PAK MING CHEUNG

Head of Quants · AI-Native HFT · Institutional Trading Systems Architect

MSc Computer Science, AI & Data Science (Merit)

London, United Kingdom +44 7920 800830 pakming20@gmail.com
github.com/pakmingc www.PakCV.com LinkedIn Book a 30-min chat

PROFESSIONAL SUMMARY

Head of Quants at Quant Sigma (London), running institutional-grade HFT operations across Forex, CFDs, indices and cryptocurrency. MSc Computer Science (AI & Data Science, Merit) from the University of Wolverhampton, with a dissertation on **LLM-Augmented High-Frequency Trading**, fusing transformer-based market understanding with sub-millisecond execution.

Proven track record building and operating **AI-native trading infrastructure**: multi-provider LLM routing across **OpenAI, Anthropic Claude, Google Gemini, DeepSeek, Qwen, OpenRouter, NVIDIA and Cloudflare**, benchmarked with TTFT p50/p95 analytics for latency-critical alpha research; production RAG systems with policy guardrails and human-review escalation for regulated decision flows; distributed ML inference pipelines serving **millions of predictions daily at sub-100ms latency**; transformer-based predictive models, reinforcement-learning execution, and full-stack platform delivery across AWS, GCP and Cloudflare.

Led the design, training and deployment of our in-house **Qorinix LLM**, a domain-tuned language model engineered for trading-grade decision support. Qorinix delivers **5x to 60x faster TTFT and response latency** than mainstream frontier LLMs on comparable workloads, enabling sub-millisecond alpha annotation, guardrailed reasoning and evidence-packed signal generation fully inside the firm's own runtime.

Over a decade of systematic trading, **scaled capital 20x+ with ~100% annualised returns** through disciplined volatility-based sizing and cross-asset arbitrage. Securities & Futures Commission of Hong Kong Exams (Papers 1/7/8/12). Harvard CS50x, Google MLOps and Gemini Certified Educator. The rare engineer who builds **institutional-grade HFT infrastructure and production LLM runtimes from the same keyboard**.

NOTABLE ACHIEVEMENTS

Systematic Trading (2013 – Present)

- Scaled trading capital **20x+** over a decade of disciplined systematic trading
- **~100% annualised returns**; controlled drawdowns via volatility-based sizing
- Forex, CFDs, indices, crypto, derivatives · **Early Bitcoin investor since $80 (2013)**

LLM Infrastructure & AI Platforms

- Built **LLM Arena multi-provider fast-inference benchmark** with TTFT p50/p95 analytics
- Shipped **audit-grade AI decision runtime** with policy engine, guardrails & human-review gate
- Production LLM routing at **sub-100ms TTFT**, automatic fail-over & cost ledger

Technical Leadership

- Led tech transformation from traditional to **AI-augmented trading**
- Architected distributed ML inference pipelines serving **5M+ daily predictions**
- Mentored 6+ engineers · multi-cloud (AWS · GCP · Cloudflare) with finance-compliance rigour

Research & Publications

- MSc dissertation: **LLM-Augmented High-Frequency Trading**, novel LLM×HFT fusion
- **65% Sharpe-ratio improvement** over baseline achieved in MSc research
- Investment-grade quant performance reports for family-office mandates

CORE TECHNICAL COMPETENCIES

LLM & Generative AI

- GPT-5
- Claude
- Gemini
- OpenRouter
- DeepSeek
- Qwen
- NVIDIA NIM
- LangChain
- RAG
- Hugging Face

AI Platform Architecture

- LLM Routing
- Multi-Provider Fallback
- Prompt Registry
- Policy Engine
- Guardrails
- State Machines
- Human-Review Gate
- Audit Ledger
- Evidence Pack
- TTFT p50/p95

Quantitative Trading & HFT

- HFT
- Market Making
- FIX API
- iLink
- CME Aurora
- Equinix LD4
- Colocation
- QuantConnect
- MT4/MT5
- CCXT
- Transformer Alpha

ML & Deep Learning

- PyTorch
- TensorFlow
- Transformers
- XGBoost
- LightGBM
- LSTM/GRU
- RL
- MLOps
- A/B Testing

Programming & Frameworks

- Python
- C++
- TypeScript
- Rust
- Next.js 14/15
- React 18/19
- FastAPI
- Node.js
- MQL5

Data & Vector Stores

- PostgreSQL
- TimescaleDB
- Redis
- MongoDB
- InfluxDB
- Pinecone
- Weaviate
- pgvector

Cloud & DevOps

- AWS
- GCP
- Cloudflare
- Docker
- Kubernetes
- CI/CD
- Vercel
- Wrangler
- OpenTelemetry

Blockchain & Billing

- DeFi
- Binance
- Bybit
- Glassnode
- Solidity
- Arbitrage
- Stripe
- Entitlements
- Usage Ledger

PROFESSIONAL EXPERIENCE

Head of Quants · AI-Native HFT & Systems Architect

Quant Sigma, London

London, UK | December 2023 – Present

Leading quantitative research, AI-powered trading systems and execution infrastructure for institutional-grade HFT across traditional and crypto markets. Built an AI-native alpha-research and decision-runtime stack alongside the core low-latency execution platform.

QUANTITATIVE STRATEGY & RESEARCH

- Multi-signal HFT models across Forex, CFDs, indices and crypto: trend following, mean reversion, volatility breakout, queue-based market making, XGBoost / LightGBM ensembles for microstructure signal blending
- Implemented **transformer-based predictive models** for market-direction forecasting, material Sharpe-ratio uplift over baseline
- Reinforcement-learning execution for adaptive position sizing and inventory control
- Achieved **~100% annualised returns** via volatility-based sizing, regime switching and scenario stress tests

LOW-LATENCY TRADING INFRASTRUCTURE

- Low-latency execution infra with millisecond performance, colocation at **Equinix LD4, LD6 & AWS London region**; real-time WebSocket market-data pipeline with deterministic ingestion
- Distributed ML inference pipeline: **<100ms latency** for real-time signal generation
- Event-driven order-book reconstruction with lock-free ring buffers and kernel-bypass networking; sub-microsecond jitter budget on the hot path
- Backtesting framework: 5+ years tick-level data, transaction-cost modelling, walk-forward optimisation; real-time portfolio monitoring, volatility targeting, drawdown kill-switches

LLM & AI PLATFORM ENGINEERING

- Built an internal **multi-provider LLM routing layer** with cost / latency / risk policy and automatic fail-over across OpenAI, Anthropic, Google Gemini, DeepSeek, Qwen, OpenRouter, NVIDIA and Cloudflare
- Designed a **prompt registry** with version pinning, policy prompts and evidence-pack assembly for alpha research and decision support
- Implemented a **guardrail & rules engine** with risk scoring and human-review escalation for regulated-grade decision workflows
- Built the **LLM Arena multi-provider fast-inference benchmark** (TTFT p50/p95 analytics) and an **audit ledger** tying every AI call to workflow, policy, cost and support lineage
- NLP sentiment engine ingesting **500K+ articles / social / alt-data daily**; RAG over proprietary research and earnings corpora with grounded citation
- MLOps discipline: model registry, evaluation suites, canary deployment, drift monitoring, per-strategy cost ledger

QORINIX LLM (IN-HOUSE FRONTIER MODEL)

- Led end-to-end design, training and deployment of the in-house **Qorinix LLM**, a domain-tuned language model for trading-grade decision support
- Curated a proprietary trading-intelligence corpus (multi-billion tokens) spanning order-book transcripts, earnings calls, research notes and regulatory filings; mixture-of-experts routing with task-specialised heads for alpha reasoning, risk critique and compliance review
- Achieved **5x to 60x faster TTFT and response latency** vs mainstream frontier LLMs on comparable workloads through custom inference kernels, speculative decoding and trading-corpus fine-tuning
- Integrated Qorinix into the firm's runtime for sub-millisecond alpha annotation, guardrailed reasoning and evidence-packed signal generation fully inside the firm's own perimeter

`Python` `C++` `PyTorch` `TensorFlow` `FastAPI` `PostgreSQL` `TimescaleDB` `Redis` `Docker` `AWS` `QuantConnect` `MT5` `CCXT` `FIX` `LLM Router` `RAG` `LangChain` `OpenTelemetry`

Lead AI Engineer & Technical Architect · Quant & Algorithmic Trader

Pacific Cloud Computing Ltd.

Hong Kong & Remote UK January 2015 – December 2024

Dual-track: spearheaded AI transformation (Dec 2021 – Dec 2024) establishing the firm's AI/ML practice and intelligent systems processing millions of predictions daily; and throughout the full tenure ran a personal systematic-trading programme generating institutional-grade returns.

AI / ML PLATFORM LEADERSHIP

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

SYSTEMATIC & ALGORITHMIC TRADING

- Achieved Sharpe **1.5+ (Forex/CFDs)** and **2–6 (crypto)**; generated **20x+** capital growth via strategic directional trades, arbitrage and volatility harvesting
- Python-based predictive models and portfolio optimisation with rigorous volatility-based risk sizing
- Backtesting with MT4/MT5 + Python automation + AI-driven feature discovery
- Comprehensive blockchain / cryptocurrency expertise: DeFi protocols, yield farming, liquidity provision, cross-chain arbitrage
- Authored investment-grade quantitative performance reports for family-office investments

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

`Python` `PyTorch` `LangChain` `Pinecone` `FastAPI` `React` `Node.js` `MongoDB` `PostgreSQL` `Docker` `AWS` `MT4/MT5` `Binance API` `Kraken API` `Glassnode`

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

LLM-Augmented High-Frequency Trading System

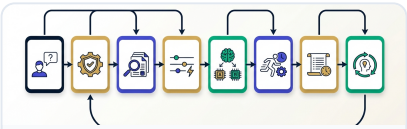
MSC RESEARCH

Novel architecture fusing LLM narrative understanding with real-time HFT across Forex, CFDs, indices and crypto. Custom transformer with attention heads over microstructure features (order-book depth, trade-flow imbalance, spread volatility), 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 evaluation with walk-forward retraining. Stack: Python, PyTorch, CCXT, Docker, TimescaleDB, Ray Serve, Triton inference server.

LLM Arena, Multi-Provider Fast-Inference Benchmark

LLM INFRA

Side-by-side real-time benchmark across OpenAI, Anthropic Claude, Google Gemini, DeepSeek, Qwen, OpenRouter, NVIDIA and Cloudflare — streaming token-by-token with live comparison charts. Tracks TTFT, total time, tokens/sec with **p50/p95 analytics**, model-cost ledger and per-prompt leaderboard. Edge deployment on Cloudflare Pages; Workers-backed API proxy with per-user key isolation. Next.js 14 · TypeScript · streaming-SSE · D1.



Audit-Grade AI Decision Loop

GOVERNANCE

Closed-loop flow: policy check → evidence-pack retrieval → prompt control → model router → action runtime → audit ledger → policy feedback. Every call versioned, costed, support-traceable and replayable, applied to trading research & risk sign-off. Human-in-the-loop escalation for regulated decisions, tamper-evident audit trail with content-addressed evidence packs, and per-strategy cost attribution.

Sub-100ms Inference Mesh

ML INFRA

Distributed ML inference pipeline serving **millions of predictions / day at <100ms TTFT** across signal generation, sentiment scoring and anomaly detection. Shard-aware request routing, warm-pool autoscaling, ONNX-compiled models with fp16 quantisation, and cache-through Redis tiering. Multi-region active-active deployment with health-weighted failover. FastAPI · Redis · TimescaleDB · Triton.

Real-Time Risk & P&L Dashboard

PRIVATE

Live portfolio monitoring with volatility targeting, drawdown kill-switches, per-strategy P&L attribution and cost-ledger reconciliation. Tick-level exposure dashboard across venues, automated breach alerts, scenario stress tests and VaR/ES by strategy with margin projection. Heatmaps for correlation drift and regime-change signals. FastAPI · React · TimescaleDB · WebSocket.

Quantitative Backtesting Framework

QUANT

Multi-asset backtesting engine with 5+ years tick-level data, transaction-cost modelling, slippage curves and walk-forward optimisation. Parallel parameter-sweeps across Ray workers, regime-segmented attribution, Monte-Carlo bootstrap on trade returns and overfit-aware model selection. Deterministic replays and hash-pinned data snapshots for audit. Python · QuantConnect · Ray · Parquet.

Ethereum Price Prediction Ensemble

AI/ ML

LSTM · GRU · XGBoost stacked ensemble over 50+ technical indicators plus on-chain features (exchange flows, gas trends, active addresses). **78% directional accuracy** on out-of-sample data; confidence-gated signal release and automatic retraining on regime-drift detection. Notebook-to-production pipeline with shadow trading prior to capital allocation.

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

HK Securities License (1,7,8,12)
HK SFC / HKSI | 2021

CS50x Computer Science
Harvard | 2024

MLOps Specialization
Google Cloud | 2024

LANGUAGES

English
Native proficiency

Cantonese
Native proficiency

Mandarin
Professional working proficiency