Daniel Tsiang

London, UK

dan_tsiang@hotmail.co.uk | https://danieltsiang.github.io

Senior Machine Learning Engineer with 5+ years of experience architecting scalable AI systems, specialising in LLMOps, GenAI, and multi-agent architectures (MCP, RAG) within the fintech sector. Proven track record of deploying robust machine learning pipelines that automate complex workflows, mitigate prompt injection risks, and recover thousands of engineering and operational hours. Adept at bridging the gap between advanced AI research and high-impact production deployments.

Work Projects

AI Customer Assistant | Mobile App *2026*

- Architected an agentic AI customer assistant utilising MCP tools to autonomously execute complex banking transactions (e.g., automated payments, recurring transfers).
- Engineered strict input/output LLM guardrails to neutralise prompt injection and jailbreak vulnerabilities, ensuring enterprise-grade security and strict topic adherence.
- Decreased customer support tickets by 20% and drove feature discovery via deep-linking, directly improving user autonomy and product engagement.

Enterprise RAG Engine | Customer Operations *2026*

- Engineered a high-precision Retrieval-Augmented-Generation (RAG) pipeline utilising Tantivy (lexical) and FAISS (semantic search) to dynamically query volatile knowledge bases.
- Slashed specialist team escalations by 50% by implementing an advanced retrieval and reranking architecture that guaranteed highly accurate, context-aware LLM generations.
- Eliminated manual knowledge retrieval bottlenecks for customer service agents, drastically reducing time-to-resolution metrics.

Automated Call Insight Generation Pipeline *2025*

- Integrated Google Cloud Gemini LLM and AWS Transcribe into Starling Bank's core Java platform to automate summarisation and insight extraction from customer call transcripts.
- Recovered 8,000+ hours monthly by entirely eliminating manual call summarisation workflows.
- Reduced post-call write-up latency by 40% (from 7 to 4 minutes), directly driving a 60% acceleration in average customer service response times (from 37s to 14s).

Multimodal Data Extraction & LLM Optimisation *2025*

- Deployed multimodal LLM architectures to automate unstructured data extraction from visually complex documents (IDs, cheques), populating production databases in real-time.
- Decreased data entry error rates by 15% and saved 500+ hours weekly by replacing legacy manual data entry processes.
- Orchestrated Kubeflow pipelines on GCP to systematise prompt engineering and hyperparameter tuning, monitoring precision, recall and inference latency to deploy highly optimised models.

Personal Projects

Venue Recommendation Agent (Multi-Agent System) (GitHub) *Jan - Apr 2026*

- Engineered a context-aware, multi-agent recommendation engine utilising Google ADK and the Model Context Protocol (MCP) to ingest real-time geospatial and categorical data.
- Integrated conversational memory, enabling the AI to retain user context to drive highly personalised multi-turn recommendation flows.

Personal Finance MCP Application (GitHub) *Feb 2026*

- Awarded 2nd place at the Claude Code Hackathon (London) for developing a natural language banking interface.
- Built bidirectional agent communication streams via Skybridge and Claude Code, allowing real-time LLM-driving execution of banking actions directly through interactive React UI components (charts, forms, widgets).

Experience

Starling Bank
London

Senior Machine Learning Software Engineer
Aug 2024 – Present

- Established enterprise LLMOps and ML strategies, setting architectural standards for deploying highly available generative AI models across the bank's services.
- Pioneered the introduction of AI security guardrails, mitigating prompt injection risks and ensuring compliance with financial sector regulations.

Machine Learning Software Engineer
Jul 2023 – Jul 2024

- Architected and scaled ML microservices using Docker and Kubernetes, successfully integrating LLMs and traditional ML models into the core Java banking platform to automate fraud detection and customer operations.
- Designed comprehensive MLOps pipelines, utilising OpenTelemetry and Prometheus for real-time telemetry, ensuring high-performance inference in production.
- Engineered robust backend services in Python, implementing strict linting rules, type checking, and comprehensive testing, eliminating code regressions and drastically improving system resiliency.

Junior Software Engineer
Aug 2021 – Jun 2023

- Developed a high-throughput load simulator to stress-test ML services from the Java banking platform, successfully identifying API performance bottlenecks and model drift before production releases.
- Architected serverless, cross-cloud data pipelines utilising event-driven AWS Lambda functions and GCP Dataflow to streamline ingestion and processing.
- Hardened deployment pipelines by integrating automated security vulnerability scanning in the CI with real-time Slack and GitHub alerts.

Skills

AI & Machine Learning	LLMOps, RAG Architecture, Agentic AI, Model Context Protocol (MCP), Semantic/Lexical Search (FAISS, Tantivy), Prompt Engineering, Google ADK
Programming Languages	Python, Java, JavaScript, SQL, Bash
Cloud, MLOps & Data Engineering	GCP, AWS, Kubernetes, Docker, Kubeflow, Dataflow (Apache Beam), Terraform
Frameworks & Libraries	FastAPI, Pandas, uv, PyTest, React, Guice, Mockito
Observability & Databases	OpenTelemetry, Prometheus, Grafana, BigQuery, PostgreSQL

Education

Harvard University
Remote

CS50ai & CS50x
Jan 2021 – Dec 2021

- Computer Science with Artificial Intelligence using Python

Imperial College London
London

Master of Engineering (Meng)
Oct 2014 – Jun 2018

- 1st Class Honours