Marco Chen

+1 (548) 577-5684 | z253chen@uwaterloo.ca | LinkedIn | Website

Education

University of Waterloo Sep. 2023 – Aug. 2028
Bachelor of Computer Science, Co-op **CGPA:** 3.7/4.0

- **Coursework:** OOP, DSA, Computer Organization, Compilers, Graph Theory, Probability, Statistics, Linear Algebra

Experience

Software Engineer Intern May 2026 – Aug. 2026
Geotab | TypeScript, React, PostgreSQL, Google BigQuery, C#, .NET Kitchener, ON, Canada

- Building a Driver Safety Coaching module to enable sessions that aggregate harsh driving events and bridge real-time communication with fleet managers, improving commercial fleet safety
- Implemented internal utility APIs that communicate with the MyGeotab backend and PostgreSQL database

Game Engine Development Apprentice (AIGC Cohort) Jan. 2026 – May 2026
Tencent Games Remote, Shenzhen, China

- Architected a **RAG-based** persistent memory system for autonomous NPCs using **PostgreSQL/pgvector**, enabling agents to store and retrieve past interactions with **cosine similarity search** to maintain character consistency

Software Engineer Intern Sep. 2025 – Dec. 2025
Geotab | C#, TypeScript, React, .NET, Jest, xUnit Kitchener, ON, Canada

- Automated the Government of Canada's compliance ELD test procedures into an **end-to-end xUnit test suite**, eliminating manual testing entirely and reducing test execution time by **98%** (from 2 weeks to 4 hours)
- Optimized frontend performance for **5M+** users by migrating legacy tech stack to modern **React**, rewrote app launching logic to reduce API calls by 46%
- Investigated and debugged backward compatibility bugs, improved UI consistency across different OS platforms

Software Engineer Intern Jan. 2025 – April 2025
Octopodi Technologies | Typescript, React, Next.js, Tauri, Tailwind CSS, Jest Waterloo, ON, Canada

- Built the front-end of a cross-platform desktop application using **React** and delivered a highly reusable component library from scratch, followed **WCAG** standards and implemented **i18n** to create highly accessible UI/UX
- Developed and automated 200+ unit tests using **Jest** and achieved over **90% code coverage** in all components

Projects

LLM From Scratch | Python, PyTorch, CUDA, WandB, Multiprocessing, einops

- Implemented a BPE tokenizer from scratch using multiprocessing and a max-heap with lazy stale deletion, reducing merge complexity to **O(n log m)** and achieving a processing throughput of **286k tokens/sec**
- Built a full **22M-parameter Transformer** LM with custom Linear layers, RMSNorm, SwiGLU gated FFN, scaled dot-product multi-head **Attention** with causal masking and RoPE on Q/K, achieving a **validation perplexity of 2.20** and **validation loss of 0.78** on the TinyStories dataset
- Developed a training pipeline on an H100 GPU using BF16 mixed-precision and zero-copy strided batching, sustaining **413k tokens/sec throughput** with a custom AdamW optimizer, cosine annealing, and gradient clipping

LLM Gateway | Go, PostgreSQL/pgvector, Redis, Kafka, ClickHouse, Prometheus/Grafana, Docker, Kubernetes

- Architected a multi-tenant LLM proxy in **Go**, unifying OpenAI, Anthropic, and Gemini APIs, featuring a smart router with automated fallbacks that **reduced user-facing errors by 87%** during outages
- Engineered a two-tier semantic cache on pgvector + IVFFlat with tiktoken-based budget enforcement, achieving a **42% hit ratio**, reducing API costs by **38%**, and dropping tail latency from **1.4s to 47ms**
- Built custom net/http adapters and Redis sliding-window rate limiters for full streaming control, sustaining **1,200+ RPS** with **sub-8ms P99 overhead** and **<2ms check latency** across 10K concurrent keys

Technical Skills

Languages: Python, Go, TypeScript, C#, C++, Kotlin, Java, HTML, CSS, SQL
Frameworks: chi, FastAPI, LangGraph, Next.js, React, Tailwind CSS, Jest, Django, Node.js, Flask
Tools & Platforms: Git, Linux, ClickHouse, Prometheus/Grafana, Docker, Kubernetes
Databases/Cloud: GCP, AWS, S3, Lambda, ChromaDB, PostgreSQL, pgvector, MySQL, MongoDB, Redis, Kafka
Libraries: PyTorch, WandB, Keras, OpenCV, NumPy, Matplotlib, pandas, PyAudio