

Shahab Jami

shahab@jamitech.dev | linkedin.com/in/shahab-jami | jamitech.dev

EXPERIENCE

Lead AI Systems Engineer

Jan 2023 – Present

JamiTech

Toronto, ON

- Architected & built ingestion pipeline benchmarked at **131k+ events/second** with **99.9% uptime** powering a **PIPEDA-Compliant** subscription platform for **650+ businesses**
- Reduced inference costs by $\sim 20\times$** (vs GPT-4o API) with **0 measurable accuracy loss**, by designing and developing an intelligent model routing system for multi-model execution - in collaboration with **Vector Institute**
- Developed data platform to transform **\$2.3B in previously unusable procurement data** into clean data with **100% lineage/provenance** using event-driven multi-agent architecture, exposing a **55% data quality gap**
- Maintained and extended a high-availability university web platform serving **53k+ users** with near-zero downtime
- Designed **HIPAA-compliant system architecture** for client applications under healthcare data regulations

Founder & AI Engineer

Jan 2025 – Jan 2026

Youralgo

Toronto, ON

- Engineered a **C++/Go vector quantization and scoring system** processing **$\sim 2,400$ multi-dimensional candidates/request**, with **sub-20ms p50 latency** handling real-time vector drift without execution stalls
- Reduced memory requirements/hardware node by 40%** and dropped execution costs/unit by **$\sim 89\%$ (\$0.55 to \$0.06)** via hardware optimization and model quantization pipeline (INT8/FP8) for constrained compute
- Architected, stress-tested, and deployed a bare-metal scaling strategy using capacity modeling to sustain throughput tests matching **10M DAUs** footprint constraints
- Reduced redundant execution overhead by $\sim 40\%$** by discarding generic framework abstractions (LangGraph) in favor of a specialized, execution state machine for orchestration

Course Author (Systems/C-Programming)

Oct 2024 – March 2025

Udacity

Toronto, ON (remote)

- Authored a technical curriculum specializing in production-level **C programming**, focusing on low-level systems execution, manual memory allocation, pointer mathematics, hardware profiling, and performance optimization
- Delivered engineering content adopted immediately by **1,000+ technical learners**, followed by **sustained engagement across subsequent quarters** establishing long-term resource-constrained systems best practices

Software Engineer Intern

May 2022 – Aug 2022

TD Bank (TD Securities)

Toronto, ON

- Developed and maintained low-latency microservices with event-driven architecture for real-time records
- Maintained full stack execution continuity during critical team staffing gap, independently

PROJECTS

Interpretability in LLMs/Agents | *Vector Institute*

April 2026 – May 2026

- Implemented end-to-end multi-agent tracing framework generating complete lineage across inference phases yielding **81% per-agent accuracy and 100% final decision reliability** to eliminate black-box algorithmic opacity

EDUCATION

York University

Toronto, ON

Honours Bachelor of Science (BSc) in Computer Science: focus in Machine Learning and Embedded Systems

TECHNICAL SKILLS

Languages: C, C++, Go, Python, TypeScript, JavaScript, Java, SQL, Assembly

AI/ML: TensorRT, Quantization, PyTorch, TensorFlow, LangGraph, CrewAI, Scikit-Learn, LLMs, RAG, Vector DBs

Data Technologies: MySQL, PostgreSQL, pgvector, Redis, RedPanda, ScyllaDB, Pandas, ETL Pipelines

Backend Technologies: Django, FastAPI, Spring-Boot, Axon, Node.js, GraphQL

Frontend Technologies: React, React-Native, Next.js, Vite

Tools & Platforms: Docker, Kubernetes, AWS, GCP, CloudFlare, MLflow, Git