

Yupo (Jason) Niu

Senior Backend / AI Platform 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

Senior backend/platform engineer with 10+ years building production distributed systems in Java, with Python across data and ML work — 8 years as Team Lead at a publicly traded company building **decision-automation** platforms (financial pricing engines, real-time risk scoring, approval-workflow systems on Spring Cloud + Kafka, sustained heavy concurrent load). Recently specialized in **applied LLM / AI-platform** work: shipped JobPilot, a multi-provider LLM automation pipeline (5 providers with circuit breakers and comprehensive pytest coverage); built an enterprise RAG prototype (hybrid BM25 + pgvector, validated in team testing) and an LLM-as-Judge evaluation pipeline (DeepEval, 3-dimension rubric). Best fit: senior backend / AI-platform roles building reliable production automation around LLMs. Graduated SAIT April 2026 — 3.96 GPA (Data Analytics 4.0, Integrated AI 3.92). The combination this targets: bringing LLM capability into established enterprise systems, rather than greenfield AI products.

Calgary-based · available immediately.

Technical Skills

- **Languages & Core:** Python, Go, Java, TypeScript, Shell scripting, SQL, Groovy, Maven, Gradle, JUnit, HTML / CSS / JavaScript, pytest, pytest-asyncio, Playwright, ruff, Jupyter, Git
- **Backend & Data:** Pandas, NumPy, Streamlit, SQLAlchemy, ETL Pipelines, FastAPI, Spring Boot, Spring Cloud, React, Next.js, PostgreSQL, MySQL, Redis, MongoDB, Elasticsearch, SQLite, InfluxDB, Spring MVC, Spring Security, Dubbo RPC, Netty, MyBatis, JPA, XXL-JOB, Kafka, RocketMQ, RabbitMQ, Nacos service discovery, Alibaba DTS, Power BI, Tableau, Plotly, Matplotlib, Seaborn, React Native, Vue.js, Angular, Redux, Tailwind CSS, HTMX, ETL
- **AI / LLM:** 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, AI-Assisted ("Vibe") Coding — Claude Code, OpenAI Codex CLI, Cursor, GitHub Copilot, PyTorch / Deep Learning (*coursework level*), Hugging Face, scikit-learn, XGBoost, SHAP, YOLOv8, OpenCV, MediaPipe, MCP (Model Context Protocol), pgvector, ChromaDB
- **Cloud, Containers & IaC:** Linux (CentOS), Docker, Kubernetes (kubect!), Ansible, Terraform, GCP, Vercel, Alibaba Cloud (ACK), Nginx
- **CI/CD & Observability:** CI/CD (GitHub Actions, Jenkins), Prometheus, OpenTelemetry, Grafana, ELK, JVM performance tuning, SkyWalking APM, ELK (Elasticsearch, Logstash, Kibana), SLS, ARMS APM, SonarQube, CI/CD
- **Architecture, Domain & Delivery:** Microservices Architecture, Distributed Transactions & Sagas, High Concurrency, Event-Driven & Real-Time Systems, API Gateway, RBAC, OAuth, Financial Algorithms, Risk Control, E-commerce Systems, Blockchain

AI Projects

JobPilot AI — Agentic Job Automation Pipeline

- 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 (prototype) (*open source*)

- 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 — single-query retrieval over the corpus (prototype, validated in team testing)

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 — automated scoring of customer-service conversations (personal evaluation pipeline)

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

go-mcp-bridge — MCP Bridge for LLM Agents

- Wrapped a Go ticketing backend with the **Model Context Protocol** so LLM agents (Claude Code, Codex, Gemini CLI) book, pay, refund, and validate demo tickets through MCP tools, resources, and prompts behind a human consent + scope gate — scoped expiring tokens, identity derived from the token rather than tool args, and concurrency-safe booking that prevents double-sell/double-charge

Work Experience

Havenz Tech

Aug 2025 – Apr 2026

AI / Full-Stack Developer

Calgary

- Built an internal 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

Edianyuan Inc. (publicly traded)

Mar 2016 – Apr 2024

Senior Java Backend Developer / Team Lead

Beijing

- 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
- Ran the field-service dispatch cost study in **Python**: clustered on-site work-order coordinates from Hive into engineer service territories (K-Means adapted to a great-circle assignment metric), then designed and executed five-day replays of real work orders against the production dispatch engine across engineer-headcount and start-position scenarios — Beijing at 25 engineers instead of 10 raised system auto-dispatch from 50.1% to 89.8% and cut modelled monthly cost 13.2%, a headcount effect rather than an algorithmic one
- Built **Hyperion Data Collect**, the multi-system financial-compliance ETL for HKEX IPO audit — six upstream systems reconciled (Oracle EBS, Hyperion HFM, U8C, order-billing, risk-control credit, ERP asset core), Fork/Join extraction of millions to tens of millions of rows staged in Redis, RocketMQ choreography with an Oracle state machine and request_id idempotency — cross-system consistency without a distributed lock
- 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
- Led DevOps platform build: migrated monolith → Spring Cloud microservices → Kubernetes on Alibaba Cloud ACK with containerization of the stateless services and dev/test/stage environment isolation — authored Jenkinsfile/Dockerfile templates and Jenkins CI/CD pipelines, used Ansible for provisioning, and drove kubectrl adoption across dev teams
- Stood up the observability stack (ELK + Prometheus + Skywalking APM) with self-built Kibana dashboards and log-based alerting; performed JVM performance tuning under sustained concurrent load — reduced API response times
- Owned hands-on server operations from the early stage — administered several CentOS servers with full root access and near-daily ops, personally building and deploying the dev, test, and production environments before the DevOps team scaled up

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, XGBoost, SHAP

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