

Yupo (Jason) Niu

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

Professional Summary

AI Application Engineer with 17 years shipping production systems, including 8 years building enterprise decision engines, risk models, and workflow automation at a publicly traded company. Now building the same class of systems with LLMs, RAG, and agents. Built multi-agent pipelines orchestrating 5 LLM providers with circuit breakers and 1,250+ tests, enterprise RAG with hybrid retrieval, and LLM-as-Judge evaluation scoring 0.94/1.00. Graduated SAIT in April 2026 with 3.96 overall GPA (15/16 courses A+) across Post-Diploma Certificates in Data Analytics (4.0) and Integrated AI (3.92). Canadian work authorization (PGWP).

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:** scikit-learn, XGBoost, SHAP, YOLOv8, OpenCV, Pandas, NumPy, Streamlit, ETL Pipelines
- **Production Stack:** Python, FastAPI, Go, Java, Spring Boot, Spring Cloud, React, Next.js, TypeScript, .NET/C#
- **Infrastructure:** Docker, Kubernetes, GCP, 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

- Built autonomous multi-agent system orchestrating 5 LLM providers (GLM, Qwen, DeepSeek, GROQ, Ollama) with circuit breakers, structured output via Pydantic, and automated quality gates — 1,250+ automated tests, 4 discovery sources, auto-apply via real Chrome
- Implemented LLM reliability stack: stamina retries, instructor validation, aiobreaker circuit breakers, provider fallback chains with per-task routing

Industry-AI-Flow — Enterprise RAG Platform

- Built intelligent intent recognition and multi-agent orchestration using LangChain with hybrid retrieval (BM25 + pgvector), processing 25K+ lines of enterprise documents with OCR via PaddleOCR

AI Ops Control Room — LLM Quality Evaluation

- Designed LLM-as-Judge pipeline: Qwen 3.5 simulates customer service, DeepSeek V3 evaluates on 3 dimensions — achieving 0.94/1.00 composite score using DeepEval framework

Trading Bots — LLM-Driven Automated Trading

- Built multi-user SaaS trading system where local LLMs (Qwen 2.5, DeepSeek-R1) make entry/exit decisions via customizable prompts, with TradingView webhook signals executing through IBKR API with per-user/per-symbol configuration

Fruit Ninja AI — Hand Gesture Game (Computer Vision)

- Built gesture-controlled game using real-time AI hand tracking via MediaPipe with Three.js 3D rendering, fluorescent neon trails, and adaptive performance — deployed on Alibaba Cloud ESA

Work Experience

AI Developer — Havenz Tech, Calgary | Aug 2024 – Present

- Built agentic AI workflow platform with LangChain featuring intent recognition, multi-agent orchestration, pgvector hybrid retrieval, and OCR document processing
- Developed monorepo BMS: ASP.NET Core 8 backend, Next.js 15 admin panel, React Native mobile app with real-time SignalR communication
- Implemented CI/CD automation using GitHub Actions with Docker + GCP deployment

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

- Built core financial algorithm engine (NPV, ACPI, ROC) powering all rental pricing and buyout decisions — the decision engine behind the company's primary revenue stream
- Designed real-time risk control system with Alibaba DTS data pipelines + Kafka streaming, multi-dimensional credit scoring models, and automated risk alerting
- Architected decision automation platform with multi-level approval workflows (M1→M2→M3→city manager), quota pool management, and 10+ configurable policy types
- Led 3 development teams (Risk Control, E-commerce, DevOps), built unified RBAC permission center and custom Dubbo RPC framework serving all internal systems
- Migrated monolith→microservices→Kubernetes, built observability stack (ELK + Prometheus + Skywalking APM) — 10K+ concurrent users, significantly reduced API response times

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

- Independently shipped Bitcoin payment system and optimized blockchain confirmation time 40%
- Integrated with major exchange APIs (Huobi, Binance, OKEx) for automated arbitrage trading

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)