TIMOTHEE KABONGO NKWAR

AI / ML Engineer — LLMs, RAG, and Production ML Systems

Nicosia, Cyprus • Open to Remote • Relocation

timotheenkwar@gmail.com • linkedin.com/in/timothee-nkwar • github.com/TimotheeNkwar • timotheenkwar.me

PROFILE

Innovative and proactive Data Scientist with a builder mindset, delivering end-to-end ML products from ideation to production. Strong track record in model governance, scalable pipelines, financial analytics, and cross-functional leadership.

TECHNICAL SKILLS

- **Languages:** Python • SQL • Bash
- **ML & Deep Learning:** Scikit-learn • XGBoost • LightGBM • PyTorch • Hugging Face Transformers • Pandas • NumPy
- **LLM & RAG:** LangChain • OpenAI API • Ollama • Pinecone • Hybrid Retrieval (BM25 + Dense) • Reranking • QLoRA / LoRA Fine-tuning • vLLM • Prompt Engineering • Agentic AI
- **MLOps & Backend:** FastAPI • Flask • Docker • CI/CD (GitHub Actions, GitLab) • MLflow • pytest • REST APIs
- **Cloud & Data:** GCP (BigQuery, Cloud Run, Vertex AI) • Supabase/PostgreSQL • MongoDB • Railway • Heroku
- **Tools:** Git • Linux • Jupyter • Matplotlib • Seaborn

PROJECTS

Medical LLM Fine-Tuning: TinyLlama 1.1B on MedQA-USMLE **2025 – Present**

- Fine-tuned TinyLlama-1.1B-Chat on MedQA-USMLE (∼10K Q&A pairs) using QLoRA (4-bit NF4, rank-16, alpha-32) via `peft` and `SFTTrainer` (TRL); trained with 1 batch size, 8 gradient accumulation steps, paged 8-bit optimizer, and gradient checkpointing on a 4GB VRAM GPU.
- End-to-end pipeline: automated data cleaning + instruction formatting via `datasets`, MLflow tracking (loss, ROUGE-L, MC-accuracy over 3 epochs), and model consolidation via `merge_and_unload()` reducing deployment footprint by 60%.
- Deployed a FastAPI inference service with a rule-based safety layer (medical keyword filtering + context pattern matching); containerized with Docker on GCP Cloud Run; full pytest + mocked model testing in CI/CD.

DataCraft Conversational AI Chatbot **2025 – Present**

- Built a production-grade LLM chatbot (Google Gemini 2.5 + LangChain) with Supabase article retrieval and dynamic project context scraping via Firecrawl API for context-aware, knowledge-grounded conversations.
- Real-time SSE streaming over Flask with rate limiting (10 req/min), token-level backpressure, exponential backoff for 429s, and graceful degradation; conversation management with auto language detection (24 langs), session-based multi-turn MongoDB memory (20-message pruning), 30-min timeout, and enriched per-message logging.
- Context window optimized via content compression (articles to 3,000 chars, project context to 2,600 chars) with intelligent query routing (article detection + project question classification) to maximize relevance within token budgets.

SpaceX Launch Success Prediction **June 2025**

- Predicted launch success across 205 SpaceX missions (88% accuracy) using Random Forest and Logistic Regression; features (launch year, rocket type, site) fetched via SpaceX REST API v4, stored in SQLite, enriched via SQL joins, and one-hot encoding.
- Built an interactive Plotly/Dash dashboard with real-time filtering, surfacing yearly trends, per-site success rates (KSC LC 39A: 94.8%, CCSFS SLC 40: 87.3%, VAFB SLC 4E: 90.0%), mission breakdowns across 3 rocket types and 4 sites, and a Folium geographic map of all launch pads.

WORK EXPERIENCE

Software & IT Assistant | Software Development Office — Cyprus International University **October 2025 – January 2026**

- Worked on a diverse range of data projects, from chatbot RAG system to Attendance QR code system.
- Implemented geographic data collection to prevent students outside the university from scanning attendance, and visualized this data to monitor and verify location-based restrictions.
- Participated and testing for Attendance QR code system.

AI & Software Intern | Software Development Office — Cyprus International University **June 2025 – August 2025**

- Collaborated with a software team to develop an intelligent chatbot that handles large datasets.
- Performed data processing and analysis to improve chatbot accuracy and performance.

EDUCATION

Cyprus International University **Nicosia, Cyprus**

Bachelor of Data Science January 2023 – January 2027

CGPA: 2.8 / 4.0

CERTIFICATES

IBM Data Science Professional Certificate Credential Apr 2025

Hands-on projects using Python, SQL, data visualization, and applied machine learning.

RAG and Agentic AI — IBM Credential Dec 2025

Retrieval-Augmented Generation (RAG), agentic AI architectures, and LLM-based workflows for real-world applications.

Google Project Management Professional Certificate Credential Feb 2026

Project planning, risk management, Agile methodologies, and stakeholder communication.

Microsoft AI & Machine Learning Engineering Professional Certificate Credential Feb 2026

Model development, evaluation, and deployment using Microsoft AI/ML tools.

LANGUAGES

French: Native **English:** Professional working proficiency