Ryan Zhang

cs.ryanzhang@gmail.com | linkedin.com/in/ryanzhang-link | github.com/rytuph

ML and SWE work experience with high-impact results from startup, large corporation and research domains. ML research experience with end-to-end model R&D in PyTorch. Applied MLE/MLOps/SWE experience developing and deploying containerized models into active company workflows with FastAPI and Docker. SWE experience with optimizing low-level algorithms in C and building backend systems and CI/CD pipelines in a fast-paced startup environment. Open to SWE/MLE internships in-person or remote across North America.

Skills

Languages: Python, C, C++, C#, Java, SQL (MySQL, PostgreSQL), Bash, JavaScript
Frameworks/Libraries: PyTorch, FastAPI, Scikit-learn, Pandas, NumPy, TensorFlow, PyTorch Lightning, OpenCV, React, Next.js, Node.js, Express.js, REST API
Tools & Processes: Docker, AWS, EC2, DynamoDB, Lambda, IAM, API Gateway, Git, CI/CD, GitHub Actions, Weights & Biases, Jupyter Notebooks, Hugging Face, CUDA, MongoDB, Agile
Concepts: Deep learning, Generative AI, Computer Vision, MLOps, Optimization, Infrastucture, Algorithms

Work Experience

Machine Learning Engineer Intern   May 2025 – Present
BMW Group | *Python, PyTorch, FastAPI, Docker, WandB, Hugging Face, Pandas*   *Munich, Germany*
- Finetuned text-to-image generative AI model (FLUX) with PyTorch to additionally accept car keypoint image inputs to generate controllable, high quality 3D car models to save designer teams 4-16 weeks of car design prototyping
- Enhanced generated car quality by 39% by applying (LoRA) Parameter Efficient Fine-Tuning (PEFT) to FLUX
- Enabled precise vehicle keypoint map control, improving benchmarks by 88% by modifying the Transformer's input layer to accept keypoint image conditioning along with text inputs (dual conditioning)
- Built containerized REST API (FastAPI) service with Docker achieving <200ms P95 latency via async handling
- Developed CI/CD pipeline using GitHub Actions to automate testing, deployment of service, ensuring stable, versioned model integration into non-technical designer team workflows

Machine Learning Research Intern   May 2024 – Aug 2024
National Research Council Canada | *Python, PyTorch, Pytorch Lightning, Pandas, WandB*   *Ottawa, Ontario*
- Advanced novel battery material discovery by training Meta's Flow Matching Generative AI (FlowMM) model on a proprietary research dataset to generate prospective crystal structures as replacement molecules for Li+ batteries
- Improved stability by 15% and reduced inference cost 8x vs previous diffusion GenAI model by finetuning FlowMM hyperparameters with PyTorch Lightning on HPC clusters
- Quadrupled data preprocessing speed of 50,000+ structure dataset by redesigning single-threaded pipeline to a parallel and memory-efficient architecture in Python
- Boosted battery candidate (DFT) hit-rate by 30% by inferencing and validating dataset of 10K+ crystal structures

Software Engineer Intern   May 2023 – Aug 2023
Quantropi | *C, Python, Docker, GitHub Actions, CI/CD, AWS*   *Ottawa, Ontario*
- Rebuilt bignum library in C to accelerate arithmetic operations (kernels) in Post-Quantum Cryptography products
- Optimized kernels by 20x by replacing naive algorithms with asymptotic and hardware efficient methods
- Ensured reliable deployment by building a CI/CD pipeline with GitHub Actions, Docker and AWS EC2 Graviton to validate cross-platform portability and automate benchmarking (8–128 bytes)
- Developed differential fuzzer in Python to verify functional equivalence of new kernels against past implementations

Projects

Touchless | Winner of uOttaHacks Overall (600+ participants), Healthcare Challenge, Accessibility Hack Winner
- Built gesture and voice interface with OpenCV, Mediapipe & ML to reduce contact in high-risk settings
- Built secure backend with AWS (IAM, API Gateway) to send Flask data to DynamoDB for real-time analytics

SQLidify | Winner of McGill Hacks Cohere challenge (400+ challengers)
- Built fullstack SQL injection detector by finetuning Cohere's Classify LLM model, achieving 92% accuracy
- Developed the classification pipeline in Python/Flask and a user-facing review interface in React

Education

Carleton University | *Bachelor of Computer Science AI & ML Stream (Honours)*   Sept 2022 – May 2027
GPA: 3.85/4.0 (A)   *Ottawa, Ontario*
Extra Curriculars: Founder of CU Tennis Club, Exchange at Ludwig Maximilian University of Munich