Justin Cui

Toronto, ON, CA justin.cui@mail.utoronto.ca +1 416 827 9628 linkedin.com/in/Justin-Cui

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% improvement in facepoints prediction accuracy.

– 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% and enhancing scalability.

– Architected end-to-end fine-tuning pipeline for conversational AI systems, replacing in-context learning with supervised fine-tuning on 5k+ curated dialogue pairs.

– Refactored skin diagnosis repository using object-oriented design, consolidating recurring patterns into a unified sign registry and score normalization framework, enhancing code maintainability and consistency.

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

– Refined eyeliner extraction through advanced k-means clustering and adaptive HSV thresholding techniques, boosting accuracy in distinguishing facial makeup regions and optimizing the overall inference pipeline.

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.

– Conducted comprehensive system testing including unit tests (95% coverage) and integration tests with simulated synthetic users using LLMs.

– 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% adoption rate within first deployment month.

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.
- **Geometric Approach to Query Performance Prediction**
  Explored a geometric approach to predicting query performance in information retrieval by analyzing embedding vector properties in high-dimensional space. Leveraged volume characteristics and spatial relationships to guide query reformulation and improve search effectiveness.

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)