Caleb Kan

+44 07828210751 | hello@calebkan.com | linkedin.com/in/caleb-kan | github.com/caleb-kan | calebkan.com

Education

Imperial College London — London, United Kingdom
Degree: Master of Engineering – MEng, Computing (Artificial Intelligence and Machine Learning) — Sep 2023 – Jun 2027
Grade: First-Class Honours (Year 1, Year 2)

Skills

Languages: **Native**: English, Mandarin, Cantonese
Programming Languages: C, Haskell, HTML/CSS, Java, Kotlin, Prolog, Python, Scala, SQL, TypeScript
Frameworks: Django, Expo, FastAPI, Flask, LangChain, React Native, Streamlit
Data & Infra: Firebase, Kafka, Milvus, MongoDB, MySQL, Qdrant, Snowflake, Supabase
Developer Tools: CI/CD, CMake, Docker, GDB, Git, JetBrains Suite, Jupyter, Linux, Valgrind, Vercel, VS Code

Experience

Software Engineer Intern (Agent Execution ML Team) — Jul 2025 – Sep 2025
HubSpot — London, United Kingdom

- Enabled AI agents memory by architecting an end-to-end refinement platform using Kafka, Snowflake, Qdrant (semantic search), and Python APIs, capturing 11k+ daily executions and applying LLM preference learning for personalised outputs.
- Improved AI agent research company web page tool reliability by 97.5% (error rate 40% to 1%) via resilient domain parsing and URL normalisation; performed statistical timeout analysis on the generate image tool, achieving 0% error rate.
- Enhanced platform robustness with JSON Schema validation and strict type checking, eliminating 40% of prior execution errors by blocking hallucinated parameters, which significantly improved reliability across HubSpot's AI agent ecosystem.

Software Engineer Intern (Marginal AI Team) — Aug 2024 – Oct 2024
Midas Advisory — London, United Kingdom

- Automated non-operating expense data processing from the top 15 U.S. banks using open-source LLMs, web scraping, data-source APIs, and Ray parallelisation, cutting processing time from hours to minutes and boosting accuracy to 97%.
- Implemented a Milvus vector-database with BAAI/bge-m3 embeddings and RRFRanker hybrid search, pairing a 93% accurate small LLM for data retrieval with a large LLM to deliver structured insights at enterprise scale reliably.

Projects

Team Up London | TypeScript, React Native, Expo, Supabase, Google Maps API, Vercel — May 2025 – Jun 2025

- Built a full-stack mobile application using React Native and TypeScript with Supabase backend, enabling recent graduates and working professionals to discover social sports communities and impromptu games. Awarded 2nd place out of 57 teams.
- Established a complete CI/CD pipeline with Jest testing, Expo EAS Build for cross-platform app generation, and Vercel-hosted distribution platform, implementing agile methodology with iterative design-feedback cycles.

WACC (Compiler Project) | Scala — Jan 2025 – Mar 2025

- Developed a WACC compiler frontend, building a lexer and parser for syntax analysis, creating an abstract syntax tree representation, implementing a symbol table for semantic checks, and designing descriptive error reporting to aid debugging.
- Implemented the backend using TAC intermediate representation with ARM32/AArch64 support, architecture-specific dependencies, code optimisations (constant propagation, folding, control flow analysis), and a standard math library.

PintOS (Operating System Project) | C — Oct 2024 – Dec 2024

- Enhanced OS kernel functionality by implementing timer-based thread synchronisation, advanced priority scheduling with priority donation, and the BSD scheduler, ensuring efficient multitasking and thread management.
- Developed a virtual memory subsystem, including paging, frame management with second-chance eviction, and supplementary page tables, enabling support for user programs, memory-mapped files, and stack growth.

ARMv8 Emulator, Assembler, and Visualiser | C — May 2024 – Jun 2024

- Engineered a cycle-accurate ARMv8 emulator and two-pass assembler, implementing precise register, memory, and instruction management including robust error handling in compliance with ARMv8 specifications.
- Developed a SDL2-based GUI for real-time emulation visualisation, featuring drag-and-drop assembly parsing and dynamic rendering of CPU states, registers, memory maps, and ALU.

AI Research Agent | Python — Aug 2023 – Sep 2023

- Developed an AI research agent using LangChain, OpenAI's LLM, Serper API, BeautifulSoup4, and X API for automated search, web scraping, cited summaries, and content posting. Integrated MongoDB Atlas for geolocation tracking.
- Implemented a Streamlit web app enabling anyone to use the AI research agent, integrating geospatial visualisation to map worldwide usage using geolocation data stored in MongoDB Atlas. Analysed engagement patterns of 70+ global users.