

ETSUB W. FELEKE

Dallas, TX | etsubwende@gmail.com | [linkedin.com/in/etsub-wendwosen-feleke](https://www.linkedin.com/in/etsub-wendwosen-feleke) | [etsubwfeleke.github.io](https://github.com/etsubwfeleke)

SUMMARY

AI Engineer with 3+ years designing, building, and shipping production LLM and machine learning systems, spanning multi-agent orchestration, retrieval-augmented generation, deep learning, and predictive analytics. Hands-on across the modern AI stack: LangGraph multi-agent workflows, custom MCP servers, vector search and embeddings, prompt engineering, tool calling, model evaluation, and fine-tuning, deployed on serverless cloud infrastructure with MLOps guardrails. Dual 4.0 GPA in Data Science (M.S.) and Economics (B.A.), with a record of turning research-grade models into reliable, measurable products.

TECHNICAL SKILLS

LLMs & Agentic AI: LLM Integration (GPT-4, GPT-4o, Claude), Agentic AI, Agentic Orchestration, Multi-Agent Systems (LangGraph, ReAct, Tree-of-Thought), Tool / Function Calling, Prompt Engineering, Context Management, Model Evaluation Frameworks, LLM Benchmarking & Observability

Retrieval & Vector Search: RAG, Embeddings, Semantic Search, ChromaDB, FAISS, Pinecone, Milvus, pgvector, HNSW / ScaNN Indexing, Knowledge Graphs

MCP & Systems Integration: MCP Server Development (FastMCP), REST APIs, Pydantic API Contracts, LLM Tool Orchestration, External Tool Calling, System Integration

Deep Learning: PyTorch, TensorFlow, CLIP (ViT), ResNet, Computer Vision, Transformers, Fine-Tuning, Transfer Learning, SHAP Explainability

Machine Learning: Scikit-learn, Predictive Modeling, Time-Series Forecasting (ARIMA), Random Forest, XGBoost, OLS / Ridge / Lasso Regression

NLP & Text Mining: Text Classification, Embeddings, Semantic Search, Sentiment Analysis, Summarization

MLOps & Deployment: Airflow, MLflow, Docker, Containerization, Google Cloud Run (GCP), Azure, CI/CD, ETL / ELT, Apache Spark (PySpark), Model Monitoring & Drift Detection, FastAPI, Serverless, End-to-End Deployment

Languages & Tools: Python, SQL, JavaScript, R, Bash, Git, React

PROFESSIONAL EXPERIENCE

Machine Learning Engineer | JSI (John Snow, Inc.)

Jul 2022 - Dec 2024

Global Health & Development Organization, Addis Ababa, Ethiopia

- Designed and deployed end-to-end machine learning pipelines for structured and unstructured healthcare data, automating ingestion, cleaning, feature extraction, and model training.
- Built and deployed predictive and forecasting models (XGBoost, Random Forest, ARIMA) for patient attendance, immunization drop-offs, and disease trends, improving intervention targeting accuracy by 25% and program risk identification by 17%.
- Delivered a custom ML system ("M-flow") integrating time-series forecasting with live health program data, enabling near real-time monitoring and proactive resource allocation across clinics.
- Embedded LLM-assisted analytics into healthcare reporting, automating summarization of large reports for faster interpretation by policy makers.
- Implemented model monitoring and guardrails (drift detection, alerting) alongside statistical validation (cross-validation, bootstrap resampling, ROC analysis) to keep deployed models reliable and robust under noisy, shifting conditions.
- Optimized data storage and ETL pipelines (SQL, PySpark, Azure) to scale across millions of patient records, reducing query times by 40% and enabling real-time dashboard updates.
- Built and deployed interactive Tableau and Power BI dashboards integrating predictive signals, giving program managers actionable insight for intervention planning.
- Served as the primary technical liaison between data engineering, program, and field teams, translating model outputs into operational decisions that reached millions of beneficiaries.
- Performed exploratory data analysis and text-based data mining to surface actionable insights that informed policy recommendations and resource allocation across health programs.
- Partnered with health professionals and cross-functional teams to validate solutions against requirements, cutting reporting errors by 22% and reaching 95% alignment with stakeholder needs.
- Ensured compliance with data privacy and security regulations through anonymization, encryption, and access control for sensitive patient data.

Data Analyst | Glocal Management Partners

Aug 2021 - Jan 2022

Consulting & Microfinance, Addis Ababa, Ethiopia

- Created statistical models and interactive Tableau and Power BI dashboards, transforming complex operational data into actionable business intelligence for senior stakeholders.
- Applied time-series analysis and predictive modeling via Python APIs to identify market trends and forecast impacts, directly supporting investment strategy decisions.
- Built and optimized ETL pipelines in Python, SQL, and Azure for structured and unstructured client data, improving data integrity and accelerating analytical processing across portfolios.
- Processed and analyzed large datasets with Python (Pandas, NumPy) and SQL, applying statistical validation to ensure reliable, decision-ready outputs.
- Optimized data storage and processing on Azure for faster computation of complex statistical models, enabling independent remote delivery.
- Collaborated with cross-functional teams to communicate findings, document solutions, and align deliverables with business objectives and stakeholder needs.

SELECTED PROJECTS

RedInk, AI Multi-Agent Grading Assistant

Apr 2026

React, FastAPI, LangGraph, GPT-4, Claude, Cloud Run

- Architected and shipped a full-stack, production AI grading platform (React frontend, FastAPI backend, Google Cloud Run) using LangGraph to orchestrate 4 specialized agents with human-in-the-loop review, cutting grading time 80% (15 to 3 minutes per student) across 15+ file types including PDFs, source code, and Jupyter notebooks.
- Integrated GPT-4 and Claude for multi-model grading, delivered bulk grading for 100+ students through an interactive TA validation UI, and achieved 10 to 20 second inference latency on auto-scaling serverless infrastructure.

Cognitive Canvas, AI Knowledge System via MCP

Dec 2025

FastMCP, ChromaDB, GPT-4o-mini, Pydantic

- Architected and deployed a production-ready custom MCP server integrating OpenAI GPT-4o-mini, a ChromaDB vector database, and a Python backend, enabling Claude and GitHub Copilot to store, semantically search, and query a live knowledge base through natural-language tool calls.
- Designed strict Pydantic API contracts for type-safe tool execution between the LLM and backend, demonstrating reliable end-to-end agentic AI system deployment.

Auto Bench, Multi-Tool Agentic LLM Benchmark Suite

Dec 2025

LLMs, MCP, Tool Integration, Evaluation

- Built a benchmark framework using MCP tools to evaluate LLM agent performance across multi-tool reasoning tasks including quiz generation, summarization, and content extraction.
- Integrated LLMs with external APIs and tools to strengthen reasoning chains, and designed structured evaluation protocols and scoring rubrics to systematically surface model strengths, weaknesses, and failure modes.

Deep Learning Content Moderation, Near-Duplicate Video Detection

Spring 2026

CLIP (ViT-B/32), ResNet-50, Computer Vision

- Built a near-duplicate video detection system using CLIP (ViT-B/32) and ResNet-50 as feature extractors to identify modified re-uploads of banned content across 884 videos and 14 transformation types, achieving 0.905 mAP and 0.901 F1.
- Implemented a tiered moderation framework with threshold tuning and a perceptual-hashing baseline; frame-efficiency analysis retained 98% performance at 5x processing speed, with CLIP reaching 100% detection on structural modifications where ResNet dropped to 25%.

AI Adoption and Firm Performance, S&P 500 Empirical Analysis

Spring 2026

Random Forest, OLS / Ridge / Lasso, SHAP, Panel Data

- Investigated whether AI adoption intensity (patent activity, technological breadth, labor signals) predicts revenue growth across S&P 500 firms using longitudinal panel data with Random Forest, OLS, Ridge, and Lasso regression and SHAP-based explainability.
- Found AI adoption did not improve revenue prediction beyond financial controls and that patent ownership is highly concentrated (Gini = 0.87), suggesting incumbents use AI patents as a market-defense mechanism rather than a growth engine.

iTracker, Deep Learning Gaze Prediction App

Aug 2025

Deep Learning, Computer Vision, Fine-Tuning, Web Deployment

- Fine-tuned a computer vision model for gaze prediction on image datasets and deployed it as a fully interactive web application, taking a deep learning model from training through to a live production environment.
- Improved accuracy through data preprocessing and transfer learning, demonstrating end-to-end ownership from experimentation to deployment.

EDUCATION

M.S. Data Science, University of North Texas, Denton, TX

Aug 2024 - May 2026

GPA 4.0 | Coursework: Deep Learning, NLP, Machine Learning, Artificial Intelligence, Software Development for AI, Big Data, Data Mining

B.A. Economics, Unity University College, Addis Ababa, Ethiopia

Oct 2019 - Jul 2023

GPA 4.0 | Coursework: Advanced Econometrics, Statistical Methods, Computational Economics, Game Theory, Time-Series Analysis, Data Analytics