Richard Zhu

: Richardzhu09 | : Richard Zhu | : Richard Zhu | Markham, ON

Experience

WAT.ai x Bindwell EPA Consulting Agent (YC S25) Waterloo, ON

Software Engineer *January 2025 - Present*

- **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.

WellCare Insurance Markham, ON

Software Engineer *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.
- **Engineered an algorithmic ingestion engine** for Ontario insurance forms via pdfplumber, achieving 100% data integrity and a 90% latency reduction (<1s/document) over LLM-based extraction methods.

Leading Aces Academy Toronto, ON

Software Developer *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 an automated forum-crawling system** for data extraction, utilizing recursive link traversal and keyword-density analysis to identify and categorize high-intent lead signals.
- **Developed a real-time WebSocket bridge** between Django and Godot, utilizing a custom messaging protocol to support low-latency data synchronization.

Projects

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.

Lumen Parser | *Python, ChromaDB, Transformers, PyMuPDF, Qwen, Jina*

- **Engineered a hierarchical retrieval pipeline** using markdown-header parsing to minimize noise via child-leaf retrieval while maximizing context through parent-node expansion.
- **Implemented Late Chunking (Jina v3)** via a custom ChromaDB embedding function to preserve global context and long-range dependencies across token boundaries before vector pooling.
- **Implemented hybrid search** via Reciprocal Rank Fusion (RRF) of semantic and BM25 retrieval, ensuring high-precision keyword retention and conceptual relevance.

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, Flask, Redis, Docker, AWS, GCP, REST APIs
Frontend: React, Vite, Tailwind CSS, Playwright (E2E), Shadcn/UI

Education

University of Waterloo Waterloo, ON

Honours Mathematics, Combinatorics and Optimization, Co-op *September 2022 – April 2027*

President's Scholarship | 93% admission average *May 2022*