Nicholas Biiy Bwalley

AI & LLM Engineer | Full-Stack Software Engineer | RAG & MCP Systems Specialist

Nairobi, Kenya | **+254 714 394 332** | **nickbiiybwalley@gmail.com**

LinkedIn : **https://www.linkedin.com/in/nick-bwalley-49220a269**

GitHub: **https://github.com/NickBwalley**

PROFESSIONAL SUMMARY

An Innovative **Full-Stack Software Engineer and AI Engineer** with **5 years** of experience in designing, developing, and deploying scalable AI-driven applications. An Expert in developing **Retrieval-Augmented Generation (RAG)** Chatbots, **LLM integration** with tools like (OpenAI, Anthropic Claude, LLaMA), MCP and **cloud-native architectures** on AWS. Proven success in **vector database search**, **structured output modeling (Pydantic)**, and building **end-to-end AI solutions** that reduce costs, accelerate delivery, and boost performance. Well-versed in creating **intelligent automations** using **n8n** and **Flowise**, leveraging webhooks and seamless system integrations to streamline complex workflows. Adept at leading cross-functional teams, architecting complex back-end systems, and translating business needs into production-ready AI/ML applications. I have built several automation pipelines and AI chatbots that increased operational efficiency by **up to 40%**, improved response accuracy by **35%**, and reduced manual processing time from hours to **minutes** through end-to-end AI and integration solutions.

CORE TECHNICAL COMPETENCIES

AI & ML Technologies:
LLMs (OpenAI GPT-5, Claude, LLaMA, Grok, Groq), RAG Pipelines, MCP, LangChain, LangGraph, LangSmith, CrewAI, HuggingFace, Google-Colab, Transformer architecture, MoE (Mixture of Expert), TensorFlow, Keras, PyTorch, Prompt Engineering, Fine-tuning using tools like (LoRA/QLoRA, PEFT), NLP, Deep Learning, Creating Agentic AI Workflows, AI Agents, Structured Output (Pydantic).

Databases & Search Technologies:
Supabase, PostgreSQL (Vector Search + FTS), MongoDB, MySQL, ChromaDB, FAISS, Pinecone, AstraDB, Redis Streams.

Back-End Development Technologies:
Next.js, Python (FastAPI, Flask), Node.js, Express.js, REST APIs, OAuth 2.0, JWT Authentication, Microservices, Apache Kafka, SSR, ORM, Postman, Bash, cURL

Front-End Development Technologies:
ReactJS, Redux, Next.js (SSR/SSG), TailwindCSS, DaisyUI, Streamlit, Gradio.

Cloud & DevOps Technologies:
Vercel, Render, Docker, Netlify, GitHub Actions, AWS (EC2, S3, Bedrock), MLOps, Serverless Architectures, Microservices architecture, Bash, Linux terminal.

Automation Tools & Workflows:
Flowise, n8n, MCP, RAG, Webhooks, Agile/Scrum, Vite, Claude Desktop, Warp, Cursor, Windsurf, GitHub Copilot, Notion, NotebookLM.

PROFESSIONAL EXPERIENCE

AI Engineer & Full Stack Software Engineer | Remote | Mar 2020 – Present

- Delivered 15+ AI-powered applications for e-commerce, fintech, and real-estate clients, automating workflows and cutting operational costs by up to 40% and successfully done and completed over 170 projects in the past. Check it out here.
- Built production-grade AI RAG chatbots with n8n, Flowise & Webhooks using OpenAI GPT-4o, and Pinecone Vector Store, Improving lead generation by 90%.
- Fine-tuned open-source LLMs (LLaMA 3.1) with LoRA/QLoRA, reducing inference latency by 30% without quality loss.
- Engineered PostgreSQL vector search & FTS pipelines for lightning-fast AI retrieval systems.
- Integrated AI microservices into MERN/Laravel stacks, reducing feature delivery cycles by 25%.
- Led Agile teams (2–5 developers), driving sprint execution, code quality, and junior mentorship.

Software Engineer Intern | Turbo G & K Networks Ltd, Nairobi | Jan 2024 – Jun 2024

- Revamped website into a React SPA (MERN Stack), cutting load time by 55% and increasing traffic by 40%.
- Migrated from monolith to microservices architecture, reducing MTTR from 30 mins to under 10 mins.
- Implemented CI/CD pipelines with Docker & GitHub Actions, achieving zero-downtime deployments.

Cybersecurity Analyst Intern | Senselearner Technologies Pvt., Remote | Sep 2023 – Dec 2023

- **Conducted end-to-end security assessments** including network scanning, vulnerability analysis, and penetration testing of web applications, leveraging tools such as Metasploitable and Hacksplaining to identify and mitigate **95% of critical threats**, enhance system resilience by **40%**, and deliver comprehensive remediation reports with actionable insights.

EDUCATION

Exchange Program – Challenge Driven Education (CDE)
KTH Royal Institute of Technology, Stockholm, Sweden | Aug 2024 – Jan 2025

BSc Business Information Technology (First Class Honours, GPA 3.8/4.0)
Strathmore University, Nairobi, Kenya | Sep 2021 – Jun 2024

Diploma in Business Information Technology (Merit)
Strathmore University, Nairobi, Kenya | May 2019 – Sep 2021

PROJECT HIGHLIGHTS

- Led AI-driven mobility optimization project in Viksjö, Stockholm Sweden & used Advanced Machine Learning Algorithms to drive insights and to inform municipality modelling transport efficiency and flow
- Built "Stora Ecom", a full-stack e-commerce platform demonstrating modern UX and scalable architecture.
- Published a research article called "AI-Powered Customer Churn Predictor" that uses Machine Learning in predicting the probability that a customer in the bank will default and leave the banking services or not based on his usage metrics. Check it out here

CERTIFICATIONS & ACHIEVEMENTS

- Published Research: AI-Powered Customer Churn Predictor, ReadyTensor (Mar 2025)
- Dean's List Award Academic Year (2021–2022 & 2022-2023), Strathmore University.
- 3rd Place Award, Strathmore Sports Day – Weightlifting (Aug 2023)

LANGUAGES

English (C2 – Native; Both Spoken & Written) | Swahili (C2 – Native; Both Spoken & Written) | French (C1 – Working Proficiency; Both Spoken & Written)

HOBBIES

Swimming, Table Tennis, Bowling, Gym Coach, 8 Ball Pool, Rubik's Cube & Card Magic.