

Hao Wu | Resume

Data Scientist/Software Engineer

Status: Data Scientist @ Argus Media
Field: Software Engineering, Data Science
Techs: R, Python, JavaScript, HTML, CSS

Didcot, UK

Summary

Data Scientist & Software Engineer with extensive experience in full-stack application development. Expert in operationalising machine learning models and building interactive, data-driven applications using Python and R. Specialized in Generative AI, including the design of RAG pipelines, multimodal embedding strategies, and high-performance vector search optimization. Adept at managing the full lifecycle of data products from research and prototyping to deployment and monitoring.

Professional Experience

Data Scientist - Argus Media	Jan 2022 - Dec 2025
* Co-developed "Ask Argus" a production RAG system: Led the critical migration of the vector database layer to OpenSearch to improve scalability and retrieval latency.	
* Successfully built and delivered a web-based solution for optimising fuel mixing worth £250,000	
* Building production ready Shiny dashboards for commodity traders using {golem}framework	
* Building custom Shiny inputs using JavaScript to enhance user experience	
* Manage user permission and data storage with MySQL and AWS S3	
* Contributing to the data preparation pipeline that feeds into the models forecasting commodity prices and distribution	
Senior Data Science Specialist - Ricardo Energy & Environment	Jan 2020 - Jan 2022
* Building R Shiny applications integrated with SQL database for data checking and ratification	
* Building bespoke R packages to simplify the process of data analysis	
* Using regression and tree based statistical models to predict air quality trends	
* Building Python CLI tool to scrape and compare data from XML files	
Data Science Specialist - Ricardo Energy & Environment	Apr 2017 - Jan 2020
* Mentoring staff on doing data analysis in R	
* Dynamic reporting, data visualisation and building tools for analysing air quality data	

Education

Ph.D. - University of Edinburgh, UK	2017
* Understanding the spatial-temporal variability of urban air pollution	
* Dispersion and statistical modelling of urban air pollution	
* Conducting air quality measurements using stationary and portable air quality monitors.	
BSc (1st Class) - University of Edinburgh, UK	2013
* Environmental and Sustainable Chemistry	
BSc - South China University of Technology, China	2013
* Applied Chemistry	

Personal Projects

Agent for writing Environment Impact Assessment report - Github	2025
* Multi-agent AI system for automated EIA report drafting: LangGraph-based workflow coordinating Planner, Interviewer, and Drafter agents to generate Chinese Environmental Impact Assessment reports section-by-section	

- * Interactive information gathering and validation: Interviewer agent collects project details through conversation, validates completeness against regulatory guidelines, and signals readiness to draft
- * Stateful workflow orchestration with persistence: FastAPI backend with PostgreSQL state management, streaming responses and checkpointing for multi-turn conversations and progress tracking
- * Production-ready architecture: Docker containerization, Langfuse observability, structured prompts with guideline validation, and RESTful API design for frontend integration

Multimodal AI Fashion Stylist Agent - [Github](#)

2025

- * Architecture: Designed and built an end-to-end RAG application that acts as a personal stylist. Implemented a Multimodal embedding strategy using Fashion-CLIP to map user clothing images and text queries into a shared vector space, enabling semantic search (e.g., "outfit for a summer wedding") against a personal inventory.
- * Agentic Workflow: Engineered an LLM Agent capable of reasoning and Tool Use. The agent autonomously detects intent to query external APIs (e.g., fetching real-time weather data) before synthesizing a final outfit recommendation, ensuring advice is both context-aware and practical.
- * Tech Stack: Utilized Pinecone for high-performance vector storage and retrieval, and implemented a modular code-base allowing for easy integration of different LLM backends.

CLI Scanner for stocks with upcomming earnings events - [Github](#)

2025

- * Scrape Market Chameleon to get historical earnings performance
- * Using ib_async to fetch live option data to assess the liquidity of the stock options