Charlie Ruan

charlieruan.com | cfruan@cs.cmu.edu

EDUCATION

Carnegie Mellon University, Computer Science Department **Pittsburgh, PA | Aug 2023 - May 2025**
M.S., Computer Science
GPA: 4.17/4.33

Cornell University, College of Engineering **Ithaca, NY | Aug 2019 - May 2023**
B.S., Computer Science and *Operations Research*
GPA: 4.0/4.3; *Summa Cum Laude*

PUBLICATIONS & MANUSCRIPTS

(* denotes equal contribution)

- **Charlie F. Ruan**, Yucheng Qin, Xun Zhou, Ruihang Lai, Hongyi Jin, Yixin Dong, Bohan Hou, Mengshiun Yu, Yiyan Zhai, Sudeep Agarwal, Hangrui Cao, Siyuan Feng, Tianqi Chen. "WebLLM: A High-Performance In-Browser LLM Inference Engine." *Will submit to **JMLR (MLOSS)***. https://arxiv.org/abs/2412.15803
- Hongyi Jin*, Ruihang Lai*, **Charlie F. Ruan***, Yingcheng Wang*, Todd Mowry, Xupeng Miao, Zhihao Jia, Tianqi Chen. "A System for Microserving of LLMs." ***Under submission***. https://arxiv.org/abs/2412.12488
- Yixin Dong, **Charlie F. Ruan**, Yaxing Cai, Ziyi Xu, Yilong Zhao, Ruihang Lai, Tianqi Chen. "XGrammar: Flexible and Efficient Structured Generation Engine for Large Language Models." ***Under submission***. https://arxiv.org/abs/2411.15100
- Siyuan Feng*, Jiawei Liu*, Ruihang Lai, **Charlie F. Ruan**, Yong Yu, Lingming Zhang, Tianqi Chen. "Productive Deployment of Emerging Models on Emerging Platforms: A Top-Down Approach." ***Under submission***. https://arxiv.org/abs/2404.09151
- Xun Zhou, **Charlie Ruan**, Zihe Zhao, Tianqi Chen, Chris Donahue. "Local Deployment of Large-Scale Music AI Models On Commodity Hardware." ***ISMIR 2024 (LBD session)***. https://arxiv.org/abs/2411.09625
- A. Feder Cooper*, Wentao Guo*, Khiem Pham*, Tiancheng Yuan, **Charlie F. Ruan**, Yucheng Lu, Christopher De Sa. "CD-GraB: Coordinating Distributed Example Orders for Provably Accelerated Training." ***NeurIPS 2023***. https://arxiv.org/abs/2302.00845
- **Charlie Ruan**. "Approximating Martingale Process for Variance Reduction in Deep Reinforcement Learning with Large State Space." *On arXiv November 2022*. https://arxiv.org/abs/2211.15886

OPEN-SOURCE PROJECTS

MLC-LLM *Core Contributor* **GitHub ★ 19.3k ⑂ 1.6k (#4 contributor)** **Pittsburgh, PA | Jun 2023 – Present**
PI: Prof. Tianqi Chen

- Enabling universal native deployment of LLMs through machine learning compilation techniques including TVM, supporting non-conventional backends like AMD (ROCm kernels) and Apple (Metal kernels); building an LLM serving system on top of it

WebLLM *Project Lead* **GitHub ★ 14.0k ⑂ 906 (#1 contributor)** **Pittsburgh, PA | Jun 2023 – Present**
PI: Prof. Tianqi Chen

- Deploying LLMs locally in web browsers with WebGPU for GPU acceleration and WebAssembly for performant CPU computation
- Paving the way for on-device agents to automate daily in-browser tasks (e.g. drafting emails, editing documents, booking tickets)
- Widely recognized in the JavaScript/Web community (talk at Google WebAI Summit '24)

RESEARCH EXPERIENCE

Sky Computing Lab *Research Assistant* *(GPU Programming, Kernel Language/Compiler)* **Berkeley, CA | Aug 2024 – Present**
PI: Prof. Ion Stoica

- Working on a GPU kernel language/compiler for automating grid-level optimizations; extensively worked with Triton and MLIR

Catalyst Group *Research Assistant* *(Distributed Systems, LLM Serving)* **Pittsburgh, PA | Mar 2024 – Present**
PI: Prof. Tianqi Chen, Prof. Zhihao Jia

- Proposed an LLM microserving architecture that enables dynamic reconfiguration of various disaggregation and coordination patterns, including balanced prefill/decode disaggregation, KV transfer, and distributed prefix cache
- Explored other disaggregated strategies such as attention/non-attention, and long request/short request; built a distributed system that supports point-to-point remote attention with CUDA kernels and the NVSHMEM communication library
- Paper under submission*:* https://arxiv.org/abs/2412.12488

RelaxML Lab *Research Assistant (Distributed Machine Learning)* **Ithaca, NY | Sep 2022 – May 2023**

PI: Prof. Christopher De Sa

- Investigated finding provably better data permutations in distributed training with decentralized data, using recently proposed example-ordering algorithm Gradient Balancing (GraB)
- Built a distributed training system in a decentralized fashion to analyze bounds on convergence rate and consensus error
- Paper accepted by ***NeurIPS 2023***: https://arxiv.org/abs/2302.00845

Variance Reduction for Reinforcement Learning *Research Assistant (RL)* **Ithaca, NY | Dec 2021 – Sep 2022**

PI: Prof. Jim Dai

- Used reinforcement learning (RL) to optimize the algorithm of matching drivers and customers on ride-hailing systems like Uber
- Formulated the application of variance-reduction method approximating martingale-process (AMP) in proximal policy optimization (PPO) when state space is large and state transitions are uncertain; experimented on ride-hailing and multiclass queueing networks
- Manuscript available: https://arxiv.org/abs/2211.15886

INDUSTRY EXPERIENCE

Google Core ML *Software Engineer Intern (TensorFlow, Python)* **Sunnyvale, CA | Jun 2023 – Aug 2023**

- Worked on TensorFlow's Distributed Runtime team; optimized TensorFlow's asynchronous checkpoint in Keras, offloading model checkpointing to an asynchronous thread to reduce wasted TPU cycles
- Received a return offer; contributed 1700+ LOC to https://github.com/tensorflow

Google Cloud *Software Engineer Intern (Platform Engineering, Linux, C++)* **Sunnyvale, CA | Aug 2022 – Oct 2022**

- Worked on TechInfra team to deploy GPUs in data centers; implemented Linux daemons to provide GPU firmware updates and monitor GPU health with D-Bus and I2C using OpenBMC; worked with pre-production hardware with limited debugging support
- Received a spot bonus and a return offer

Amazon Robotics *Software Engineer Intern (Full-Stack, Kotlin, Java)* **Greater Boston, MA | May 2022 – Jul 2022**

- Worked on robots for warehouse automation solutions; implemented a full-stack configuration portal on Amazon Robotics's HCI software, allowing warehouse workers to personalize their interaction with the autonomous warehouse robots
- Received return offer for a full-time position

XPeng Motors *Software Engineer Intern (Sensor Fusion, Python, C++)* **Shanghai, China | Jun 2021 – Aug 2021**

- Worked on the Sensor Fusion team for XPeng's self-driving cars; processed and fused various sensor data (e.g. radars, cameras) of XPeng's self-driving cars to provide a reliable perception result

Morgina Information Technology *Software Engineer Intern (C/C++, Embedded)* **Shanghai, China | Jun 2020 – Jul 2020**

- Optimized the multi-object tracking algorithm of millimeter-wave radars installed in intersections that monitor traffic information

STUDENT ACTIVITIES

Cornell Electric Vehicles *Software Engineer (Python, ROS, Linux)* **Ithaca, NY | Aug 2019 – Mar 2022**

- Designed the ROS (Robot Operating System) for the vehicle's autonomy system; engineered a platform for communications between sensors (e.g. LIDAR, IMU) and algorithms, as well as among algorithms (e.g. vision, localization)

TEACHING EXPERIENCE

Intro to Engineering Stochastic Processes *Teaching Assistant* **Ithaca, NY | Jan 2023 – May 2023**

- Topics include: discrete-time/continuous-time Markov chain, Poisson process, queueing theory, Markov decision process
- Was the sole TA in charge of the design and office hours for a project that compares Monte Carlo simulation with neural networks

Intro to Machine Learning *Teaching Assistant* **Ithaca, NY | Aug 2021 – Dec 2021**

- Topics include: decision trees, support vector machine, kernels, neural networks, statistical learning theory, online learning, boosting

AWARDS & HONORS

Cornell Engineering Dean's Honor List (for all semesters)	**2019 – 2023**
Omega Rho Honor Society for Operations Research	**May 2023**
Undergraduate Summer Research Funding, School of Operations Research (five students selected in total)	**May 2022**
Tau Beta Pi Engineering Honor Society	**Mar 2022**