YEBOAH DAVID ZAHEMEN

AI & SOFTWARE ENGINEER

AH-29163124, Kumasi - Ghana | davidzahemenyeboah@gmail.com | **GitHub** | **LinkedIn**

Results-driven AI & Software Engineer with 2+ years of experience specializing in Large Language Models (LLMs), MLOps and Software development. Proven ability to optimize complex ML systems, implement cutting-edge NLP & Software solutions for edge devices, and lead both independent research and collaborative development.

AREA OF EXPERTISE

Large Language Models: Fine-tuning (PEFT, LoRA, QLoRA), transformer architectures, RAG and MCP
Deep Learning: PyTorch, Transformers, parameter-efficient training

Software Engineering: React & React Native, Flutter, NoSQL DB with Firebase, PostgreSQL with Supabase,
Tools & Frameworks: Hugging Face, BitsAndBytes, Docker, Azure Cloud, GCP (Google Cloud Platforms)

KEY ACHIEVEMENTS

- **AdaptIQ:** Launched full-stack AI-Learning Platform with intelligent features in 2 months, retaining and improving learning outcomes for 500+ users in first month of launch
- **Robin**: Delivered a web-first AI coding IDE unifying prototyping, learning, and building; cut time-to-first-preview by ~45% and reduced context switching ~60% in pilot (n≈100) via **Ask➜Apply** agent flows, inline diff previews, safe VFS terminal, and 0-install live previews.
- **FinSight:** Successfully built a fine-tuned LLM (~ 2B Parameters) for finance, scoring high NLP benchmarks on custom datasets.

EXPERIENCE

AI & Software Developer, AdaptIQ — **April 2025 - Present**

- Developed **AdaptIQ**, a personalized AI learning platform featuring Google Gemini-powered tutoring, intelligent schedule generation, and multi-modal learning.
- Engineered an intelligent scheduling system incorporating learning science (spaced repetition, energy optimization) and user preferences, leading to more effective study plans.
- Integrated Google Gemini API to create subject-specific AI tutors, delivering interactive chat, dynamic quiz generation, and curated resources, significantly enriching the learning experience.
- Improved application responsiveness and data management by implementing robust state management with React Context API and efficient data fetching/caching via React Query.
- **Visit the Web App here**

AI & Software Developer, Robin AI — **July 2025 - Present**

- Built and shipped **Robin**, an AI-powered web IDE (Flutter + Supabase) with Ask➜Apply agent flows, inline diff previews, safe VFS system, and 0-install live app previews via WebContainers.
- Accelerated idea➜preview by ~45% and reduced context switching ~60%; >70% of pilot users reported clearer code comprehension from chat-embedded diffs.
- Delivered Playground (seed➜artifact streaming) and Learn (notes/quizzes/topic chat), unifying build + learn without leaving the editor, ensuring a seamless UX.
- Production-minded design: RLS-backed file ops, checkpoints/rollbacks, COOP/COEP security for live previews, preview retries/observability, and a hybrid server-runner roadmap for heavier stacks.
- *Actively developing the platform, with more improvements & features to come*
- **Visit the Web App here**

AI/ML Engineer, FinSightAI April 2025 - Present

- Architected and deployed **FinsightAI**, a production-ready conversational finance AI (SmolLM2, ~2B parameters), leveraging QLoRA, PEFT, and UnSloth for memory-efficient training (70% VRAM reduction).
- Engineered an end-to-end synthetic data pipeline processing diverse financial documents, generating 45K+ high-quality training conversations (70M+ tokens) while ensuring regulatory compliance.
- Optimized inference performance by 60% through 4/8-bit quantization and PyTorch SDPA attention, enabling real-time advisory on modest hardware on HuggingFace Infra.
- Improved ROUGE and BLEU scores by 45% on complex financial queries via systematic fine-tuning on targeted datasets across 8+ financial expertise categories.
- **View the model on HuggingFace | Interactive Demo on Google Colab**

ML Engineer Intern | Boston Consulting Group (Remote, US) Nov 2024 - Dec 2024

- Led development of an enterprise-grade conversational AI system, using PEFT and quantization on LLMs, to handle 10,000+ daily interactions with 99.9% uptime.
- Engineered an efficient multi-GPU training pipeline for BlenderBot (3B parameters), reducing training time by 65% while maintaining model performance.
- Developed a robust data processing pipeline for financial documents (PDF extraction, text normalization), improving training data quality and leading to an overall 45% improvement in domain-specific response accuracy.
- Successfully integrated the LLM system into existing infrastructure via cross-functional collaboration, ensuring scalability and production readiness.

EDUCATION

Bachelors of Computer Science | KNUST Sep 2023 - Sep 2026

KNUST (Kwame Nkrumah University of Science & Technology)

- Major in Software Development, Cybersecurity & Artificial Intelligence.
- Research Paper on "Enhancing Financial Domain Performance of Small Language Models through QLoRA Fine-tuning" **(link to paper)** .

ADDITIONAL INFORMATION

- **Languages:** English, French, Russian.
- **Certifications:** View Certifications here .
- **Awards/Activities:** Top 5 in INNGEN Hackathon KNUST, Active involvement in Tech Activities including a recent Tech Bootcamp **@TechStripped,** Active Contributor to the **Research & Products Committee, KNUST Data Science Club**