PHILIP CHEUNG

Principal Forward Deployed Engineer

Applied AI · LLM Systems in Production · Financial Services

MSc Computer Science, AI & Data Science (Merit)

London, United Kingdom

PROFESSIONAL SUMMARY

I am the engineer who gets AI from a working demo into a business that has to live with it. As Principal Forward Deployed Engineer at a Swiss corporate finance advisory firm I run a five-person team and personally own the architecture, integrations and production accountability across four products, reporting weekly to a finance-led board. I build with AI coding agents under a three-gate review and evaluation process, which is how a small team ships integrations with Open Banking, Stripe, Firebase and twelve LLM providers, and I carry the pager when it breaks. Before engineering leadership I spent years on the client side: technical product management at Groupon and enterprise account management for a marketing-technology platform, working with government and multinational accounts across Hong Kong, Singapore, Taiwan and Macau. MSc Computer Science, AI and Data Science (Merit).

CORE CAPABILITIES

Discovery to Production Ownership

- Discovery to production, not demos
- SLAs, audit trail and rollback paths
- On-call; escalation point for incidents
- Scoped, dated deliveries for the board

Integration Depth

- Open Banking, Stripe, Firebase and Firestore
- Cloudflare Pages, Functions, Workers, D1, R2
- AWS and FIX connectivity
- 12 LLM providers; Groq and Cerebras LPU inference

Governance Built In

- Policy gates and evidence packs
- Human-review escalation for high-risk cases
- Per-call cost ledger
- Audit ledger on every AI call

Adoption & Stakeholder Work

- Thin working prototype plus live dashboard, not slide decks
- Weekly board reporting
- Board, investor and regulatory requirements into scoped deliveries
- Client-side background: account management and technical product management

Deployment Leadership

- Five-person team today
- 8 engineers mentored at Pacific Cloud
- Seniors ship v1, juniors iterate under review
- Architecture reviews and AI/ML working practices

How I Deliver (AI-assisted)

- Claude Code, Codex CLI, Gemini CLI, Cursor
- Eval harness and three review gates
- Evaluation criteria defined before build
- I review, debug and own what the agents produce

PROFESSIONAL EXPERIENCE

Principal Forward Deployed Engineer

Weidmann & Cie. AG

London (Hybrid) December 2024 - Present

Swiss corporate finance advisory firm building an applied-AI product portfolio: Qorinix (ultra-low-latency AI inference cloud), LaSpend (read-only AI money assistant over UK Open Banking), Fixxmi (Swiss consumer service marketplace) and an AI-native trading desk. I take each product from discovery to production with a five-person team and report weekly to a finance-led board.

- Own **end-to-end architecture and production accountability** across all four products on AWS, Cloudflare (Pages, Functions, Workers, D1, R2) and Firebase; I am the escalation point for production incidents
- **Designed and shipped the Qorinix inference control plane** with AI coding agents under a three-gate review and evaluation process: multi-tier model router across 12 providers (OpenAI, Anthropic, Gemini, DeepSeek, Qwen, Groq, Cerebras and others), streaming SSE, per-call cost ledger, usage-based billing and entitlements; targets **sub-200ms time-to-first-token**
- **Integrated LaSpend** with regulated UK Open Banking data, Stripe and Cloudflare D1: subscription detection, waste scoring and assisted cancellation, designed read-only by policy with human approval for every user-triggered action and deterministic fallback when models fail
- **Delivered Fixxmi's AI lead matching** (Next.js 15, Firebase Cloud Functions, Firestore europe-west6) across 12+ service categories in DE-CH/EN within GDPR and nFADP boundaries, with pay-per-lead Stripe credit packs
- **Built the governance loop** applied to every AI call: policy check, evidence-pack retrieval, prompt registry, model routing, audit ledger and human-review escalation for high-risk cases
- Ran a **12-provider inference benchmark** (LLM Arena, live on my portfolio site) to ground model, latency and cost decisions; introduced token tiering and self-hosted models for non-latency-critical workloads to cut API spend
- Translate board, investor and regulatory requirements into **scoped, dated deliveries**; win decisions with thin working prototypes and live dashboards rather than slide decks

AI Engineer & Technical Architect

Pacific Cloud Computing Ltd.

Hong Kong & Remote UK December 2021 - December 2024

Promoted internally from Senior Software Engineer to lead the firm's AI transformation across SaaS, analytics and market-intelligence products.

- **Built and operated a real-time ML inference service** (feature pipelines, model serving, monitoring) for the firm's analytics products, with latency and accuracy gates defined before each release
- Implemented **enterprise retrieval (RAG)** and market-intelligence NLP pipelines, including the evaluation methodology used to accept them into production
- Directed the launch of **three multi-tenant SaaS platforms**; ran architecture reviews and established AI/ML working practices
- **Mentored an agile team of 8 engineers**: seniors ship v1, juniors iterate under review
- Led **quantitative-markets decision support** research: cross-asset anomaly detection and price forecasting

Senior Software Engineer

Pacific Cloud Computing Ltd.

Hong Kong January 2015 - November 2021

- Built an **enterprise document management SaaS** with real-time collaboration for corporate clients
- Built e-commerce platform capabilities including **multi-currency payments**; the SEO programme reached top search rankings
- Pioneered the firm's **blockchain product work**: smart-contract products, NFT marketplace support and DeFi analytics

Senior Product Manager (Technical) / Web & Content Manager

Groupon.com

Hong Kong April 2013 - December 2014

- Led **platform modernisation** from a monolithic to a service-oriented architecture as technical product manager, coordinating engineering delivery and business stakeholders
- Introduced an **A/B testing and SEO programme** that delivered **150% traffic growth**
- Ran digital product operations and commercial prioritisation for the marketplace: merchandising, pricing structure and merchant coordination in an Agile environment

Senior Operations Manager

SoManyCall Telecom

Hong Kong March 2008 - April 2013

- Oversaw strategic and operational delivery for custom software solutions in the telecoms sector, achieving a **25% annual growth rate**
- Led a multi-disciplinary team across the full project lifecycle

CORE TECHNICAL COMPETENCIES

Machine Learning & Deep Learning

Python | PyTorch | Scikit-learn | XGBoost | TensorFlow | Transformers | Decision Trees | LSTM/GRU | Fine-tuning | Feature Engineering

MLOps & Production ML

Automated Retraining | Model Evaluation | A/B Testing | Canary Deployment | Drift Monitoring | Model Serving | ONNX | Triton | Quantisation

Backend, APIs & Microservices

FastAPI | Microservices | REST APIs | Node.js | TypeScript | React | Next.js | Event-Driven | Streaming / SSE

Cloud, DevOps & Infrastructure

AWS | GCP | Cloudflare | Docker | Kubernetes | Terraform | CI/CD | OpenTelemetry | Multi-Region Failover

Data Pipelines & Storage

PostgreSQL | Redis | TimescaleDB | pgvector | Firestore | Cloudflare D1 / R2 | Batch Pipelines | Real-Time Pipelines

Generative AI & LLM Systems

RAG | LLM Orchestration | LangChain | Multi-Provider Routing | Prompt Registry | Guardrails | Function Calling | Eval Harness

AI-Assisted Engineering

Claude Code | Codex CLI | Gemini CLI | Cursor | Eval Harness | Three Review Gates

EDUCATION

MSc Computer Science, AI & Data Science

University of Wolverhampton, UK

2023 - 2025 | Grade: Merit

Dissertation: LLM-Augmented High-Frequency Trading Strategy Development (AI-assisted three-gate workflow; 85% less strategy development time)

Modules: Deep Machine Learning, Intelligent Agents, Data Science & Mining, Applying AI, Cloud Computing, Research Methods

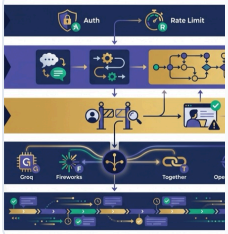
Bachelor of Business Administration

Hong Kong University of Science & Technology

1995

Marketing with Information Systems minor

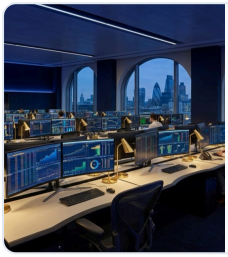
SELECTED DEPLOYMENTS



Qorinix Inference Control Plane

Designed and shipped with AI agents under my review

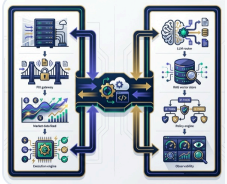
Problem: the board wanted "instant" AI responses for real-time agents, voice and trading alerts, inside a latency budget, provider cost and failover constraints.
Shipped: multi-tier router across 12 providers, streaming SSE, per-call cost ledger, usage-based billing and entitlements.
Outcome: targets sub-200ms TTFT (p50), measured as TTFT p50/p95 per provider in the LLM Arena benchmark; token tiering cut API spend for non-latency-critical work.



LaSpend, Open Banking Assistant

Designed and shipped with AI agents under my review

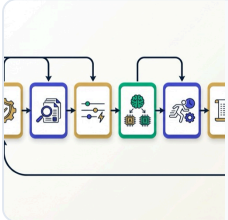
Problem: read-only AI money assistant over FCA-regulated Open Banking data; PII minimisation, never moving money.
Shipped: Open Banking API, Stripe, Cloudflare Pages + Functions + D1; human approval for every user-triggered action, deterministic fallback, verified action receipts.
Outcome: live in production; measured by recurring-spend detection precision and cancellation completion rate.



Fixxmi, AI Lead Matching

Designed and shipped with AI agents under my review

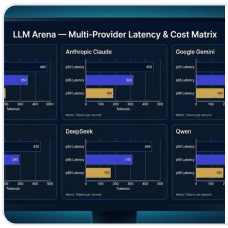
Problem: Swiss consumer marketplace needed AI lead matching across 12+ categories in DE-CH/EN under GDPR and nFADP, with Firestore europe-west6 residency.
Shipped: Next.js 15, Firebase Cloud Functions, Firestore, pay-per-lead Stripe credit packs.
Outcome: live; measured by match acceptance rate.



Audit-Grade Governance Loop

Built

Problem: every AI call across four products must be policy-checked, costed and auditable, with a human path for high-risk cases.
Shipped: policy check, evidence-pack retrieval, prompt registry, model routing, action runtime, audit ledger, human-review escalation, policy feedback.
Outcome: applied to every AI call; measured by audit completeness and escalation rate.



Portfolio Assistant, personal site

Built

Problem: a live, inspectable proof of hands-on work: a chatbot that must answer fast, stay inside staged daily budgets and hold a hardened system prompt.
Shipped: Next.js edge runtime on Cloudflare Pages; five-provider cascade (Groq, Cerebras, Gemini, DeepSeek, OpenRouter) with timeouts and fallback; signed-cookie rate limiting.
Outcome: live on my portfolio site; measured by fallback rate and p95 response time.

SELECTED CERTIFICATIONS

MLOps Specialization
Google Cloud | 2024

Gemini Certified Educator
Google | 2025-2028

CS50x Computer Science
Harvard | 2024

SFC Licensing Exams (Papers 1/7/8/12)
HKSI, Hong Kong | 2021

Open to Principal Forward Deployed, Applied AI and AI Solutions Engineering roles. Full UK right to work | Based near London | Hybrid or remote UK | Languages: English, Cantonese, Mandarin

Portfolio on request

Last Updated: September 2026 · References available upon request