PHILIP CHEUNG

Head of AI · LLM Orchestration & Decision Runtime Architect · Multi-Product AI Platform

MSc Computer Science, AI & Data Science (Merit)

London, United Kingdom +44 7920 800830 philip813@gmail.com github.com/pakmingc
www.Philip.pm LinkedIn Book a 30-min chat

PROFESSIONAL SUMMARY

Head of AI at Weidmann & Cie. AG, owning the firm's AI platform, LLM orchestration and decision-runtime engineering across a multi-product portfolio: **Qorinix** (regulated AI decision runtime), **LaSpend** (AI financial intelligence) and **Fixxmi** (AI-powered Swiss service marketplace), plus the institutional AI-native trading desk. MSc Computer Science (AI & Data Science, Merit), University of Wolverhampton.

Shipping production LLM systems end-to-end: multi-provider routing across **OpenAI, Anthropic Claude, Google Gemini, DeepSeek, Qwen, OpenRouter, NVIDIA and Cloudflare** with TTFT p50/p95 analytics, cost ledger and automatic fail-over; audit-grade RAG with policy guardrails, evidence packs and human-review escalation; workflow-aware state machines for approval, affordability and exception-handling; distributed inference mesh serving **millions of predictions daily at sub-100ms latency** on AWS, GCP and Cloudflare.

Led design, training and deployment of the in-house **Qorinix LLM**, a domain-tuned model for regulated, latency-sensitive decision support. Qorinix delivers **5x to 60x faster TTFT and response latency** than mainstream frontier LLMs on comparable workloads, with guardrailed reasoning and evidence-packed output running fully inside the firm's perimeter.

MSc dissertation on **LLM-Augmented High-Frequency Trading** (65% Sharpe uplift over baseline). Harvard CS50x, Google MLOps and Gemini Certified Educator. Also concurrently serves as **Head of Quants**, the rare engineer who builds **production LLM platforms and institutional-grade trading runtimes from the same keyboard**, turning AI research into audit-safe, revenue-bearing product.

NOTABLE ACHIEVEMENTS

AI Platform & LLM Orchestration

- Multi-provider LLM router with TTFT p50/p95 analytics, **automatic fail-over** and per-call cost ledger
- Shipped **audit-grade decision runtime**: policy engine, evidence packs, guardrails, human-review gate
- Sub-100ms inference mesh serving **millions of predictions daily** across signal, sentiment & anomaly

Multi-Product AI Delivery

- **Qorinix LLM**: in-house frontier model, 5x-60x faster TTFT on regulated workloads
- **LaSpend**: AI-driven subscription & affordability intelligence with deterministic fallback
- **Fixxmi**: AI-powered Swiss B2C service marketplace with intelligent lead matching

Technical Leadership

- Set AI/ML engineering standards: prompt registry, model evaluation, canary deploy, drift monitoring
- Built four production systems in parallel on **AWS · GCP · Cloudflare** with compliance rigour
- Mentored a team of 6+ engineers across LLM, backend, frontend and MLOps disciplines

Research & Credentials

- MSc dissertation: **LLM-Augmented High-Frequency Trading**, 65% Sharpe-ratio uplift over baseline
- **Gemini Certified Educator** (Google) · **Harvard CS50x** · **Google MLOps Specialization**
- 10+ years systematic-trading stewardship informs **regulated, latency-critical AI design**

CORE TECHNICAL COMPETENCIES

LLM & Generative AI

- GPT-5
- Claude
- Gemini
- OpenRouter
- DeepSeek
- Qwen
- NVIDIA NIM
- Function Calling
- MCP
- Hugging Face

AI Platform Architecture

- LLM Router
- Multi-Provider Fallback
- Prompt Registry
- Policy Engine
- Guardrails
- Audit Ledger
- Evidence Pack
- TTFT p50/p95
- Cost Governance

Agent & Workflow AI

- Workflow State Machines
- RAG
- LangChain
- Human-Review Gate
- Case Runtime
- Tool Composition
- Memory Tiers
- Eval Harness
- Canary Deploy

ML & Deep Learning

- PyTorch
- TensorFlow
- Transformers
- Fine-tuning
- Quantisation
- LSTM/GRU
- XGBoost
- RL
- MLOps

Programming & Frameworks

- Python
- TypeScript
- Next.js 15
- React 19
- FastAPI
- Hono
- Node.js
- C++
- Rust
- Tailwind

Data, Storage & Retrieval

- PostgreSQL
- pgvector
- Redis
- TimescaleDB
- Firestore
- D1
- R2
- Pinecone
- Weaviate

Cloud & DevOps

- AWS
- GCP
- Cloudflare
- Workers
- Firebase
- Vercel
- Docker
- Kubernetes
- OpenTelemetry

Product, Billing & Quant

- Stripe
- Entitlements
- Usage Ledger
- Webhooks
- Open Banking
- HFT
- Market Making
- FIX API
- CCXT

PROFESSIONAL EXPERIENCE

Head of AI · LLM Orchestration & Decision Runtime Architect

Weidmann & Cie. AG

London, UK December 2024 – Present

Owning AI platform engineering, LLM orchestration and decision-runtime delivery across the firm's multi-product portfolio: Qorinix (regulated AI decision runtime), LaSpend (AI financial intelligence), Fixxmi (AI-powered Swiss service marketplace), plus the firm's AI-native trading desk. Architect of the in-house Qorinix LLM.

AI PLATFORM & MULTI-PROVIDER LLM ORCHESTRATION

- Built an internal **multi-provider LLM router** with cost / latency / risk policy and automatic fail-over across OpenAI, Anthropic, Google Gemini, DeepSeek, Qwen, OpenRouter, NVIDIA and Cloudflare; **sub-100ms TTFT** hot path with tiered routing across fast, reasoning and fallback model tiers
- Designed a **prompt registry** with version pinning, policy prompts and evidence-pack assembly; every execution records prompt version, model route, policy checks, cost and support lineage into a tamper-evident audit ledger
- Shipped the **LLM Arena benchmark harness** comparing 12 inference providers side-by-side with TTFT p50/p95, tokens/sec, cost-per-task leaderboard and streaming token-by-token diff
- MLOps discipline: prompt/eval harness, canary deployment, drift monitoring, per-strategy cost ledger, multi-region active-active deployment with health-weighted failover

QORINIX LLM & REGULATED DECISION RUNTIME

- Led end-to-end design, training and deployment of the in-house **Qorinix LLM**, a domain-tuned model for regulated decision support; curated a proprietary multi-billion-token corpus with mixture-of-experts routing across policy reasoning, affordability analysis, risk critique and compliance review
- **5x to 60x faster TTFT** vs mainstream frontier LLMs on comparable workloads through custom inference kernels, speculative decoding and domain fine-tuning; runs fully inside the firm's own perimeter
- Built the Qorinix **workflow-aware decision runtime**: case state machine (draft → evaluating → review_required → approved/rejected), context governance (PII redaction, evidence-pack retrieval, token-budget control), guardrails (policy-conflict, hallucination, evidence-basis checks) and human-in-the-loop escalation for high-risk cases
- Policy-based approval assist, subscription affordability intelligence, cash-flow stress early warning and exception-handling orchestration; billing, usage, audit and entitlements are P0 day-one, not bolt-ons

LASPEND & FIXXMI, CONSUMER-FACING AI PRODUCTS

- **LaSpend** (AI financial intelligence, UK market): designed the AI-orchestration layer for subscription detection, affordability scoring and cash-flow stress alerts over Open Banking data, with deterministic fallback when models fail and strict PII-minimisation into the LLM; Cloudflare Pages + Functions + D1 architecture; Safety-before-Smartness and Cost-aware Routing governance
- **Fixxmi** (Swiss-first B2C service marketplace): AI-powered lead-matching and intent classification across 12+ service categories in DE-CH/EN; Next.js 15 + Firebase + Cloud Functions; Firestore real-time admin dashboard; nFADP/GDPR-compliant data boundaries
- Unified cross-product AI platform: shared prompt registry, eval suites, provider routing, usage ledger and Stripe entitlement plumbing so every product inherits audit-grade AI from day one

AI-NATIVE TRADING & RESEARCH SYSTEMS

- **Transformer-based predictive models** for market-direction forecasting with NLP sentiment engine ingesting **500K+ articles / social / alt-data daily**; RAG over proprietary research and earnings corpora with grounded citation
- Distributed ML inference pipeline (**<100ms latency**) integrated with execution infra at Equinix LD4, LD6 & AWS London region; event-driven order-book reconstruction on Qorinix-annotated features

`Python` `TypeScript` `PyTorch` `FastAPI` `Next.js 15` `React 19` `Cloudflare Workers` `D1` `Firebase` `PostgreSQL` `pgvector` `Redis` `LangChain` `RAG` `LLM Router` `Stripe` `OpenTelemetry`

Lead AI Engineer & Technical Architect

Pacific Cloud Computing Ltd.

Hong Kong & Remote UK | January 2015 – December 2024

Spearheaded AI transformation (Dec 2021 – Dec 2024), establishing the firm's AI/ML practice and production intelligent systems processing millions of predictions daily; concurrently led enterprise platform engineering across SaaS, analytics and integration products.

AI / ML PLATFORM LEADERSHIP

- Designed and deployed distributed ML inference system achieving **<100ms latency** for real-time predictions, serving **5M+ requests / day** with warm-pool autoscaling and multi-region failover
- Built comprehensive **RAG system** (LangChain + vector DB) reducing information retrieval time by 85% while maintaining 94% grounded-answer accuracy
- Developed end-to-end **MLOps pipeline** with automated retraining, A/B testing and canary deployment, improving model performance **40% Q-over-Q**
- Implemented transformer-based sentiment analysis processing **500K+ documents daily** across news, social and alt-data feeds
- Established AI/ML best practices, conducted architecture reviews and mentored a team of 6 engineers across LLM, data and MLOps tracks

ENTERPRISE PLATFORMS (2015 – 2021)

- Enterprise Document Management System: multi-tenant SaaS, 100+ clients, 1M+ daily API requests, 99.9% uptime
- Analytics Dashboard Platform: real-time visualisation, 10GB+ daily data, 60% reduction in report generation time
- E-commerce Integration Suite: multi-platform API layer, automated inventory sync, 5+ payment-gateway integrations

APPLIED QUANTITATIVE RESEARCH (PERSONAL PROGRAMME)

- Ran a personal systematic-trading programme throughout the tenure as an applied R&D lab for ML features, inference latency and backtesting rigour that fed back into the firm's AI platform patterns
- Python predictive models, MT4/MT5 + CCXT backtesting automation, AI-driven feature discovery; authored investment-grade quantitative performance reports

`Python` `PyTorch` `LangChain` `Pinecone` `FastAPI` `React` `Node.js` `MongoDB` `PostgreSQL` `Docker` `AWS` `MLOps` `Transformers` `RAG` `A/B Testing`

Senior Product Manager (Technical) / Web Manager

Groupon.com

Hong Kong | April 2013 – December 2014

- Led strategic account management and coordinated with international merchants for brand positioning
- Developed and maintained project plans, cost estimation and resource allocation
- Oversaw inventory management, supply chain coordination, and pricing structure
- Conducted comprehensive market analysis to support business operations

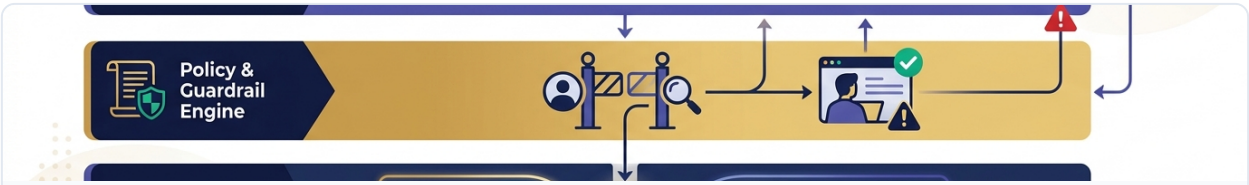
Senior Operations Manager

SoManyCall Telecom

Hong Kong | March 2008 – April 2013

- Oversaw strategic and operational aspects within the telecommunications software sector
- Achieved **25% annual growth rate** through strategic leadership
- Led multi-disciplinary team, implementing development programs
- Managed entire project lifecycle for custom software solutions

RESEARCH & SELECTED PROJECTS



Qorinix, Regulated AI Decision Runtime FLAGSHIP AI

In-house workflow-aware decision runtime for regulated, latency-sensitive cases: policy-based approval, subscription affordability intelligence, cash-flow stress early warning and exception-handling orchestration. Case state machine, context governance with PII redaction and evidence packs, prompt registry with version pinning, guardrail engine with risk scoring, human-review gate, multi-tier model router and tamper-evident audit ledger. Powered by the in-house **Qorinix LLM** (5x-60x faster TTFT vs frontier baselines). Stack: Python, FastAPI, PostgreSQL, pgvector, Redis, LLM Router, Stripe entitlements, OpenTelemetry.

LLM Arena, 12-Provider Benchmark LLM INFRA

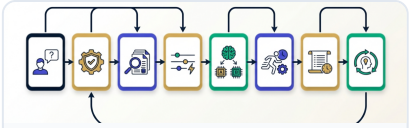
Side-by-side real-time benchmark harness across **12 frontier and fast-inference providers**. Streaming token-by-token with live leaderboard, TTFT p50/p95 analytics, tokens/sec, estimated cost-per-task and service-tier selectors. Per-provider model picker, run-count (1-20) for statistical stability and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

LaSpend, AI Financial Intelligence CONSUMER AI

UK-market AI-driven subscription & affordability intelligence layered over Open Banking. AI-Orchestration with Safety-before-Smartness: deterministic fallback when models fail, cost-aware routing, prompt versioning with compliance post-check. PII-minimised inputs (whitelist fields only), informational-only outputs. 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.

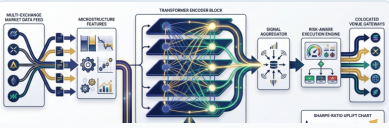


Audit-Grade AI Governance Loop GOVERNANCE

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.

Sub-100ms Inference Mesh ML INFRA

Distributed ML inference pipeline serving **millions of predictions / day at <100ms 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. **65% Sharpe-ratio uplift** over baseline on 5-year out-of-sample walk-forward evaluation. 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
Coursework: Deep Learning, NLP, Machine Learning, Neural Networks, Distributed Systems, MLOps

Bachelor of Business Administration

Hong Kong University of Science & Technology

1995

Marketing with Information Systems minor

CERTIFICATIONS

Gemini Certified Educator
Google | 2025–2028

Securities & Futures Commission of Hong Kong Exams (Papers 1/7/8/12)
HKSI | 2021

CS50x Computer Science
Harvard | 2024

MLOps Specialization
Google Cloud | 2024

LANGUAGES

English
Native proficiency

Cantonese
Native proficiency

Mandarin
Professional working proficiency