Yupo (Jason) Niu

AI Engineer | Calgary, AB, Canada | yupoca24@gmail.com | linkedin.com/in/yupo-niu | github.com/DISSIDIA-986 | portfolio.dissidia.tech | Chinese native · English fluent

Professional Summary

PGWP work-authorized — no employer sponsorship required. Permanent full-time roles in Canada.

AI Engineer with 17 years building production systems — 8 years as Team Lead at a publicly traded company designing decision automation (financial pricing engines, real-time risk scoring, approval workflow platforms) and now applying that same automation foundation to LLM and agentic systems: multi-agent orchestration, enterprise RAG, and LLM evaluation pipelines. Shipped a multi-provider LLM pipeline (5 providers with circuit breakers and comprehensive pytest coverage), an enterprise RAG platform with hybrid BM25 + pgvector retrieval over a multi-format document corpus, and an LLM-as-Judge evaluation pipeline (Qwen 3.5 + DeepSeek V3 on DeepEval, 3-dimension rubric). Graduated SAIT April 2026 — 3.96 GPA across Post-Diploma Certificates in Data Analytics (4.0) and Integrated AI (3.92), 15/16 courses A+.

Technical Skills

- **AI & LLM Systems: Agentic AI / Multi-Agent Orchestration, LangChain, LangGraph, RAG (pgvector, ChromaDB), Structured Output (Pydantic/instructor), LLM Reliability (circuit breakers, provider fallback chains), LLM-as-Judge / DeepEval, Prompt Engineering, MCP**
- **Machine Learning & Data: PyTorch, Hugging Face, scikit-learn, XGBoost, SHAP, YOLOv8, OpenCV, MediaPipe, Pandas, NumPy, Streamlit, SQLAlchemy, ETL Pipelines**
- **Production Stack: Python, FastAPI, Go, Java, Spring Boot, Spring Cloud, React, Next.js, TypeScript, .NET/C#**
- **Infrastructure: Docker, Kubernetes, GCP, Vercel, Alibaba Cloud, CI/CD, GitHub Actions, Prometheus, OpenTelemetry, Grafana, ELK**
- **Data Systems: PostgreSQL, pgvector, MySQL, Redis, MongoDB, Elasticsearch, ChromaDB, SQLite, InfluxDB**

AI Projects

JobPilot AI — Agentic Job Automation Pipeline (live demo, login required; code private)

- **Shipped autonomous multi-agent system orchestrating 5 LLM providers (GLM, Qwen, DeepSeek, GROQ, Ollama) with comprehensive pytest coverage, 4 discovery sources, and end-to-end auto-apply via headless Chrome**
- **Built LLM reliability stack with circuit breakers, stamina retries, Pydantic/instructor structured output, and per-task provider fallback chains — kept pipeline running through real provider outages and rate limits**

Industry-AI-Flow — Enterprise RAG Platform

- **Built LangChain-based intent recognition and multi-agent orchestration with hybrid retrieval (BM25 + pgvector) over a multi-format enterprise document corpus (PDF, DOCX, scanned), ingested via PaddleOCR — replaced manual lookup workflow with single-query answers**

AI Ops Control Room — LLM Quality Evaluation

- **Designed LLM-as-Judge evaluation pipeline (Qwen 3.5 simulates customer, DeepSeek V3 evaluates) using DeepEval with a 3-dimension rubric — replaced manual QA review of customer-service conversations with automated scoring**

Trading Bots — LLM-Driven Automated Trading

- **Built multi-tenant SaaS trading system: local LLMs (Qwen 2.5, DeepSeek-R1) generate entry/exit decisions from TradingView webhook signals, executed through IBKR API with per-user, per-symbol prompt configuration**

Fruit Ninja AI — Hand Gesture Game (Computer Vision)

- **Shipped gesture-controlled WebGL game using MediaPipe real-time hand tracking + Three.js 3D rendering with adaptive performance tuning; deployed to Alibaba Cloud ESA edge network**

Work Experience

AI / Full-Stack Developer (Part-time, ~20 hrs/wk) — Havenz Tech, Calgary | Aug 2025 – Apr 2026

- **Built an agentic AI / RAG workflow prototype on LangChain — intent recognition, multi-agent orchestration, pgvector hybrid retrieval, and PaddleOCR document ingestion — for automated document triage (validated in team testing)**
- **Owned the entire React Native mobile app (iOS + Android) for RISE, a Calgary sports-venue membership system — delivered to the client and published to the App Store and Google Play (React Native + TypeScript)**
- **Built CI/CD on GitHub Actions + Docker + GCP Cloud Run — cut deploy from manual 30+ min to automated 3 min per PR**

Senior Java Backend Developer / Team Lead — Edianyun Inc. (publicly traded), Beijing | Mar 2016 – Apr 2024

- **Built financial algorithm engine (NPV, ACPI, ROC) powering every rental-pricing and buyout decision — the decision engine behind the company's largest revenue stream**
- **Designed real-time risk control system with Alibaba DTS + Kafka streaming, multi-dimensional credit scoring models, and automated alerting — flagged high-risk applicants pre-approval to reduce default exposure**
- **Architected decision automation platform: 4-level approval workflow (M1→M2→M3→city manager), quota pool management, and 10+ configurable policy types — replaced manual approval chains with automated routing**
- **Led 3 teams (Risk Control, E-commerce, DevOps), built unified RBAC permission center and custom Dubbo RPC framework adopted by every internal system**
- **Migrated monolith → microservices → Kubernetes and stood up the observability stack (ELK + Prometheus + Skywalking APM) — sustained 10K+ concurrent users and reduced API response times**

Senior Java Developer — JiuLing Hou Credit, Beijing | Jun 2015 – Mar 2016

- **Built core backend for a consumer-finance credit platform; designed REST APIs and relational data models for the credit-application workflow (Java / Spring / MyBatis, MySQL, Redis)**

Full Stack Developer — Bitmain Technologies, Beijing | Jun 2014 – Jun 2015

- **Solo-shipped Bitcoin payment system and cut payment crediting/settlement latency ~40%**
- **Integrated multiple cryptocurrency exchange APIs to run automated arbitrage trading across venues**

Senior Java Developer — AsiaInfo Technologies, Beijing | Jul 2011 – Jun 2014

- **Built the PC web portal for Henan Mobile serving 10M+ regional telecom users; designed ETL data pipelines and BI reporting systems**

Java Developer — Zhengzhou GaoHong Softcom, Zhengzhou, China | Jul 2007 – Jul 2011

- **Developed and maintained Henan Mobile's mobile self-service web portal and related business modules (Java/J2EE)**

Education

Post-Diploma Certificate in Integrated Artificial Intelligence — SAIT, Calgary | Sep 2025 – Apr 2026

- 3.92/4.0 GPA | Computer Vision, Predictive Analytics, AI Governance, Human-Centred AI
- Capstone: Applied AI Projects

Post-Diploma Certificate in Data Analytics — SAIT, Calgary | Sep 2024 – Apr 2025

- 4.0/4.0 GPA | Statistical Analysis, Predictive Analytics, Business Intelligence
- Capstone: ThyroidSentry (97.4% cancer recurrence prediction with XGBoost + SHAP)

Diploma in Software Technology — Northwest University, Xi'an, China | Sep 2005 – Jul 2007