

PHILIP CHEUNG

AI Product & Transformation Lead

Financial Services | Fintech | Applied GenAI

🎓 MSc Computer Science, AI & Data Science (Merit)

📍 Reading, United Kingdom



PROFESSIONAL SUMMARY

AI Product and Transformation Lead with more than 10 years across financial services, fintech, quantitative markets and digital platforms, currently Head of AI at a Swiss corporate finance advisory firm. Translates commercial and financial-services problems into prioritised AI use cases, then owns product vision, MVP scope, roadmaps, requirements, evaluation criteria and governance through to adoption. The current portfolio spans an ultra-low-latency AI inference cloud (Qorinix), a read-only AI money assistant over Open Banking (LaSpend), a Swiss consumer marketplace (Fixxmi) and an AI-native trading desk. Bridges executives, users, domain experts, technical contributors and vendors, and accelerates discovery and delivery through AI-assisted prototyping and coding agents. Comfortable deep in the technical detail, from LLM, RAG and agentic patterns to inference latency and cost trade-offs, evaluation harnesses and cloud delivery: able to shape requirements, evaluate outputs and direct governed delivery with confidence. MSc Computer Science, AI and Data Science (Merit).



CORE CAPABILITIES



AI Product & Transformation

- AI strategy and use-case prioritisation
- Product vision, MVP definition and roadmaps
- Business cases, value hypotheses and adoption planning
- Transformation operating models



Product & Delivery Leadership

- Product discovery and user needs
- Requirements, PRDs, user stories and acceptance criteria
- Backlog and release prioritisation
- Cross-functional delivery coordination
- Stakeholder and executive communication



Responsible AI & Governance

- Output and model evaluation
- Human-review workflows
- Data, privacy, auditability and risk considerations
- Provider and vendor evaluation
- Quality, cost and operational controls



Financial Services Domain

- Corporate finance and M&A workflows
- Investment analysis
- Quantitative markets and trading
- Fintech operations
- Risk, compliance and regulated processes



AI Solution Fluency

- LLM and RAG concepts
- Agentic and workflow-based AI
- API, data integration and cloud-delivery concepts
- AI-assisted prototyping
- Evaluation and quality assurance



How I Deliver

- Problem-first use-case selection
- Smallest product slice that proves value
- Evaluation criteria defined before build
- AI-assisted prototyping and coding agents
- Quality, cost and latency governed from day one



PROFESSIONAL EXPERIENCE

Head of AI

Weidmann & Cie. AG

📍 London, UK (Hybrid) 📅 December 2024 - Present

Swiss corporate finance advisory firm (M&A and growth advisory) building an applied-AI product portfolio: Qorinix (ultra-low-latency AI inference cloud), LaSpend (read-only AI money assistant over Open Banking), Fixxmi (Swiss consumer marketplace), pricing and insurance comparison workflows, and an AI-native trading desk.

- Owns the firm's **AI strategy and use-case portfolio** across its product lines, prioritising investment against commercial value, risk and regulatory constraints; presents strategy and progress to a finance-led management board
- Owns **0-to-1 product direction for Qorinix**, the firm's ultra-low-latency AI inference cloud targeting **sub-200ms time-to-first-token (p50), 100-200+ tokens per second and sub-10ms cache-hit latency** for latency-sensitive workloads such as real-time agents, voice and trading alerts: MVP scope, phased delivery, release gates and go-live criteria across the web, control-plane and AI runtime layers
- Defined the platform's **evaluation and governance model**: latency, quality and cost benchmarks, provider routing and failover, per-call cost ledger, usage-based billing and entitlements, with human-review workflows where risk requires
- Shaped consumer-facing propositions: **LaSpend**, a read-only AI money assistant over regulated Open Banking data (subscription detection, waste scoring and assisted cancellation, never moving customer money, human-in-the-loop for user-triggered actions), and **Fixxmi** AI lead matching across 12+ service categories (DE-CH/EN) within GDPR and nFADP boundaries
- Translates business, operational and regulatory requirements into **executable platform plans** with senior stakeholders, and directs delivery through technical contributors and **AI-assisted development workflows**
- Commissioned an internal **12-provider inference benchmark** to ground model selection, latency and cost decisions; leads a cross-functional team of five across development and financial analysis

AI Engineer & Technical Architect (AI Transformation & Delivery Leadership)

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.

- Defined and prioritised the firm's **AI use-case portfolio and adoption roadmap**, moving AI from experiments into production capabilities; oversaw the launch of **three multi-tenant SaaS platforms** with subscription management, usage analytics and automated billing
- Directed delivery of the firm's real-time ML inference capability, a platform that served **millions of daily predictions at sub-100ms latency**; set requirements, acceptance criteria and evaluation gates
- Shaped the firm's **enterprise retrieval (RAG)** and market-intelligence NLP capabilities, defining evaluation methodology and quality expectations for domain-specific queries
- Oversaw the **model-improvement cycle** (retraining, A/B testing and monitoring) and provider selection across the ML estate
- Established AI/ML working practices, ran architecture reviews and **mentored an agile development team of 8 engineers**
- Guided **quantitative-markets decision support**: cross-asset anomaly detection and price-forecasting research feeding trading workflows

Senior Software Engineer

Pacific Cloud Computing Ltd.

📍 Hong Kong 📅 January 2015 - November 2021

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

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



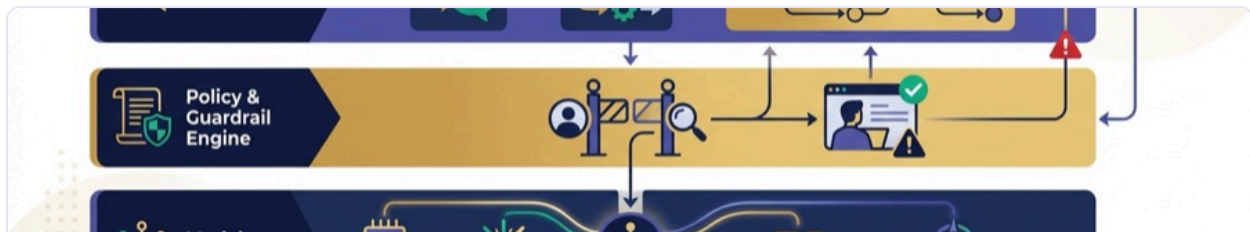
Data Pipelines & Storage

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



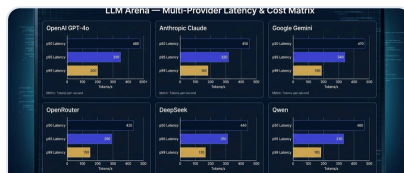
Generative AI & LLM Systems

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



Qorinix, Ultra-Fast AI Inference Cloud FLAGSHIP AI

High-speed AI inference cloud for latency-sensitive workloads: real-time agents, voice assistants, trading alerts and conversational AI. Targets **sub-200ms TTFT (p50)**, **100-200+ tokens/sec** and **sub-10ms cache-hit latency** at a fraction of frontier-API cost; internal benchmark targets place it 4-12x faster than leading low-cost providers and 6-12x faster than frontier APIs. Multi-tier model router with provider failover, streaming SSE, per-call cost ledger, usage-based billing and entitlements. Stack: Python, FastAPI, PostgreSQL, Redis, LLM Router, Stripe entitlements, OpenTelemetry.



LLM Arena, 12-Provider Benchmark LLM INFRA

Side-by-side real-time benchmark harness across **12 frontier and fast-inference providers**. Streaming token-by-token with live leaderboard, TTFT p50/p95 analytics, tokens/sec, estimated cost-per-task and service-tier selectors. Per-provider model picker, run-count (1-20) for statistical stability and quick TTFT-turbo profiles. Next.js 14, TypeScript, server-side API isolation, streaming-SSE.



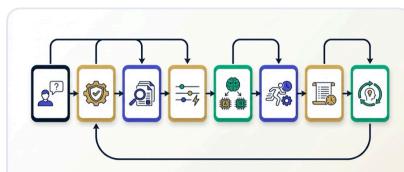
LaSpend, AI Money Assistant CONSUMER AI

Read-only AI money assistant over regulated UK Open Banking: **subscription detection, recurring-spend dashboard, waste scoring and assisted cancellation** with verified action receipts. Never moves customer money; human-in-the-loop for all user-triggered actions; PII-minimised inputs and deterministic fallback when models fail. Cloudflare Pages + Functions + D1 architecture, React 19 + Vite + Tailwind frontend.



Fixxmi, AI Swiss Service Marketplace CONSUMER AI

Swiss-first B2C service-job lead marketplace with AI-powered intent classification and lead matching across 12+ categories in DE-CH / EN. Pay-per-lead monetisation with Stripe Credit Packs, nFADP / GDPR-compliant data boundaries, Firestore-backed real-time admin dashboard. Next.js 15 static export + Firebase Cloud Functions (Node 20) + Firestore europe-west6.



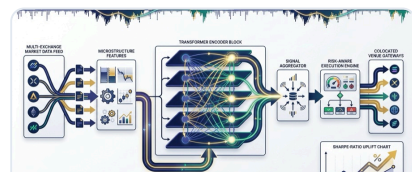
Audit-Grade AI Governance Loop GOVERNANCE

Closed-loop flow applied across every firm AI product: policy check, evidence-pack retrieval, prompt control, model router, action runtime, audit ledger, policy feedback. Every call versioned, costed, support-traceable and replayable. Content-addressed evidence packs, per-product cost attribution, human-in-the-loop escalation for high-risk cases.



Low-Latency Inference Mesh ML INFRA

Distributed ML inference pipeline serving **hundreds of predictions per day at <200ms TTFT** across classification, sentiment, affordability scoring and anomaly detection. Shard-aware routing, warm-pool autoscaling, ONNX-compiled models with fp16 quantisation, cache-through Redis tiering, multi-region active-active with health-weighted failover. FastAPI, Redis, TimescaleDB, Triton.



LLM-Augmented HFT (MSc Research) RESEARCH

Novel architecture fusing LLM narrative understanding with real-time HFT. Custom transformer with attention heads over microstructure features, ensemble gating and latency-aware inference scheduling that down-routes heavy heads on tight time budgets. AI-assisted development workflow with a **three-gate validation framework** delivered an **85% reduction in strategy development time**. Python, PyTorch, CCXT, Ray Serve, Triton.



EDUCATION



MSc Computer Science, AI & Data Science

University of Wolverhampton, UK

2023 - 2025 | Grade: Merit

Dissertation: LLM-Augmented High-Frequency Trading Strategy Development

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 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 AI Product, AI Transformation and Head of AI roles across financial services, fintech and applied GenAI.

Full UK right to work | Based near London (Reading) | Hybrid or remote | Languages: English, Cantonese, Mandarin

 **Portfolio available on request**

Last Updated: July 2026 · References available upon request