

RAYNAULD RONO

AI ENGINEER | AGENTIC AI | LLM APPLICATIONS

raynauldkipkoech@gmail.com | +254 11415 3834 | www.raynauldronoh.space |
github.com/RaynauldR-afk

PROFESSIONAL SUMMARY

AI Engineer and software developer focused on practical LLM-powered applications, agentic workflows, retrieval-augmented generation (RAG), semantic search, and evaluation pipelines. Experience spans software development, AI engineering, and agentic AI roles. Portfolio work includes an Enterprise Knowledge & Research Agent, semantic caching, document intelligence, and LLM evaluation concepts using Python and local language models. Focused on reliable, maintainable AI systems with measurable quality, latency, and user value.

TECHNICAL SKILLS

Programming: Python, REST APIs, unit testing, debugging, Git/GitHub, Pytorch

LLM and Agents: LangChain, LangGraph, LlamaIndex, OpenAI, prompt engineering, tool-based workflows

Retrieval and Data: RAG, semantic search, embeddings, vector databases (ChromaDB, FAISS), document chunking, FastAPI

ML and Operations: PyTorch, Hugging Face, model evaluation, latency/coverage metrics, MLflow (include where proficient)

SELECTED AI PROJECTS

Enterprise Knowledge & Research Agent | Python, LangGraph, LangChain, Pinecone, Neo4j

- Portfolio project for a knowledge assistant that retrieves relevant information and synthesizes context-aware answers through a structured agent workflow.
- Integrated a local Ollama model and validated API connectivity; built a testable client layer and handled connection errors and empty prompts.

LLM Evaluation Pipeline | Python, unittest, FastAPI, Huggingface

- Implemented an evaluation runner for multiple test cases, answer generation, keyword-coverage scoring, and per-case reporting.
- Added unit tests for input validation, multi-case ordering, coverage calculations, and generation-failure handling.

Semantic Caching Agent | LangGraph, LangChain, Claude, FAISS

- Project direction: compare incoming queries semantically and reuse a cached response when similarity passes a defined threshold, reducing repeated model calls.
- Evaluate cache hit rate, answer correctness, latency saved, and false-match rate.

Document Intelligence / RAG Assistant | ChromaDB, embeddings, Python, OpenAI, Milvus

- Project direction: ingest and chunk documents, index embeddings, retrieve relevant passages, and generate answers grounded in retrieved context.
- Evaluate retrieval relevance, groundedness, response quality, and retrieval latency.

PROFESSIONAL EXPERIENCE

AI Engineer | Handshake AI • Aug 2025 – Aug 2026

- Designed and refined LLM-powered workflows, including prompt strategies, retrieval steps, and tool use, to support reliable AI-assisted tasks.

- Evaluated model outputs against quality criteria; investigated hallucinations, inconsistent responses, and failure cases, then documented improvements.
- Built and maintained Python components and API integrations for AI features, with attention to error handling, logging, testing, and maintainability.
- Collaborated with product and engineering stakeholders to translate requirements into scoped AI solutions and communicate trade-offs in quality, latency, and cost.

Agentic AI Engineer | Code Parrot • Apr 2024 – Jun 2025

- Developed agent workflows that decomposed user requests into steps and coordinated LLM calls, tools, and retrieval components.
- Integrated APIs and knowledge sources into agent workflows, including input validation, timeout handling, and graceful failure paths.
- Implemented retrieval-augmented generation patterns such as document chunking, embedding-based search, context assembly, and source-grounded responses where appropriate.
- Tested agent behavior across representative tasks, refined prompts and routing logic, and documented limitations and edge cases for maintainers.

Junior AI Engineer | Wiru • Mar 2023 – Mar 2024

- Prepared and validated datasets for AI experiments, checking data quality, formatting, missing values, and consistency before use.
- Supported model and prompt experiments by defining test examples, comparing outputs, recording observations, and summarizing results.
- Built Python scripts and reusable components for data processing, inference, and evaluation, with clear input/output handling.
- Assisted with integrating AI capabilities into applications and troubleshooting issues involving dependencies, API responses, and model behavior.

Junior Software Developer | Koko Networks • Nov 2021 – Jan 2023

- Implemented and maintained application features using Python and related software tools, following existing code conventions and project requirements.
- Developed or updated REST API endpoints and service integrations, including request validation and clear error responses where required.
- Investigated bugs, reproduced issues, wrote or updated unit tests, and verified fixes before changes were merged or released.
- Used Git-based workflows to manage changes, participate in code reviews, document technical behavior, and coordinate work with other developers.

CERTIFICATIONS & PROFESSIONAL DEVELOPMENT

- Microsoft Certified: Azure AI Engineer Associate (AI-102) — Azure AI solution design and deployment.
- AWS Certified Machine Learning Engineer – Associate — production ML workflows on AWS.
- Databricks Certified Machine Learning Associate — ML lifecycle and platform fundamentals.
- Hugging Face course certificates in NLP/Transformers and Agents — practical LLM and agent foundations.
- DeepLearning.AI short courses in LangChain, RAG, and evaluating/debugging LLM applications.

EDUCATION

Bsc. Mathematics and Computer Science | Jomo Kenyatta University of Science and Technology | 2017-2021

ADDITIONAL DETAILS

- Portfolio: www.raynauldronoh.space
- GitHub: <https://github.com/RaynauldR-afk>
- References available upon request (optional).