First-author

Toronto, CA email@example.com +1 000 000 0000 linkedin.com/in/username

Education

University of Toronto, Bachelor of Applied Science 09/2021 - 05/2026

Machine Intelligence, cGPA: **3.94**/4.00

– Relevant Courses: Data Structures & Algorithms, Probability and Statistics, Reinforcement Learning, Digital and Computer Systems, Introduction to Machine Learning, Matrix Algebra and Optimization, Probabilistic Reasoning, Software and Neural Networks, Introduction to Databases.

Publications

– **Retrieval-Augmented Conversational Recommendation with Prompt-Based Semi-Structured Natural Language State Tracking**
First-author, *ACM SIGIR*, 2024.

Experience

SWE/SDE/MLE/MLops Intern, Modiface – Toronto, CA May 2024 - Present

– Engineered a robust data processing pipeline to augment a synthetic 3D-face dataset (25M images) using stable diffusion with Python and Shell scripting, achieving a 17

– Developed a seamless model conversion tool to translate in-house computer vision models from PyTorch to TensorFlow, ensuring consistent cross-platform integration, reducing latency by over 10

– Optimized inference performance of generative AI models (including GANs and diffusion-based models) by integrating OpenVINO, reducing latency by over 40

– Finetuned Stable Diffusion LoRA models to improve task-specific GenAI performance, improving accuracy by 40

– Spearheaded the redesign and development of an internal computer vision model profiling tool's metric panel using TypeScript. This dashboard provided real-time monitoring of key performance indicators such as FPS, memory usage, and track time, enhancing system observability and performance tuning.

SWE/SDE/MLE/Applied ML/ML Research intern, Data-Driven Decision Making Lab (UofT) – Toronto, CA April 2023 - September 2023

– Designed and implemented a RAG-based chatbot system leveraging LLM APIs for dynamic natural language interactions and personalized responses.

– Developed core infrastructure in Python with 5+ API integrations, implementing rate limiting and caching mechanisms to handle 500+ RPM.

– Built a custom data pipeline using FAISS vector database to process and embed 1M+ product entries for real-time retrieval.

– Retrieval-Augmented Conversational Recommendation System (SIGIR 2024, 1st Author): Developed an open source RAG-based conversational recommendation system powered by GPT-4.

– Implemented two-tower dense retrieval with FAISS indexing 5M+ item embeddings for real-time similarity search.

Software Engineer Intern, Voith Hydro – Montreal, CA May 2022 - September 2022

– Engineered a SharePoint infrastructure solution to centralize engineering resources, deployed to 200+ users across the engineering department

– Automated material specification analysis by creating Python scripts (Pandas/NumPy) to process CSV datasets, generating standardized reports that reduced manual review time.

– Conducted user research interviews with 15+ engineers to optimize UI/UX flow, resulting in 95

Skills

– **Programming:** Python, C, MATLAB, PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face Transformers, NumPy, Pandas, Matplotlib, Seaborn, OOP, Bash/Linux, Git, Docker, CI/CD, ML Ops, GCP, ONNX, OpenCV, Postgres, mySQL

– **Languages:** English (Native/Bilingual Proficiency), Mandarin (Native/Bilingual Proficiency), German (Limited Work Proficiency)