

BOB ZHENG

☎ 647-869-2600 ✉ b63zheng@uwaterloo.ca 🌐 bobzheng.com 📧 shenyan-zheng 📄 BobZheng00

Technical Skills

Languages: Python, JavaScript/TypeScript, SQL, C/C++, Kotlin, Bash Script
Frameworks/Libraries: FastAPI, Vite, Next.js, React, Express.js, Flask, Django, Node.js, Streamlit
Data and ML Technologies: PyTorch, Transformers, ADK, LangGraph, Huggingface, Pandas, NumPy, scikit-learn
Tools: AWS, Azure, GCP, Linux, Git, Docker, Kubernetes, PostgreSQL, MongoDB, Jenkins

Work Experience

Software Engineer

Sept 2026 – Dec 2026

Carta | Data Ecosystem

Waterloo, CA

- Improving data quality for a **Snowflake Cortex** text-to-SQL agent served over **MCP**: profiling client queries, enriching semantic views with metrics, dimensions, and verified questions, and building benchmarks and regression tests.
- Streamlined API docs **CI/CD** with **CircleCI** validation, automated upstream PRs, and **Jenkins** auto-deploy on merge.

Undergraduate Research Assistant (with Prof. Edith Law)

May 2026 – Present

University of Waterloo | HCI Lab

Waterloo, CA

- Investigating **student-directed multi-agent orchestration** for learning: shipped a platform where a student coordinates a persistent study companion and per-subject tutor agents against self-defined goals.
- Designed the **conversation-history** layer keeping long sessions in-context: rolling **compaction** folds old turns into a running summary behind an advancing watermark, over a scoped memory store the agent mutates via remember/recall/update tools.
- Made the **LLM** an actor over the domain via typed **tool-calling**: structured tools for quest-board writes, memory, workspace files with **PDF** extraction, and assessment logging, so agent turns persist real state rather than text.

Software Developer

Jan 2026 – April 2026

TD | Enterprise Innovation

Toronto, CA

- Built an **Accessibility Scanner** browser extension to automate enterprise-wide web accessibility and compliance checks across internal web properties, using **Vite** and **Express.js** in **TypeScript**.
- Architected a **PostgreSQL** database with optimized indexing for efficient scan result retrieval and eligible domain lookup; maintained an enterprise **CI/CD** pipeline in **Jenkins** with **Vitest** unit tests, coverage thresholds, and security scans.
- Studied **ROS** and experimented with a Humanoid Service Robot's 2D **LiDAR** and sonar sensors, tuning costmaps and path planning to prototype in-person service scenarios such as customer greeting and appointment check-in.
- Contributed to a PoC multi-agent design system using **Google ADK** and **Anthropic** models, designing autonomous agentic flows where **ADK LLM agents** coordinate via **A2A** and **function calling** to delegate sub-tasks across specialized agents.
- Built collaboration features on **GCS** and **Firestore**, and enabled bidirectional sync between code prototypes and **Figma** via **MCP** and **REST API**, with custom parsers that compressed Figma context by 50% for efficient LLM consumption.

NLP Research Assistant (with Prof. Usman Naseem)

Sep 2025 – August 2026

Macquarie University

Remote

- First-authored **VISPA** (<https://arxiv.org/abs/2601.12758>), a training-free pluralistic alignment framework accepted to **EMNLP 2026 Main Conference**, replacing costly multi-model ensembles with modular value-steered generation.
- Engineered an inference system using DeBERTa-based **NLI** classifiers to rank human values, executing inference-time **latent space manipulation** through **PyTorch** by injecting value directions into intermediate LLM activations.

AI/ML Engineer

Sept 2024 - May 2025

Eon Media Corp.

Toronto, CA

- Built a **Flask** backend for pipeline health monitoring and MLOps: dataset management and automated training/evaluation.
- Shipped transcript chunking with **all-MiniLM-L12**; fine-tuned **Mistral-7B** to summarize media into searchable segments.
- Cut **AWS EKS** pod creation and runtime cost via container right-sizing, algorithmic optimizations, and efficient **AWS API** usage; improved scalability with **Docker**, **Kubernetes**, and **AWS ECR/EKS**.
- Developed multi-player tracking for sports video using **Deep SORT** with a custom **ImageNet**-based feature extractor; integrated team/jersey-number recognition and key-moment detection.

Projects

ORIS: interactive AI tutoring platform | LiveKit, Anthropic, Next.js, React

Jan 2025 - May 2026

- Architected an agentic voice tutor on **LiveKit Agents** with custom tool-calling for material navigation, pacing, and Q&A handoff, letting the agent drive monologue flow autonomously without constant user input.
- Designed a sequence-ordered state-sync protocol over **LiveKit** data channels for real-time slide and annotation updates across browser, **FastAPI** orchestrator, and agent worker, with request/ack confirmation on agent-initiated actions.
- Shipped an automatic course indexing pipeline: ingest study materials, chunk and embed, index in **FAISS** for **RAG**; generated course outlines and slides via **Claude (Anthropic API)** with structured tool use.

Education

University of Waterloo

Sep 2023 - Apr 2028

Bachelor of Computer Science | CGPA: 93.66

Waterloo, CA

- Certifications:** UR2PhD Undergraduate Research Training Course; AWS Certified AI Practitioner