Benjamin Chislett

chislett.ben@gmail.com | bchislett@nvidia.com

Experience

Senior Software Engineer *Jun 2025 - Present*
NVIDIA, Toronto, ON
- Actively maintaining and developing vLLM as a core committer responsible for speculative decoding, structured outputs, and asynchronous scheduling features
- Delivered key performance optimizations for vLLM, such as overlapped execution for asynchronous scheduling and both kernel-level and algorithmic optimizations for speculative decoding
- Performing original research for novel speculative decoding techniques, and implementing new techniques from latest literature into vLLM and training infrastructure

Software Engineer *Sep 2024 - Jun 2025*
CentML, Toronto, ON
- Developed targeted optimizations on top of vLLM for LLM inference workloads
- Contributed major features to vLLM, with a focus on speculative decoding and vision-language model support
- Acquired by NVIDIA in June 2025

Research Intern *May 2022 - Jan 2023*
Dynamic Graphics Project, University of Toronto
- Researched intrinsic triangulations and coarsening applications for multigrid methods on 3D surfaces
- Developed a novel high-performance intrinsic mesh decimation library in C++ and Python
- Co-authored "Surface Simplification using Intrinsic Error Metrics" in ACM Transactions on Graphics Journal, featured at SIGGRAPH 2023

Research Intern *May 2021 - Sep 2021*
EcoSystem Research Lab, University of Toronto
- Researched auto-scheduling for machine learning compilers in TVM and low-level optimizations for out-performing cuBLAS in CUDA workloads
- Implemented and evaluated novel auto-scheduling algorithms based on the AutoTVM and Ansor
- Received NSERC Undergraduate Student Research Award (USRA) funding

Machine Learning Engineer *Apr 2021 - Aug 2021, July - Aug 2023*
Activeloop AI, California (Remote)
- Designed infrastructure for a cloud machine learning data platform in Python
- Developed machine learning solutions using PyTorch, Tensorflow, and AWS cloud computing
- Architected distributed tensor database systems in C++ and Python, designed and implemented database abstractions and core functionality

Software Developer *Sep 2018 - Jul 2019, May 2020 - Dec 2020*
Mysa Smart Thermostats, St. John's, NL
- Developed and maintained a full-stack TypeScript web interface for an AWS backend
- Authored a suite of libraries for interacting with AWS at multiple layers of abstraction
- Developed various new features for a React-Native mobile application
- Architected a data pipeline used to create a data lake and perform analytics using IaC and SQL

Research Intern *Jul 2019 - Sep 2019*

Okinawa Institute of Science and Technology, Okinawa Prefecture, Japan

- Researched the Compressive Split-Step Fourier Method for solving Gross-Pitaevskii systems. Performed mathematical modeling and validation, optimized numerical GPU solvers, and developed theoretical foundations for order-of-magnitude speedups using compressed frequency methods.
- Maintained GPUE: a CUDA/C++ application for simulating Quantum effects of superfluids
- Authored GPUE.jl: a high performance JuliaLang-GPU implementation of GPUE

Education

First-Year Doctoral Fellow, Computer Science *Sep 2023 - Sep 2024*

EPFL, Lausanne, Vaud, Switzerland

- Member of Wenzel Jakob's Realistic Graphics Lab
- Teaching Assistant for Advanced Computer Graphics (CS440), Spring 2024 Semester
- Developed custom high utilization hardware-accelerated machine learning primitives for differentiable rendering in PTX/CUDA
- Left the program after first year to pursue applied opportunities in industry

Honours Bachelor of Science, Computer Science *Sep 2019 - May 2023 (05/23)*

University of Toronto, Scarborough, ON

- Cumulative GPA: 3.96
- ICPC North America Finalist, 2020/21

Teaching Assistant at University of Toronto

- Artificial Intelligence (CSCD84),
  Principles of Programming Languages (CSCC24) *Winter 2022/23*
- Computability and Computational Complexity (CSCC63),
  Algorithm Design and Analysis (CSCC73), Computer Graphics (CSCD18) *Fall 2022*
- Introduction to the Theory of Computation (CSCB36), Linear Algebra 2 (MATB24) *Summer 2022*
- Introduction to the Theory of Computation (CSCB36) *Fall 2021*
- Linear Algebra 1 (MATA22), Introduction to Computer Science 2 (CSCA48) *Winter 2020/21*