Ben Cheng

437-362-4394 | bencheng0609@gmail.com | www.linkedin.com/in/bencheng2 | https://bencheng2.github.io

Education

University of Toronto Ontario, Canada

Specialist in Computer Science (Co-op), Major in Statistics (GPA: 3.92 Dean's List) *Sep. 2021 – June 2026*

Work Experience

5G ML Software Developer May 2024 – Present

Ericsson

- Applied Deep Q-Network (DQN) and Proximal Policy Optimization (PPO) to 5G SRS scheduling and antenna calibration tasks, supporting commercial deployments with Bell, T-Mobile and CMCC
- Integrated AI inference and RL training modules into a C++ Real-time Embedded Communication System
- Contributed to the research, implementation, and field validation of AiXellerator, leading to up to a 20% improvement in real-time downlink throughput and a 10% gain in spectral efficiency (Link)
- Constructed GitLab CI/CD pipelines for automated testing and built InfluxDB & Grafana Time-Series visualization workflows for Radio Access Network (RAN) performance

Full Stack Developer Sep 2023 – Apr 2024

Centre for Social Services Engineering

- Developed a high-performance semantic recommendation engine using React and GraphDB, featuring advanced multi-condition filtering, paginated results, and a responsive Material UI built with TypeScript
- Accelerated SPARQL query execution through Lucene full-text indexing and optimization techniques, reducing average search latency by over 60% while improving system robustness and scalability
- Engineered full-stack integration between Ontotext GraphDB and a NodeJS-based backend, enabling efficient knowledge graph traversal and real-time data access through modularized data pipelines

Software Engineer May 2023 – Apr 2024

Middleware Systems Research Group

- Architected and implemented a modular web platform leveraging Django CMS and Nginx, featuring a custom search engine built with scikit-learn for enhanced query relevance and extensibility
- Developed a robust CI/CD pipeline with GitHub Actions to automate testing, Docker-based deployment, and periodic renewal of TLS certificates for secure server operations
- Integrated Google Scholar API with MySQL and Redis, enabling real-time synchronization and efficient management of publications across research projects and collaborators

Research Experience

Research Assistant (Diffusion + Cellular Automata) May 2025 – Present

Stony Brook University

- Investigated the integration of Neural Cellular Automata (NCA) into diffusion-based image generation frameworks
- Designed and implemented models to replace the CNN backbone in UNet with multi-layer NCA, and conducted experiments on CIFAR-10 and CelebA-HQ datasets demonstrating high-quality generation performance
- Utilized the Accelerate parallel computing library to enable multi-GPU parallel training with mixed-precision optimization improving GPU utilization and speeding up experimentation
- Supervised by Professor Chenyu You at Stony Brook University

Research Assistant (Energy System + Graph Algorithm) July 2024 – May 2025

UAlberta Electrical and Computer Engineering

- Explored the potential of grid-level visualization for efficient, region-aware carbon emission analysis, utilizing Alberta's electricity network as a starting point with the goal of scaling to national or global applications
- Designed and implemented a carbon emission tracing algorithm to attribute emissions to individual generators and regions, enabling Finer-grained Analysis of grid behavior and supporting provincial or national-scale assessments
- Conducted topology-based analysis of the electricity network with Power Dispatch Model, leveraging Multi-Source BFS and Linear Programming (LP) to examine power flow, generator contribution, and consumption allocation
- Supervised by Professor Yize Chen at the University of Alberta

Research Assistant (Medical Time-Series Machine Learning) Jan 2025 – Apr 2025
Vector Institute, University Health Network

- Explored the effectiveness of Feedforward and Recurrent Neural Networks in predicting Tacrolimus blood-level for liver-transplant patients after administration, focusing on modeling temporal and non-linear clinical patterns
- Built and trained custom neural network models in PyTorch on anonymized patient datasets from University Health Network, evaluating predictive accuracy and clinical applicability
- Aimed to support physician decision-making in post-transplant immunosuppressive therapy by delivering personalized, data-driven dosage recommendations grounded in statistical modeling
- Supervised by Professor Michael Brudno and Professor Rahul Krishnan at the Vector Institute

Research Assistant (CSP + Wi-Fi Signal Positioning + LLM) Jan 2024 – Apr 2024
Dynamic Graphics Project (DGP) Lab

- Developed a WiFi-based indoor positioning cartography system capable of estimating spatial layouts and reasoning relative room positions using signal strengths and constraint satisfaction problem (CSP) techniques
- Explored integration of LLMs for generating descriptive map annotations, organizing measurement tasks, and automating operator decision-making during training and deployment
- Pursued cost-effective automation of assistive indoor map generation by leveraging WiFi signal analysis, CSP-based spatial reasoning, and LLM-guided planning to minimize manual intervention and domain-specific configuration
- Supervised by Professor Khai Truong at the Dynamic Graphics Project Lab

Publications

(First Author) Open Datasets for Grid Modeling and Visualization Apr 2025
Ben Cheng, Yize Chen *arXiv*

- Applied Linear Programming and Topology Analysis to infer electricity grid structures from public datasets
- Reconstructed time-resolved nodal demand and supply distributions to visualize Alberta's electricity flow
- Pioneered data-driven approaches for regional-to-national power flow and carbon emission distribution analysis

Projects

Student Assessment Prediction | *PyTorch, Numpy, Pandas, Scikit-learn, Scipy* (Nov 2023 – Dec 2023)

- Participated in a ML project focused on predicting learning efficiency using Item-Response Theory (IRT)
- Conducted comprehensive analyses using multiple algorithms, including KNN and Matrix Factorization, to improve prediction accuracy of students' performance

MovieReviewShare | *React, NodeJS, Express, MongoDB, Redux, Tailwind CSS, Azure* (July 2023 - July 2023)

- Developed a movie-review platform application using NodeJS, ExpressJS and RESTful API
- Deployed backend and database server on Microsoft Azure and MongoDB Atlas cloud platform

SoftStore | *React, Spring Boot, Spring Data JPA, Material UI, TypeScript, Redis* (May 2023 - June 2023)

- Architected a comprehensive e-commerce platform leveraging Spring Boot, React, Spring JPA, and Redis
- Integrated Sa-token authorization and authentication framework, aiming to strengthen data security

CUDA Convolution Image Processing | *CUDA, OpenCV, C++, MSVC, CMake* Aug 2023 - Sep 2023

- Implemented Grayscaling and Gaussian blurring functions through C++ convolution computing
- Optimized the performance by over 30% through GPU parallel computing in the Nvidia CUDA platform
- Utilized MSVC and CMake to structure, build, test, and package the application

Technical Skills

Languages: Java, Python, C, C++, MATLAB, CUDA, JavaScript, TypeScript, HTML, CSS, PHP, R, Go
AI/ML Platform: PyTorch, Scikit-learn, Huggingface, Accelerate, Numpy, Pandas, Scipy
Deep Learning: Diffusion, NLP, LLMs, UNet, CNN, GAN, ViT (Vision Transformer), DQN, PPO, LSTM, Time-series
Database: MySQL, PostgreSQL, MongoDB, Redis, SPARQL, GraphDB, Lucene, InfluxDB
Backend Frameworks: NodeJS, Express, Spring Boot, Django, Spring JPA, MyBatis
Frontend & UI: React, React Native, Material UI, Tailwind CSS, Redux, Expo, Selenium, Jest
DevOps & Tools: Linux, Git, Docker, GitLab CI/CD, Gerrit, Grafana, Nginx, Jenkins, CMake, Makefile, MSVC
Algorithms & Optimization: Linear Programming, Multi-Source BFS, Constraint Satisfaction Problem (CSP)
Others: Microsoft Azure, Google Cloud Platform, Google Scholar API