Kevin Timothy Muller

AI Engineer | Backend & Data

ktimothymuller@gmail.com · London, United Kingdom · LinkedIn · GitHub · Portfolio

Summary

I build AI systems that reach production and stay there. At Zebra Technologies I own the GenAI platform behind a global sales organisation — RAG assistants, LangGraph agents and NL-to-SQL analytics on Google Cloud. I like the unglamorous half most: cutting a 12.7-hour pipeline to minutes, collapsing six LLM calls into one, proving a number right before anyone acts on it. Backend and data are mine end to end; interfaces I ship with AI coding tools.

Technical skills

Languages: Python, TypeScript, SQL, C++, R

AI & ML: LangGraph, LangChain, RAG, Vertex AI, Gemini, OpenAI, PyTorch, TensorFlow, FAISS, evals

Backend & data: FastAPI, asyncio, REST APIs, BigQuery, Firestore, Cloudflare D1, Drizzle ORM, pytest

Cloud & delivery: GCP (Cloud Run, Vertex AI, Cloud Build), Cloudflare Workers, Docker, CI/CD, Git

AI-assisted delivery: Claude Code, Cursor, spec-first and subagent-driven development, custom agent tooling

Experience

AI Application Developer → Enterprise Systems Analyst II 12/2025 – Present

Zebra Technologies · Contract (12/2025 – 03/2026) → Full-time (04/2026 – present), London

GenAI systems & agents

- Own the GenAI platform behind a global sales organisation: 9 production services on Google Cloud and 487 commits across 18 repositories.
- Migrated the customer Support Assistant off AWS onto GCP and rebuilt it around a LangGraph router over two Vertex AI Search datastores, with Gemini-first inference and OpenAI fallback.
- Built a LangGraph Salesforce Q&A agent over BigQuery with SQL self-correction on dry-run, AES-GCM-encrypted session history and a bilingual detect-and-translate wrapper.
- Shipped a Voice-of-Customer RAG assistant end to end: a plan-then-fan-out query planner over Vertex AI Search, Gemini answers citing SharePoint, and a web front end behind Ping SSO.
- Rebuilt an AI deck generator to fit copy into 355 measured slots across 42 brand-approved designs, grounded in the sales retrieval corpus with a web-search evidence leg.

Data & cost engineering

- Cut the GenAI ingestion pipeline from ~62,000 per-record BigQuery MERGEs to one batched MERGE per table, turning a 12.7-hour run into minutes and removing $3–5 of DML cost on every run.
- Collapsed a natural-language-to-SQL analytics endpoint from 4–6 sequential LLM calls into a single Gemini Pro call with parallel BigQuery execution, and put a classifier in front that cut input tokens by 60–80%.

Sales impact & integrations

- Built and launched an LLM sales-playbook service (FastAPI, Cloud Run Jobs) to a 33-rep London cohort, lifting weekly conversations per rep 66% and meaningful conversations 46%.
- Integrated Outreach.io over a server-to-server auth chain with 16 read-only scopes, validating call duration against recorded media across 2,073 calls.

AI Product Engineer 05/2025 – 12/2025

AI stealth startup

- Designed and built a full-stack video-to-text application end to end, letting universities and businesses digitise physical documents from a single recorded scan.
- Engineered the pipeline: on-device C++/OpenCV frame extraction and text detection, printed-vs-handwritten routing across two OCR engines, and vision-LLM quality correction.

Machine Learning Intern 09/2022 – 12/2022

FIS Global Business Solutions

- Led the analytical engine for ‘FIS Ecoglobe’ at the InnovateIN48 competition, placing top 4 nationally.
- Built a multilayer ESG classifier pairing sentiment analysis with feature mapping into Environmental, Social and Governance domains, and a CNN that scored each article’s impact.
- Sourced the training corpus through the X (Twitter) and LinkedIn APIs and BeautifulSoup scraping.

Education

MSc Data Science and Analytics 09/2023 – 09/2024

University of Leeds, United Kingdom

- Dissertation: “Deducing molecular structure of plastics from flow data using deep learning”, supervised by Prof. Daniel J Read.
- Showed an LSTM recovers the molecular weight distribution of unimodal recycled plastics from rheology data within the 15% industry tolerance, verified with noise-injection and restricted-frequency robustness tests; documented where bimodal samples fall short.
- Highlight projects: 30-class image classification with an MLP and CNN in PyTorch; drone-delivery optimisation using MCDA (VIKOR) and linear programming.

BTech Electronics and Computer Engineering 07/2019 – 05/2023

Vellore Institute of Technology, Chennai

- Dissertation: “Underwater Autonomous Tank Cleaning Rover”, published on arXiv:2304.08185.
- Built the SLAM navigation module on ROS Gazebo on a Raspberry Pi, driven from a browser UI that sent target coordinates from a pre-loaded RViz map.

Projects

- JukeBoxxed (jukeboxxed.pages.dev): a real-time multiplayer party game on Cloudflare Durable Objects with the WebSocket hibernation API, backed by D1 and R2.
- JobBot: a job-application agent orchestrated with LangGraph and Playwright that answers only from FAISS-backed retrieval, gating low-confidence answers to a human and writing JSONL traces for audit.
- Us, Daily: a production journaling app for two on Cloudflare Workers with D1, KV, R2 and Vectorize, Better Auth, Drizzle ORM, a nightly cron and Claude-written captions that fall back deterministically when no key is set.
- ModulatorX (github.com/KennyTheBoss05/ModulatorX): a native FL Studio controller plugin in C++/JUCE with a drawable, tempo-synced LFO/envelope, MIDI and audio-threshold triggering and lock-free UI-to-audio handoff.

Recognition

InnovateIN48 top 4 nationally · VIT Datahub Datathon 3rd place · AWS Certified Cloud Practitioner · arXiv:2304.08185