KETO ETEFA GEMECHU

ketoetefagemechu@gmail.com | 213-709-2430 | LinkedIn | Portfolio

EDUCATION

Pomona College | *Bachelor of Arts in Computer Science* **May 2025**

Coursework: Machine Learning, Algorithms, Data Structures, Linear Algebra, Statistical Modeling

SKILLS

AI & Systems: PyTorch, TensorFlow, **Hugging Face Transformers**, OpenAI API, **Prompt Engineering**, **RAG**, Model Inference APIs, FastAPI/Flask, REST, **Microservices**, **JWT Auth, AWS, Docker, PostgreSQL, Git, CI/CD**

WORK EXPERIENCE

The Bot Company— R&D Specialist (Robotics & ML Systems)

San Francisco, CA | Nov 2025 – Present

- Designed and executed structured validation protocols for ML-driven robotic systems, generating high-quality labeled datasets used in supervised training, model evaluation, and regression testing cycles.
- Analyzed system behavior across extended test sessions to identify failure modes, edge cases, and performance degradation patterns, contributing actionable insights that informed model retraining and system parameter tuning.
- Partnered with ML and robotics engineers to reproduce inconsistencies and validate fixes, improving reliability and reducing recurring behavioral defects in production test environments.

Somulecu — Software Developer, (Blockchain & Product Development)

Los Angeles, CA | Aug 2024 – Present

- Designed and deployed a secure blockchain desktop client (Rust + C++), supporting **50+** early users and achieving **99.7%** uptime through structured error handling and resilience-focused architecture.
- Architected asynchronous, event-driven system components using **Tokio and Boost.Asio,** improving concurrency handling, reducing latency, and strengthening fault tolerance for scalable multi-node deployments.
- Containerized application builds using **Docker** and integrated **CI/CD pipelines** to automate testing and ensure reproducible cross-environment releases.

Sematha Technologies — Product Development Intern (Backend)

Remote | May 2024 – Apr 2025

- Designed and maintained **RESTful APIs** (Python, Flask, HTTP/JSON) with structured validation and centralized error handling, reducing client data issues by **35%.**
- Modeled and optimized **PostgreSQL** schemas to enforce data integrity and improve query **efficiency** across production workflows.
- Implemented authentication, input validation, and automated unit/integration **tests** within **CI/CD** pipelines to ensure secure, reliable deployments.

PRODUCTS & PROJECTS

Pai-Claremont, Claremont, CA (Flutter + Dart + Firebase + Firestore + Mockito) Apr - Jun 2024

- Led development of a production mobile data platform (Flutter + Firestore) integrating real-time cloud services, improving data transmission efficiency by **50%.**
- **Designed** structured Firestore schemas and validation logic, reducing data integrity errors by **67%** across pilot studies.
- Implemented secure **authentication** and role-based access controls using **Firebase Auth** to protect research submissions.
- Developed automated **unit and widget tests** to ensure application stability and reduced load times **by 2 seconds**, improving usability scores **by 30%.**

College Application Support App (GPT-4 + Multi-modal Generative AI) Jan 2024 -Present

- Designed and deployed scalable **LLM inference microservices** (REST/HTTP, JSON) processing **500+** long-form essays in production with asynchronous request handling and structured output validation.
- Implemented **retrieval-augmented generation (RAG) pipelines** using embeddings and semantic retrieval to ground responses in rubric-aligned criteria, reducing hallucination rates and improving output determinism.
- Developed an automated evaluation framework incorporating rubric-based scoring, failure-case logging, and regression testing to monitor model performance across prompt iterations.
- Containerized inference services using Docker and integrated CI/CD pipelines to ensure reproducible deployments and controlled release updates.

AutoPost Agent – AI Content Project (Python, LangChain, FastAPI) Jul- August 2025

- Designed and deployed an **agentic LLM microservice** (FastAPI, REST/HTTP) that processes **long-form videos** through multi-step prompt orchestration and structured output validation, reducing **content repurposing time** from ~**6 hours to under 5 minutes.**
- Containerized the service using **Docker** to enable reproducible deployments and production scalability, while optimizing latency, token usage, and output consistency through controlled **A/B experimentation** (2× engagement lift).