Jason Hon

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

Education

University of Waterloo 2023 - 2027

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

- **Courses**: Compilers, Data Structures, Algorithms, Operating Systems, Computer Architecture, Object Oriented Prog.
- **Awards**: CS Upper Year Scholarship, President's Scholarship of Distinction, President's Research Award (CAD $8000)
- **Extracurriculars:** Hack The North, Computer Science Club, UWCSA, YouTube Channel (20k+ subscribers)

Experience

ML Stack Engineer Intern Jan 2026 - Apr 2026

Cerebras Systems

- Incoming Winter 2026, working with Cerebras' **MLIR** graph compiler in **C++** to lower LLMs

ML Systems Research Assistant Jul 2025 - Apr 2026

University of Waterloo

- Accelerating diffusion models on edge devices with techniques like **quantization**, **hybrid model**, and **masked attention**
- Benchmarked Wan2.1 Text-to-video models (1.3B, 14B, and hybrid) by analyzing **L1** distance and **MSE** across intermediate noise steps, and evaluated the resulting product's quality using **PSNR** scores

Compiler Engineer Intern Jan 2025 - Apr 2025

Huawei Canada

- Integrated a **MLIR** pass to find optimal tensor parallelization strategies for **PyTorch** models based on device topology
- Optimized attention-layer tensor parallelization plan search by implementing a **C++ integer linear programming** solver, achieving a **15×** speedup in search time
- 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%**

Compiler Research Assistant Apr 2025 - Aug 2025

University of Waterloo

- Researched and integrated a static analysis to the **Scala compiler** to ensure the safe initialization of global objects, detecting and resolving **10+** errors of language bug reports on Github
- Enforced **partial ordering** and **initialization time irrelevance** in the compiler, reducing debugging cycles for developers by catching an additional **25%+** of initialization errors

Software Engineer Intern Sept 2025 - Dec 2025

Super.com

- 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

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, CUDA
Technologies: MLIR, LLVM, PyTorch, Git, Docker, Jenkins, Bash, Unix, FastAPI, Kubernetes, AWS