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.
- **Elaborative Subtopic Query Reformulation for Query-Driven Recommendation**
  Co-author, *ACM SIGIR*, 2025 (under review).

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
- Developed a Python-based fine-tuning pipeline integrating existing training conversation pairs and formatting scripts to enhance beuaty chatbot by replacing the in-context learning approach
- Finetuned Stable Diffusion LoRA models to improve task-specific GenAI performance, improving accuracy by 40

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.
- Architected the application using object-oriented design patterns, creating a modular system with 10+ interchangeable components for different NLP workflows.
- 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.
- Delivered technical demonstrations to 20+ Meta engineers and executives, showcasing system architecture and RAG features.

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

Projects

- **Subjective Summarization with PCA + LLM** Developed a novel approach to summarize opinion-based datasets using LDA + LLMs to identify key topics and opposing viewpoints. Produced structured summaries that capture stance distributions and core arguments while maintaining interpretability.
- **LLM Powered Resume Modifier for Job Descriptions** Engineered a multi-agent system using Google's Gemini AI to tailor LaTeX resumes. The system parses resumes, condenses job descriptions, and employs relevance scoring and content selection to optimize work experiences and extras. Integrated a VLM feedback loop for formatting refinement and visual consistency.

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)