

Shukun Zhang

Toronto, ON | +1 437-869-1299 | s968zhan@uwaterloo.ca | [linkedin.com/in/shukun-zhang](https://www.linkedin.com/in/shukun-zhang)

SUMMARY

Software / AI engineer and Waterloo MDSAI student who ships finished products—enterprise Agentic RAG and a production multi-tool AI agent platform. Hands-on building AI agents that call tools via Model Context Protocol (MCP), reason with ReAct/AutoGen, and harden real-world agent behavior with evaluation and fail-closed execution. Seeking an AI engineering internship / co-op to design, build, and ship agentic AI platform capabilities from early prototypes to production.

EDUCATION

University of Waterloo - Master of Data Science and Artificial Intelligence (Co-op) 2026 Sept - Dec 2027 (Expected)
Recognition: Vector Scholarship in AI Recipient
Tufts University - 5-Year Combined BS/MS: B.S. Mathematics & Computer Science; M.S. Data Science 2020 Sept - May 2021
GPA: 3.85/4.00; **Recognition:** Magna Cum Laude; Dean's List; Tufts Alumni Scholarship.

TECHNICAL SKILLS

Programming & Backend: Python, C++, TypeScript, SQL, FastAPI, PostgreSQL, Redis
ML & Deep Learning: PyTorch, CUDA, TensorFlow, Scikit-learn, OpenCV, NumPy, Pandas, MindSpore
LLM & Agent Systems: Transformers, RAG, AI Agents, LangChain, LangGraph, AutoGen, ReAct, Qwen, DeepSeek
Retrieval & Infrastructure: FAISS, BM25, RRF, CrossEncoder, BGE-M3, Knowledge Graphs, Celery, Azure, Docker

PROFESSIONAL EXPERIENCE

Talkweb Information System Co., Ltd Hangzhou, China
AI Algorithm Engineer Mar 2025 - Nov 2025

- **Built a lightweight C++ on-device AI Agent** framework extending AutoGen with OpenAI APIs/MCP; developed an **AI Agent Security Posture Management system** with fine-tuned DeBERTa and rule/LLM MCP analysis.
- Developed a cloud-edge AI model protection framework with **AES-GCM** and digital envelopes; integrated **llama.cpp** and **ONNX Runtime** with OpenSSL-based edge reconstruction.

MatriQ (Formerly KT-Quantum Technology) Hangzhou, China
Software Engineer / AI Engineer Jan 2024 - Jan 2025

- Contributed to growth from **Angel to A+ funding (1B+)** within one year by developing automated AI and quantum-control infrastructure.
- Achieved **99.85% F1** with a PyTorch/CUDA atom-state classifier trained on **14,751 images** via statistical labeling; 99.73% precision and **0.55 ms inference/50 images**.
- Reduced waveform-generation latency **13x (200 ms to 15 ms)** by profiling bottlenecks and re-engineering a Python pipeline in C++, optimizing memory, buffers, and synchronization.
- Built a **hybrid quantum-AI control backend** integrating ML inference, data acquisition, digital I/O, and rearrangement algorithms via Labscript/Qt, supporting **45x45 arrays / 1,000+ error-free atoms**.

Huawei Technologies Co., Ltd. (Huawei 2012 Labs) Shenzhen, China
Software Development Engineer Nov 2021 - Nov 2023

- Developed **AI model security** mechanisms in MindSpore, including encryption/decryption, adversarial defense, and dynamic model obfuscation; filed a related patent.
- **Supported MindSpore CC EAL2+ certification** through formal security artifacts and compliance modules; **led the Trusted AI SIG** and drove secure-AI features across framework components.

PROJECTS

KnowledgeHub AI | [View Project](#) Toronto, Canada
Founder & ML Engineer Nov 2025 - Present

- Built and shipped an **enterprise Agentic RAG platform** for technical/SWIFT documentation, combining query rewriting, routing, reranking, evaluation, and web fallback for cited answers.
- Improved retrieval with **BGE-M3 + BM25/RRF + CrossEncoder** and graph RAG; built multilingual PDF/DOCX/JSON ingestion, multi-KB FAISS indexes, and FastAPI/WebSocket serving.

Cinegent AI | [View Project](#) Toronto, Canada
Founder & ML Engineer Jan 2026 - Present

- Built and deployed a **production AI video agent** automating commercial video ads from script → storyboard → image → video using LangChain ReAct and **16 tools**.
- Developed Qwen/Wan **multimodal generation and agent-memory pipelines**; hardened long-running workflows with Celery, SSE, retries, and context/tool-loop safeguards.

PUBLICATIONS

Model Processing Method, Model-based Data Processing Method, and Related Apparatus | [Patent](#) | Co-Second Author, US20250245302A1 (Jul 2025) | [Link](#)
Ultra-metric Path Distances for Hyperspectral Image Clustering | [Paper](#) | First Author, Remote Sensing 13(5), 955 (2021) | [Link](#)