQ for real-time data ingestion across distributed systems
- Implemented observability (Grafana + Prometheus) and optimized pipelines with Numba/CUDA, achieving 4x throughput improvement

Software Engineer · Pertsol *Nov 2024 – June 2025*

- Engineered real-time target proximity prediction system via REST APIs, reducing location lookup latency by 60%
- Built speaker diarization pipeline and optimized ingestion scripts for audio processing at scale

Software Engineer Intern · Dyma.ai *July – Nov 2024*

- Built production agentic RAG system for document retrieval, improving retrieval accuracy by 35% through systematic benchmarking

Research Intern · Ernst & Young (EY) *Jan – July 2024*

- Developed ML-based document retrieval and fraud detection system, reducing false positive rate by 40%

PROJECTS

Distributed ML Training & Inference Platform *Python · PyTorch · Kubernetes · Docker · LoRA · Quantization*

- End-to-end ML training pipeline with LoRA fine-tuning, 8-bit quantization, and Kubernetes orchestration for GPU scaling

Real-Time Issue Processing System *React · TypeScript · FastAPI · WebSockets · Docker Compose · Nginx*

- Event-driven full-stack system with real-time WebSocket streaming, agentic workflows, Docker Compose + Nginx, built for high availability

EDUCATION

BTech, Electronics & Telecommunications — Vishwakarma Institute of Information Technology (SPPU, Pune) *2020–2024 · CGPA: 8.69*