

Aiden (Chin Wei) Mak

AI Engineer — LLM Agents · RAG · Evaluation & Observability · Python/FastAPI

Toronto, ON | (204) 962-2406 | mcwaiden000@gmail.com | [linkedin.com/in/mcwaiden](https://www.linkedin.com/in/mcwaiden) | github.com/aidenmak0624 | aidenmak.com

TECHNICAL SKILLS

AI Engineering: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), agentic AI & multi-agent orchestration, LangChain / LangGraph, tool & function calling, Model Context Protocol (MCP), prompt engineering

Evaluation & LLMops: promptfoo evals, LangSmith tracing, guardrails & hallucination suppression, prompt/agent regression testing

Languages & Backend: Python, TypeScript, JavaScript, SQL, FastAPI, Node.js, REST APIs

Vector & Data: Pinecone, ChromaDB, PostgreSQL / pgvector, Redis, SQLite, embeddings, chunking & retrieval tuning

Deployment: Docker, GCP Cloud Run, Supabase Edge Functions, CI/CD (GitHub Actions), React / Next.js

EXPERIENCE

Zokforce – AI Engineering Intern (AI Consultancy)

Jan 2026 – Present

Volunteer → Intern (since May 2026) · Toronto, ON

- Built a **Python** clinical-dialogue audio generation tool converting de-identified clinical notes and case briefs into labeled multi-speaker audio to test the **STT/ASR** pipeline of a **government-funded** medical-scribe project
- Researched clinical datasets (ACI-Bench, MTS-Dialog) and commercial **TTS engines** (Azure Speech, MiniMax, Google, ElevenLabs), **advising leadership** on the STT testing direction for Cantonese/English clinical audio
- Partnered with senior engineers on **issue triage and release stability**, reproducing bugs, verifying fixes, and contributing to **QA planning** and release-readiness decisions
- Ran **acceptance testing** for an LLM-observability platform, logging defects with reproduction steps and screenshots

Independent AI Engineer – Self-Directed Projects

Jun 2025 – Present

Toronto, ON · Remote

- Shipped **4+ AI applications** to live production URLs, owning scoping, build, deployment, and evaluation

Canadian Museum for Human Rights – Applied AI Developer

Sep 2024 – Apr 2025

University of Manitoba Partnership

- Built an NLP-driven human-rights education app with the **Gemini LLM API**, delivering **6 interactive learning modules**
- Gathered requirements from museum stakeholders in an **Agile** team, shipping iterative releases on an academic-year timeline

PROJECTS

HR Intelligence Platform – AI Engineer

[GitHub](#) | [Live Demo](#) | Jan–Feb 2026

Python, LangGraph, MCP / FastMCP, ChromaDB (RAG), GCP Cloud Run, Prometheus / Grafana

- Cut manual HR policy lookups to a single question by routing them through **8 LLM agents** and a citation-grounded **RAG** assistant
- Exposed **37 tools** to any MCP client via a **FastMCP server** (8 resources, 5 prompts) over stdio and HTTP transports
- Protected 8 agent workflows from regressions with **1,909 unit, integration, and workflow tests** gating every merge in CI
- Kept the deployed service debuggable with **GCP Cloud Run** CI/CD and Prometheus, Grafana, and LangSmith tracing

DecisionEase – AI Engineer

[GitHub](#) | [Live Demo](#) | Apr–May 2026

Next.js, React, TypeScript, FastAPI, Multi-Agent (A2A), pgvector

- Designed a **6-agent system** turning messy user context into ranked next-best-action recommendations via **A2A** coordination
- Kept agent runs bounded with a **4-layer memory** model, per-agent token budgets, and 5000 ms timeouts
- Made semantic recall fast with **1536-dim pgvector embeddings** and query-optimized indexes
- Shipped the full stack live with a React/Next.js frontend and **Python/FastAPI** backend across **27 REST routes**

ReviewLensAI – Creator / AI Engineer

[GitHub](#) | [Live Demo](#) | Mar 2026

React, Supabase Edge Functions, Pinecone, OpenAI GPT-4o, RAG, promptfoo

- Suppressed ungrounded AI claims by requiring citations, verified by a **promptfoo eval suite (6/6)** and **181/181 Vitest tests**
- Turned raw customer reviews into ranked complaint themes via a **3-worker agentic pipeline** over Pinecone (cosine, top-K=8)
- Let non-technical users start from whatever they have by accepting **4 ingestion modes** (CSV, paste, URL, image) into one **RAG** index

The Golden Fork – Software Engineer

[GitHub](#) | [Live Demo](#) | Feb 2026

Next.js, OpenAI, Pinecone, WebSocket, Stripe, Radix UI

- Built a **RAG** menu assistant with **OpenAI embeddings** and **Pinecone** that answers dietary questions from real menu data and turns browsing into orders through structured add-to-cart suggestions
- Made **Stripe** webhooks the authoritative order trigger, creating kitchen tickets only after server-side payment confirmation and pushing live order status over **WebSocket** so staff never refresh mid-service
- Served customer, kitchen, and manager surfaces from one **REST API** backend of 10+ routes

EDUCATION

University of Manitoba

Winnipeg, MB

Bachelor of Science, Honours Computer Science – GPA: 3.5

Sep 2020 – Jun 2025

Relevant Coursework: Artificial Intelligence, Software Engineering, Databases, Data Structures & Algorithms, Machine Learning