

Hwang Wei Ju

Toronto, ON | hwangweiju@gmail.com | +1 647-865-5552
williamhwang.com | linkedin.com/in/hwang-wei-ju | github.com/williamhwangwei-ju

Education

University of Toronto

BASc in Computer Engineering, GPA: 3.75/4.0, Dean's Honour List

Toronto, ON
Expected June 2027

- **Minors:** Artificial Intelligence Engineering; Engineering Business

Experience

Undergraduate Researcher — Analog In-Memory Computing (AIMC)

May 2026 – Aug 2026

York University

Toronto, ON

- Reduced lifetime model-quality degradation by **27%** versus the strongest baseline by developing a sensitivity-aware placement algorithm for LLM weights across AIMC tiles.
- Lowered all-analog GPT-2 perplexity by **99%** (3,146 to 31.5) through hardware-aware fine-tuning.
- Identified a **540×** noise-sensitivity spread by profiling all 49 GPT-2 projections, pinpointing the LM head as the bottleneck.
- Quantified GPT-2 lifetime degradation by building a tile-aging simulator modeling drift, thermal variation, and faults.

Compiler Engineer Intern

May 2025 – Apr 2026

Huawei Technologies Canada

Toronto, ON

- Reduced CPU power consumption by **2.8%** in gem5 by developing an LLVM pass for AArch64 branch-predictor gating.
- Improved Instructions Per Cycle (IPC) by **1.65%** by extending the TRRIP cache-replacement policy to the data cache in gem5.
- Increased IPC by **1.1%** on mobile workloads by implementing runtime-guided instruction prefetching in gem5.
- Streamlined trace and ELF generation for **20+ performance engineers** with automated pipelines adopted team-wide.

Full-Stack Software Engineer Intern

May 2024 – Aug 2024

BTG Pactual

São Paulo, Brazil

- Cut financial-data processing time by **40%** by automating migrations with AWS Lambda serverless pipelines.
- Shipped a full-stack banking feature, enabling clients to modify payment amount, date, payer, and payee in real time.
- Decoupled payment processing for independent service scaling by building AWS SQS and SNS workflows in Terraform.

Publications

- **Hwang Wei Ju**, Fatemeh Gholami, Roman Genov, and Amirali Amirsoleimani, "Sensitivity-Aware Placement Sustains All-Analog GPT-2 Inference on Aging In-Memory Tiles," under review at *IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2027.

Projects & Extracurriculars

University of Toronto Robotics Association — Computer Vision Engineer

Sep 2022 – May 2024

- Reached **91.6% IoU** at **12.8 ms/frame** by training a deep-learning lane-segmentation model on 1,174 hand-labeled images.
- Achieved an **85%** ramp-navigation success rate by designing a system with dual-LiDAR ramp detection and waypoint planning.

Alzheimer's Detection Deep Learning Model

[GitHub](#)

- Achieved **96.5%** accuracy on four dementia stages by fine-tuning ResNet152, versus 74.3% for Random Forest.
- Implemented Optuna hyperparameter search over learning rate, batch size, and training epochs to optimize accuracy.

Geographic Information System (GIS) Software

[GitHub](#)

- Reduced average traveling-courier route cost by **14.6%** using multistart search, route perturbation, and simulated annealing.
- Accelerated map rendering by **7×** through multithreading, improving interface responsiveness.

Technical Skills

- **Machine Learning:** PyTorch, TensorFlow, Hugging Face Transformers, AIHWKit, LLM inference, computer vision
- **Languages:** Python, C, C++, TypeScript, JavaScript, Java, SQL, Bash, Verilog, AArch64 Assembly, MATLAB, HTML/CSS
- **Systems & Compilers:** LLVM, compiler optimization, gem5, computer architecture, performance profiling, Linux, ROS
- **Web, Cloud & Tools:** React, Node.js, NestJS, REST API, AWS, Terraform, Docker, Git, CI/CD, Azure DevOps