

Endale Yohannes

Addis Ababa, Ethiopia | endaleyohannes8@gmail.com | +251955259074

LinkedIn | Github | Leetcode

Summary

Software Engineer specializing in Generative AI (LLMs, Retrieval-Augmented Generation / RAG), machine learning, and scalable Python back-end services. Strong in building end-to-end AI products—from ingestion and vector search to deployment and monitoring. Fast learner, strong problem solver, and effective collaborator in challenging environments.

Core Skills

- **Backend:** Python, FastAPI, Django, REST APIs, async workflows, automation scripting
- **GenAI / ML:** RAG, LangChain, LangGraph, Hugging Face, PyTorch, TensorFlow, Scikit-learn
- **Data / ETL:** Pandas, NumPy, MongoDB, SQL, data quality checks, batch processing
- **Vector Databases:** Pinecone, FAISS, Qdrant (self-hosted)
- **Cloud / DevOps:** AWS (Lambda, ECS/Fargate, ECR, S3, API Gateway, CloudWatch), Docker, Docker Compose, Linux server hardening, backups, load balancers, DNS, Cloudflare
- **Frontend:** React, Next.js, TypeScript, Tailwind CSS
- **Tools:** Git, Jira, ClickUp

Experience

Python Developer / Backend Engineer (GenAI + DevOps) | Canvas Prints | 03/2024 - Present

- Built **Retrieval-Augmented Generation (RAG)** systems embedding product/knowledge data into vector stores (Pinecone) and streaming context-aware outputs via LLMs.
- Orchestrated **large-scale ETL workflows** (hundreds of millions of rows) using **Pandas + MongoDB + AWS Batch**, improving reliability and refresh performance
- Designed and maintained back-end services/APIs supporting AI features, automation, and integrations (FastAPI + AWS).
- **Linux server management & secure setup:** hardened servers (principle of least privilege, firewalling, SSH hardening), and operationalized production-ready environments.
- **Self-hosted Qdrant on secured servers** for vector search workloads; configured persistence, monitoring, and safe operations practices.
- **DevOps & deployment ownership:** deployed multiple apps using Docker/Docker Compose, set up reverse proxies/load balancers, and standardized deployment workflows.
- Implemented DNS and delivery optimizations using **Cloudflare** (DNS management, performance and reliability improvements).
- Set up **backup strategies** and operational safeguards (restore-ready processes) for critical services and data.
- Hosted a **video generation model on a GPU server**, handling deployment and runtime stability for production usage.

Full Stack Python Developer | Holistic Reading LLC | 08/2023 - 12/2024

- Engaged in full-stack development and automation projects.
- Implemented CI/CD pipelines to improve development workflow.

Top Rated Freelance Developer | Upwork | 08/2023 - Present

- Worked on various projects including back-end and full-stack development.
- Delivered high-quality solutions to clients, ensuring project requirements were met and exceeded.

Web Developer (Team Lead) | Escalate | 03/2023 - 08/2023

- Led a team building the Community Portal (Next.js / TypeScript); project won a 150,000 ETB innovation grant.

Head of Education Intern | Africa to Silicon Valley (A2SV) | 04/2023 - 11/2024

- Planned curricula, mentored 20+ students, and ran weekly algorithm contests.

PROJECTS (SELECTED)

- **Least Resistance Habit System (Personal Project) (2026):** Full-stack habit tracker (**Next.js**, **FastAPI**, **PostgreSQL**) with **auth**, habit **CRUD**, **daily check-ins**, Today/Dashboard views, **LangGraph AI agent** for natural-language logging, **public/private blog**, **CSV export**, and **Docker-based deployment**.
- **Enterprise RAG Knowledge Chatbot (2025):** LangChain + LangGraph orchestration, Pinecone vector DB, follow-up query rewriting, dynamic top-k, streaming responses.
- **End-to-End RAG Agent Template (2024):** reusable service template for ingest → embed → chat/search APIs; production deployment workflow.
- **Local-Language Threat Detection (LSTM) (2024):** RNN-based classifier for local-language text; real-time moderation support.
- **EMG Signal Classification for Bionic Hand (2023):** CNN-based EMG gesture classification; 92% accuracy across 4 gestures.

Education

BSc in Software Engineering(AI stream), Addis Ababa University | **GPA:** 3.7 / 4.0
09/2019 - 07/2024 | Addis Ababa, Ethiopia

Introduction to Data Science and Competitive Programming, Africa to Silicon Valley (A2SV)
12/2021 - 12/2022 | Addis Ababa, Ethiopia

Achievements

- **Awarded 150,000 ETB (Ethiopian Birr)** as part of the Community Portal team for delivering the most impactful project.
- **Top 0.01%** Award from the National exam of Ethiopia from around 350000 students, 2019.
- **800+** problems solved on Leetcode.

Interests

- Participating in Competitive Programming Contests.
- Taking part in Data Science Hackathons.
- Exploring nature through visits to rivers, forests, and mountains.