Jiaming Pei

+1-778-708-0116 | jiamingpei16@gmail.com | Vancouver, BC (open to Toronto) | Eligible for co-op work in Canada
jiaming-theodore-pei | jmpei | jmpei.dev

Education

Northeastern University — Vancouver, BC
Master of Science in Computer Science — GPA: 3.85/4.0 — *Sept. 2024 – Aug. 2027*

Minzu University of China — Beijing, China
Bachelor of Economics in Finance — *Sept. 2020 – Jun. 2024*

Technical Skills

Languages: Python, TypeScript/JavaScript, SQL, Go, Java
GenAI / LLM: RAG (retrieval-augmented generation), Amazon Bedrock (Knowledge Bases, Converse), vector search (OpenSearch Serverless), LangChain, prompt & tool-calling, LoRA / fine-tuning, HuggingFace Transformers
ML: PyTorch, scikit-learn, model quantization (PTQ / QAT), pandas, NumPy
Backend / Cloud: AWS (Lambda, DynamoDB, S3, AppSync GraphQL, ECS, SQS, RDS), AWS CDK, Terraform, FastAPI, Docker, REST APIs, microservices, MySQL / MongoDB, Git, Linux

Projects

CanPlan — RAG Task-Planning Backend | *TypeScript, AWS CDK, Bedrock KB, OpenSearch, DynamoDB, Lambda*

- Built the RAG step-generation backend for CanPlan, a real-client assistive app for people with brain injury or dementia and their caregivers, turning a high-level task into grounded, citation-backed steps. The GraphQL resolver retrieves the top-4 passages from a Bedrock Knowledge Base (OpenSearch Serverless vector index), then calls Bedrock Converse (Claude Sonnet)
- Provisioned the serverless backend as AWS CDK (TypeScript): AppSync GraphQL with Cognito JWT auth, a single-table DynamoDB design (composite PK/SK over 9 entity types), and per-domain Lambda resolvers
- Deployed across two AWS regions via CDK cross-region references. Built a corpus-ingestion pipeline (chunked text + metadata → S3 → Bedrock ingestion job) that populates the vector store

Financial Sentiment Analysis Agent | *PyTorch, HuggingFace, LoRA, FastAPI, Docker, LangChain, HuggingFace Spaces*

- Fine-tuned DistilBERT with LoRA on FinancialPhraseBank (4,846 samples), training only 1.31% of parameters (887K of 67.8M) to reach weighted F1 = 0.83 on the held-out test split. Produced a 3.4 MB adapter from a 4.9-minute run on an Apple M3 Pro
- Evaluated model calibration with one-vs-rest reliability curves, surfacing negative-class overconfidence (predicted confidence 0.96 vs 0.73 accuracy) as a candidate for temperature scaling
- Served the model via FastAPI with lifespan loading (p95 < 30 ms warm CPU inference) and Docker, and shipped a public Gradio demo on HuggingFace Spaces. Built a LangChain agent (gpt-4o-mini, tool-calling) that calls the deployed `/predict` endpoint plus NewsAPI to answer financial questions end-to-end

CRDNN Voice-Activity Model Compression | *PyTorch, SpeechBrain, static/dynamic PTQ, QAT, distillation, Core ML*

- Compressing an on-device voice-activity-detection (VAD) model (CRDNN, iPhone Core ML target), diagnosed why textbook quantization barely shrank the 0.435 MB model: dynamic PTQ reaches only `nn.Linear` (1.2% of params), leaving the Conv2d + GRU stack (98.8%) in FP32, and QAT scoped to that 1.2% only cut F1
- Went beyond the course to implement static PTQ, a GRU→LSTM swap (making the recurrent weights quantizable), and knowledge distillation, compressing the model 0.435 MB → 0.185 MB (−57%) at F1 0.943 vs 0.959 baseline (E4). A distilled student reached 0.050 MB, 8.7× smaller
- Built a reproducible E0–E6 size/F1/latency benchmark with a train-split-disjoint calibration set and an FP32-LSTM control to attribute F1 shifts. Moved measurement off shared-CPU Colab to a local arm64 host (the iPhone Core ML `qnnpack` target) for stable latency

Experience

Quantitative Developer Intern — Jul. 2023 – Sep. 2023
Benefits Mutual Asset Management Co., Ltd. — *Beijing, China*

- Pulled daily Wind-terminal market data and wrote Python to compute technical indicators for the team's trading strategies
- Ran daily strategy backtests over historical price data and packaged the outputs for review
- Produced monthly strategy performance, P&L, and risk summary reports for the portfolio-management team