Ella Afonso

AI SOFTWARE ENGINEER | FULL-STACK, LLMs, RAG AND AGENTIC SYSTEMS

| 07927114754 | https://github.com/Ella-Afonso | ella.afonso@outlook.com | Leeds / willing to relocate

Professional Summary

AI Software Engineer with an MSc in Artificial Intelligence and consulting experience within the NHS. Develops full-stack applications, APIs, machine learning pipelines, RAG, and multi-agent systems using Python, FastAPI, Flask, Next.js, SQL/NoSQL, and cloud technologies. Integrates Qwen and Claude into applications, validates structured outputs, measures quality and cost, and employs testing, CI/CD, logging, health checks, and observability to enhance reliability. Combines technical autonomy with requirements management, Agile practices, and stakeholder communication.

Technical Expertise

Programming and full-stack development	Python, JavaScript, HTML, CSS, FastAPI, Flask, Next.js, Streamlit, Jinja2, Pydantic, SQLAlchemy, REST APIs, JSON, Git and GitHub
LLMs, agents and RAG	Qwen/DashScope, Claude API, OpenAI API, Ollama, LangChain, Astra DB, embeddings, vector search, retrieval, structured outputs, prompt templates, routing, adjudication, MCP and human-in-the-loop
Testing and quality	pytest, unittest, FastAPI TestClient, Playwright, Jest, unit/API/integration tests, schema validation, regression probes, code review, and technical documentation.
Reliability and operation	timeouts, retries, rate-limit handling, structured logging, health endpoints, monitoring, observability, configuration and secrets management, failure assessment, and responsible AI
Cloud, databases, and delivery	Azure Data Factory, ADLS Gen2, Databricks/Lakeflow, Spark, Azure Functions, Event Hubs, Azure SQL, Key Vault, AWS, Alibaba Cloud, Docker, GitHub Actions, CI/CD, SQLite, MongoDB, Oracle and Tablestore.
AI-assisted development	Claude Code, Cursor, Windsurf, and GitHub Copilot for decomposition, scaffolding, debugging, testing, and documentation, with diff reviews, testing, and validation against official documentation.

Key Engineering Achievements

PDF Query RAG: Semantic document search with LangChain and Astra DB

- Implemented PDF ingestion, reading and chunking, embedding generation, indexing in an Astra DB vector database, and similarity search based on the user's question embedding.
- Orchestrated the retrieval flow using LangChain to fetch relevant context prior to generating a response, separating the document preparation phase from the query phase and preparing the architecture for the evaluation of Recall@k, groundedness, and citations.

Quorum: Multi-agent system for the auditable screening of systematic reviews

- Designed and implemented three specialized screeners - routing, adjudication, and a "skeptic" variant - using Python, FastAPI, MCP, and Qwen via DashScope; each decision produces validated JSON, a confidence score, a rationale, an abstention status, and an audit trail.
- Published a proof-of-deployment on Alibaba Cloud Function Compute, OSS, and Tablestore, and measured precision, recall, F2, parse rate, cost per record, and human review load.
- On a holdout set of 328 records, the selected configuration achieved precision, recall, and F2 scores of 0.875 and reduced the review workload by 57%; on a second dataset of 440 records, it maintained fixed prompts and thresholds and documented instances where the baseline outperformed the partnership, converting the negative result into a conditional deployment rule.

Parkinson's Telemonitoring: API-served ML application with CI/CD and observability

- Built a reproducible pipeline using 5,875 records and 18 vocal variables to predict UPDRS scores, incorporating data preparation, StandardScaler, training, consistent model/scaler serialization, and evaluation via MAE, MSE, RMSE, R^2, and adjusted R^2.
- Enabled inference via Flask/Jinja2 and API endpoints, using Gunicorn and Docker; automated testing and delivery with GitHub Actions, secrets, and health checks.
- Added management of timeouts, retries, and rate limits, input validation, structured logging, monitoring, and observability; documented clinical limitations to prevent interpretation as a diagnostic tool.

Explore additional AI, software engineering and data projects on GitHub : https://github.com/Ella-Afonso

Experience

Innovation Partner | University of Leeds | 09/2025 - Present

- Delivered cross-disciplinary workshops combining AI and Neuroscience, presenting technical concepts in an accessible way to varied audiences.
- Acted as a key point of contact for students and staff, understanding user needs and guiding engagement with digital tools and innovation resources.

Data Consultant | NHS | 04/2025 – 07/2025

• Analysed and reconciled 20,000+ transactional records across legacy systems and Oracle Cloud, identifying critical data defects and supporting migration accuracy and data integrity.
• Translated ambiguous stakeholder requests into validation rules, exception lists, ad-hoc analyses, monthly controls, and reporting, communicating defects, risks, and priorities via Power BI and Excel.
• Worked in an Agile environment with Azure DevOps and data/business user teams, documenting findings and supporting decision-making within a client setting.

Founder and Mentor | Rooted In Code | 04/2024 – 10/2025

• Founded a tutoring initiative providing tutoring to students from minorities, to foster their confidence and skills in algorithms, basics of machine learning, programming for data science and mathematics and statistics.
Clients: Holyrood Academy Sixth Form's EAL Department
• Worked with Holyrood Academy Sixth Form's EAL Department to enhance educational programmes for students whose second language is English, developing tailored support strategies to ease their integration into UK education and empower them through STEM -focused mentorship.

Previous experience: Senior Scientist - ALS (2022); Scientist Intern - Coalesce Product Development (2021), involving interdisciplinary collaboration and a 30% improvement in a validation process; previous laboratory roles at Yeo Valley Farms and Oscar Mayer.

Education

MSC Artificial Intelligence | University of Leeds | 2024 - 2026 1st

• Developed practical experience in Data Science, Deep Learning, Robotics, Knowledge Representation, Ethics, NLP, Transformers, and intelligent systems; scores of 84% and more across various modules.

BSC Neuroscience | University of Lancashire| 2019 - 2023 2:1

• Understanding of how to design, analyse and interpret research studies, also acquired knowledge of how various diseases and trauma to the nervous system affect behaviour
• Dissertation on theta waves and memory using a systematic review and meta-analysis.

Additional training:

• MLOps in Production - DeepLearning.AI; Foundation Models and Generative AI - MIT OpenCourseWare; Machine Learning Bootcamp - Stanford Online; AI Engineering - IBM (ongoing).

References and project examples available upon request as these are currently on GitHub