Shahab Jami

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

Experience

Lead AI Systems Engineer Jan 2023 – Present
JamiTech *Toronto, ON*

- Led architecture for large-scale data platform to automate transforming **$2.3B in previously unusable financial procurement data**, with event-driven multi-agent system into clean data with **100% lineage & provenance**, uncovering a **55% disclosure gap**
- **Reduced inference costs by ∼20×** (vs GPT-4o API) as technical lead by designing an intelligent model routing system for LLM-agent orchestration with **0 measurable accuracy loss**, in collaboration with **Vector Institute**
- Designed architecture and built a **PIPEDA-Compliant** subscription platform and data pipeline benchmarked to sustain **131k+ events/second** to improve logistics for **650+ businesses** reliably with **99.9% uptime**
- 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*

- Measured sustained performance at **10M DAU throughput** for AI content generation platform, Architected and built end-to-end, through capacity modelling and stress testing of bare-metal scaling strategy
- Engineered VRAM hot-swap pipeline **Reducing costs from $0.55 to $0.06 (∼89%) per video and memory requirements per node by 40%** with GPU optimization, quantization, and refactored inference orchestration
- **Reduced token usage by ∼40%** by leading a transition from LangGraph-based pipelines with a custom multi-step agent orchestration architecture
- **Achieved sub-20ms p50 latency** for a **real-time** recommendation system, evaluating **∼2,400 candidates per request** and handling vector drift by engineering a custom C++/Go vector quantization system

Course Author Oct 2024 – March 2025
Udacity *Toronto, ON (remote)*

- Collaborated with cross-functional teams to author a production-focused C programming course covering low-level systems concepts including memory management, performance optimization, and systems-level programming
- **Adopted by 1,000+ learners** immediately, followed by **sustained engagement across subsequent quarters**

Software Engineer Intern May 2022 – Aug 2022
TD Bank (TD Securities) *Toronto, ON*

- Contributed to full-stack development of internal systems using Spring Boot microservices and React (TS & JS)
- Worked within an event-driven architecture leveraging Axon Framework for distributed system coordination
- Maintained development continuity during a temporary staffing gap, independently progressing system work and priorities

Education

York University Toronto, ON
Honours Bachelor of Science (BSc) in Computer Science with a focus in Machine Learning

Technical Skills

Languages: Python, TypeScript, JavaScript, Java, C, C++, Go, SQL
AI & ML: PyTorch, TensorFlow, TensorRT, LangGraph, CrewAI, Scikit-Learn, LLMs, RAG, Vector Databases
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