Joy Chen

+44 7990 093175 | joy.jovanna.s@gmail.com | github.com/Joynnncode | linkedin.com/in/joychennn | Portfolio | United Kingdom

PROFESSIONAL SUMMARY

AI Engineer who builds full-stack, AI-powered products end to end - RAG pipelines, LLM integration, and pluggable multi-provider architectures shipped as live, deployed apps. Treats LLM features as systems to be measured rather than demos: request-level tracing for latency and token cost, and automated eval sets wired into CI as release gates. Uses Claude Code daily as part of a regular AI-assisted engineering workflow. Background training and evaluating deep learning models on medical imaging data brings research-grade rigour around evaluation, edge cases, and reproducibility to AI engineering work. Comfortable across the stack: Next.js/TypeScript front ends, FastAPI/ASP.NET Core services, Postgres/Drizzle data layers, and PyTorch/TensorFlow model work.

CORE COMPETENCIES

RAG Pipelines & LLM Integration | Prompt Engineering | Multi-Provider LLM Architecture | AI-Assisted Engineering | Full-Stack Development | Deep Learning & Model Evaluation | CI/CD & Containerised Deployment | Monitoring & Observability | Regulated & Reproducible Research

FULL-STACK AI & DEEP LEARNING PROJECTS

RadReport-Agent: Tool-Calling Clinical Imaging Agent

Tool-calling agent that answers chest X-ray questions by choosing which of six tools to invoke (DenseNet-121 classifier, PSPNet segmentation, cardiothoracic ratio, BM25 report search, patient lookup, PubMed), with every clinical statement traced back to the tool call that produced it. A 43-case eval set, 12 of them adversarial, caught the agent inventing academic citations despite picking the right tools; a grounding validator now verifies every quote against real tool output. Safety enforced in code, not in the prompt. Public de-identified data, no clinical validation.

Python, PyTorch, Tool-calling agents, Streamlit, Docker, pytest

Medical Image Analysis Platform

Built a three-tier full-stack platform (React/Vite frontend, ASP.NET Core auth/orchestration API, Python FastAPI inference service, PostgreSQL via EF Core) for uploading CT scans and viewing AI segmentation results in an interactive 3D browser viewer. Integrated two pretrained MONAI Model Zoo bundles (3D UNet, SegResNet covering 104 structures) for automated segmentation, with JWT auth and Docker Compose containerization. Scoped as a research/portfolio project, not validated for clinical use.

React, ASP.NET Core, FastAPI, MONAI, PostgreSQL, Docker

PantryChef - AI Recipe & Nutrition Assistant

Built a full-stack app turning available ingredients into ranked recipes and cooking tutorial videos, plus in-browser barcode scanning (ZXing) against Open Food Facts for Nutri-Score/NOVA health grading, and a RAG-powered Q&A assistant grounded in a hand-curated knowledge base (embeddings + cosine similarity, cited sources). Instrumented the full RAG pipeline with Langfuse tracing (latency, token cost, per-call input/output) and built a 25-question retrieval-relevance and answer-correctness eval set gating CI on RAG-relevant PRs. Designed a pluggable LLM provider layer (Groq/Anthropic/OpenAI) to isolate a public demo's API costs from real usage, with Vitest coverage and a GitHub Actions CI/CD pipeline.

Next.js, TypeScript, Vercel AI SDK, RAG, Langfuse, Vitest, Tailwind CSS, GitHub Actions

Automated Intracranial Haemorrhage Segmentation

Developed and evaluated three segmentation models on 200 non-contrast head CT volumes with voxel-level expert annotations: a 3D U-Net baseline plus improved 3D and 2D variants, with preprocessing (brain-ROI extraction, HU windowing, resampling to median voxel spacing, elastic/affine augmentation) targeted at small-lesion performance. Raised mean Dice from 0.60 to 0.66 (3D) and 0.68 (2D) and cut Hausdorff Distance from 50.0 to 29.4. Configured the CUDA 11.0 / PyTorch 1.7 environment from scratch on a remote Linux workstation (SSH, single RTX 3080) within a 10 GB VRAM budget.

PyTorch, Linux, CUDA

RESEARCH EXPERIENCE

BGI Research (China National GeneBank)

Sep 2024 - Dec 2024

Research Intern

Shenzhen, China

- Coordinated daily across three cross-functional lab teams (wet-lab, chip fabrication, QC) over a four-month workflow, aligning on protocol changes and communicating hand-offs clearly under time pressure.
- Analysed sequencing outputs in Python and maintained structured lab records under QC compliance.

Shantou University

Mar 2022 - Apr 2023

Research Assistant

Shantou, China

- Worked closely with faculty and lab partners to design, fabricate and validate a point-of-care diagnostic chip from concept through testing, documenting methodology and results for clear hand-off to future researchers.

EDUCATION

MSc Medical Imaging Science, Merit University of Manchester, UK — Sep 2023 - Sep 2024

Modules: Non-Radioisotope Imaging, Imaging in Clinical Diagnosis, Mathematical Foundations of Imaging, Mathematical Computing for Medical Imaging.

BEng (Hons) Biomedical Engineering Shantou University, China — Sep 2019 - Jun 2023

GPA 3.4/5.0 | Average 84/100 | Ranked 18/58 (Top 31%). Modules: Medical Imaging Physics, Computational Simulation in Medical Device Innovation, Applied Bioinformatics.

CERTIFICATIONS & AWARDS

Good Clinical Practice (GCP) Introduction E6(R3)	NIHR Research Delivery Network	2026
Academic Excellence Scholarship	Shantou University	

SKILLS

AI / LLM Engineering: RAG pipelines, LLM integration (Vercel AI SDK), Embeddings, Prompt engineering, Pluggable LLM providers (Groq, Anthropic, OpenAI), Tool-calling agents, Agent loops, Structured output validation, Grounding & hallucination checks

Familiar with: LangChain, LlamaIndex, Vector databases (Pinecone, Weaviate, AWS OpenSearch), AWS Bedrock, MCP (Model Context Protocol)

AI-Assisted Engineering: Claude Code - used daily as part of regular development workflow

Cloud & MLOps: AWS (Lambda, API Gateway), CI/CD (GitHub Actions), Containerised deployments (Docker Compose), ML deployment pipelines, model monitoring, logging & observability, LLM observability (Langfuse tracing), CI-gated RAG evals, adversarial eval sets, grounding verification, code-level safety constraints, automated & integration testing for AI applications

Full-Stack Development: Next.js, TypeScript, Tailwind CSS, Drizzle ORM, PostgreSQL, Auth.js, ASP.NET Core, FastAPI, Vitest, Docker

Deep Learning & Data: PyTorch, TensorFlow, MONAI, scikit-learn, NumPy, Pandas, OpenCV

Programming: Python, C# / .NET, Swift, MATLAB, R, SQL, Git

Languages: English (professional), French (conversational), Mandarin (native), Japanese (conversational)