Richard Zhu

✉: Richardzhu09 | **in**: Richard Zhu | : Richard Zhu | Markham, ON

Experience

Software Developer Markham, ON
WellCare Insurance *September 2025 - December 2025*

- Architected a fault-tolerant automation pipeline for legacy Guidewire systems using **Playwright**, utilizing **strict input normalization** to ensure deterministic data entry across complex, conditional workflows.
- Engineered a **Human-in-the-Loop (HITL)** verification system, routing automated data entry to a dashboard for manual review to guarantee **99.9% data integrity**.
- **Developed an LLM-powered parsing engine** via Anthropic's Claude and pdfplumber, standardizing unstructured layouts and reducing ingestion latency by 90% (<3s/document).

Software Developer Toronto, ON
Leading Aces Academy *January 2025 - April 2025*

- **Scaled authentication infrastructure** to a self-hosted **Keycloak (OIDC)** cluster, implementing **custom SPI providers** for schema validation across distributed Django microservices.
- **Engineered a distributed background worker system** using **Celery** and **Redis** to offload video processing tasks, ensuring system responsiveness during large batch uploads.
- **Developed a real-time WebSocket bridge** between Django and Godot, utilizing a custom messaging protocol to support low-latency data synchronization.

Projects

WAT.ai x Bindwell EPA Consulting Agent (YC S25) | *Python, FastAPI, React, OpenAI*

- **Engineered a FastAPI streaming protocol** to multiplex "source" and "answer" data types, using TextDecoder for real-time response rendering and citation mapping in React.
- **Developed a custom PDF viewer** with fuzzy-matched deep linking, allowing users to jump from AI-generated citations directly to highlighted source text within regulatory documents.
- **Integrated a GPT-4o "Judge Agent"** into the full-stack workflow, routing autonomous self-correction states to the UI to ensure 100% faithfulness across a 7-day build.

Exam Generation Engine | *LangGraph, FastAPI, Pinecone, Google Gen AI, OpenAI*

- **Architected a multi-agent generation** pipeline using LangGraph, utilizing a Map-Reduce pattern to parallelize the creation of problems, solutions, and grading rubrics.
- **Implemented an agentic RAG workflow** with Pinecone dense indices, leveraging Gemini for structural document parsing and LaTeX conversion of complex academic materials.
- **Engineered a real-time Markdown/LaTeX** streaming layer via FastAPI, routing agent outputs to a React/Vite interface for dynamic document assembly.

Advanced RAG Knowledge Engine | *Python, Flask, LangChain, ChromaDB, LLama 3, AWS, Groq*

- **Architected a two-stage retrieval pipeline** using ChromaDB and a Cross-Encoder reranker, implementing **custom chunking** to preserve tabular context and metadata for high-precision search.
- **Optimized inference latency** by self-hosting **Qwen embeddings via Transformers** locally, offloading reasoning to Llama-3-70b via Groq to minimize Time-To-First-Token (TTFT).
- **Engineered a production-grade backend** on AWS EC2, configuring Nginx as a reverse proxy to handle SSL termination and load balancing for the asynchronous **Gunicorn/Flask** application server.

Technical Skills

Languages: Python, TypeScript, JavaScript, C++, Java, Bash
AI & Data: LangChain, Hugging Face (Transformers), LangGraph, OpenAI API, Gemini, ChromaDB, RAG, pdfplumber
Backend: FastAPI, Django, Gunicorn, Nginx, Redis, Celery, Docker, AWS (EC2), REST APIs
Frontend: React, Vite, Tailwind CSS, Playwright (E2E), Shadcn/UI

Education

University of Waterloo Waterloo, ON
Honours Mathematics, Combinatorics and Optimization, Co-op | *3.7 GPA* *September 2022 – April 2027*
President's Scholarship | 93% admission average *May 2022*