

Jiaming Pei

AI Engineer, New Grad

+1-778-708-0116 | jiamingpei16@gmail.com | Vancouver, BC (open to Toronto) | Eligible to work in Canada

linkedin.com/in/jiaming-theodore-pei | github.com/jmpei | jmpei.dev

EDUCATION

Northeastern University

Vancouver, BC

Master of Science in Computer Science — GPA: 3.85/4.0

Sept. 2024 – Expected Dec. 2026

- Coursework: Machine Learning, Algorithms, Scalable Distributed Systems, Database Management Systems

Minzu University of China

Beijing, China

Bachelor of Economics in Finance

Sept. 2020 – June 2024

- Coursework: Probability and Statistics

PROJECTS

CanPlan — RAG Task-Planning Backend | AWS CDK, Amazon Bedrock, Cohere Rerank

June 2026

- Co-led CanPlan's serverless AWS backend (assistive task-planning app for brain-injury and dementia care) and sole-authored its Amazon Bedrock RAG pipeline producing grounded, cited task steps
- Replaced fixed top-K retrieval with coarse recall followed by Cohere Rerank 3.5 and a dynamic score floor, calibrated against golden queries on an 80-passage HealthLink BC corpus
- Owned the AI progress-reporting subsystem, from metric aggregation to AI-written summaries with presigned S3 downloads, on the team's shared single-table DynamoDB foundation

Financial Sentiment Analysis Agent | LangChain/LangGraph, MCP, LoRA

May 2026

- Built a LangChain 1.0 / LangGraph agent that chains news search with sentiment scoring and treats fetched article text as untrusted data, not instructions
- Evaluated the agent against a deterministic pipeline on 10 trajectory scenarios $\times$ 5 live-LLM runs: the pipeline matched its reliability (50/50 vs 49/50) with $7\times$ fewer tokens
- Served the adapter behind a FastAPI /predict endpoint used by the agent and MCP server, at 25.7 ms p95 warm-CPU inference, plus a live [Gradio demo](#)
- Fine-tuned DistilBERT on FinancialPhraseBank with LoRA to 0.83 weighted F1 (vs a 0.44 majority-class floor), training 1.31% of parameters into a 3.4 MB adapter

EXPERIENCE

Full-Stack Developer

Nov. 2025 – Present

Beijing Boyatang Education Technology Co., Ltd.

Beijing, China (Remote, Contract)

- Owned the company's online course-ordering platform as sole developer: a FastAPI service on PostgreSQL, a React client, and a staff console for orders, refunds and daily reconciliation
- Kept seats from overselling with pessimistic row locks taken in ascending id order inside the placement transaction, verified by a concurrency suite where eight threads contend for one seat and exactly one wins
- Made checkout retry-safe with a per-user idempotency key backed by a unique constraint, tested for replays, cross-user key reuse, and concurrent same-key calls

Quantitative Developer Intern

July 2023 – Sept. 2023

Benefits Mutual Asset Management Co., Ltd.

Beijing, China

- Shipped position, P&L and risk-exposure monitoring that ran daily in production for the desk's portfolio managers, flagging anomalies before the open
- Engineered 8 price-volume factors and a 1-layer LSTM that lifted out-of-sample Sharpe from 1.0 to 1.2, on market data scraped and cleaned in Python (Selenium, BeautifulSoup, pandas)

TECHNICAL SKILLS

Languages: Python, Go, SQL, TypeScript/JavaScript, Java, C

GenAI / LLM: Retrieval-Augmented Generation (RAG), LLM agents (LangChain / LangGraph), Model Context Protocol (MCP), Amazon Bedrock (Knowledge Bases, Converse), Cohere Rerank, vector databases (Amazon S3 Vectors), prompt engineering, trajectory evals, Langfuse

ML: PyTorch, Hugging Face Transformers, LoRA fine-tuning, model quantization and distillation, scikit-learn, pandas, NumPy

Backend / Cloud: FastAPI, REST & GraphQL APIs (AppSync), microservices, Docker, AWS (Lambda, DynamoDB, S3, ECS Fargate, RDS), AWS CDK, Terraform, PostgreSQL, MySQL, GitHub Actions (CI/CD), Linux