Jason Hon

+1 (226) 753 0193 | jkhhon@uwaterloo.ca | jasonhonhk | JasonH53 | jasonhon.com

Education

University of Waterloo Sept 2023 - Dec 2027

Bachelor of Computer Science, Artificial Intelligence Specialization *GPA: 90.3%*

- **Courses**: Compilers, Algorithms, Data Structures, Operating Systems, OOP, Computer Architecture
- **Awards**: CS Upper Year Scholarship, President's Scholarship of Distinction, President's Research Award (CAD $8000)

Experience

University of Waterloo Jul 2025 - Present

ML Systems Research Assistant *Waterloo, ON, Canada*

- Developed a **masked-attention** pipeline combining small and large diffusion models by computing top-K token divergence per step and selectively routing tokens in **PyTorch**, reducing inference time by **40%** while maintaining within **10%** PSNR
- Gathered initial requirements by computing **L1** distance and **cosine** similarity of attention outputs of various steps, configs

Huawei Canada Jan 2025 - Apr 2025

Compiler Engineer Intern *Markham, ON, Canada*

- Integrated a **MLIR** pass to find optimal tensor parallelization strategies for **PyTorch** models based on device topology
- Optimized tensor parallelization plan search with a **C++ integer linear programming** solver, speeding up search by **15×**
- Improved inference throughput by automating insertions of sharding and collective operations in MLIR graphs, increasing tokens per second by **8%** on NPUs
- Integrated partition templates and constraints for operators in attention and FFN, shrinking strategy search space by **40%**

University of Waterloo Apr 2025 - Aug 2025

Compiler Research Assistant *Waterloo, ON, Canada*

- Integrated static analysis to **Scala's compiler** to ensure safe initialization of global objects, resolving **10+** GitHub reports
- Enforced **partial ordering** in the compiler, reducing debugging cycles for devs by catching additional **25%+** of errors

Super.com Sep 2025 - Dec 2025

Software Engineer Intern *Remote*

- Improved ML training pipeline by optimizing the workflow for the recommender system and dynamic pricing model using **Apache Airflow, Kafka**, and **FastAPI**, increasing customer conversion rate by **8%** through more accurate predictions
- Converted **50+** legacy React class components to functional components using hooks, reducing average component render time by **22%** and cutting bundle size by **18%**

Projects

Lacs Compiler Sept 2024 - Dec 2024

- Designed and implemented a **full compiler** for a **Scala**-like language with support for closures, tail calls optimizations, and memory management via garbage collection
- Integrated DFA lexical analysis, Earley parsing, semantic analysis, register allocation, and code gen to an **IR**

Chess Engine Jun 2024 - Aug 2024

- Developed a Chess engine in **C++** with AI opponents of varying difficulty using a minimax inspired algorithm
- Strictly adhered to **Object-Oriented Design** patterns (MVC, Factory), easing feature expansion and reduced coupling

CodeyBot, UW Computer Science Club Sept 2024 - Aug 2025

- Developed and deployed CodeyBot, a Discord bot using **Docker, SQL** to a server with over **4,500** members
- Spearheaded development of a wordle-like geography guessing game in **TypeScript**, played by over **500** users

Technical Skills

Languages: C/C++, Java, JavaScript/TypeScript, Python, SQL, Scala, TableGen, HTML/CSS, CUDA
Frameworks/Libraries: LLVM/MLIR, React, NodeJS, Flask, FastAPI, NextJS, GraphQL, Tailwind, gRPC
Technologies: Git, Docker, Jenkins, Bash, Unix, AWS, GCP, MySQL, PostgreSQL, Kubernetes, Redis, Datadog, TrackJS