Bob Wei

bobqywei@gmail.com | bobqywei.github.io/me | /bobqywei | /bobqywei

Experience

Microsoft AI (from Inflection AI) — *Mountain View, CA*

Principal Member of Technical Staff (L65) — *April 2024 - Present*

- Core contributor to the **reasoning** of our best models (*MAI-1-preview*), pushing multi-turn RL in long horizon tasks
- Founding team member on the **post-training alignment** stack shipping the models behind *Microsoft Copilot*.
- Onsite engineer at OpenAI. Led the integration of **GPT4o-voice** capabilities into Copilot. Led the Microsoft AI research team that spun up **strawberry RL** internally and conducted experiments to understand its implications.

Inflection AI (acquired by Microsoft) — *Palo Alto, CA*

Member of Technical Staff — *March 2023 - March 2024*

- Core contributor in scaling up our pretraining stack for then best-in-class **1e25** *Inflection 2* model, incorporating **fp8 on H100, ZeRO-1,** and **interleaved pipeline parallelism**. Improved MFU and maintained deterministic debugging
- Co-led the data and algorithm (SFT, DPO) iteration to push math and code performance for *Inflection 2.5*
- Helped build our post-training stack for human feedback and pushed on the models backing our product **Pi**

Google X: Everyday Robots — *Mountain View, CA*

Machine Learning Engineer — *July 2022 - February 2023*

- Trained and deployed VLM's for robot perception, enabling real-time **open-vocabulary object detection**
- Led integration of perception models into end-to-end RL stack, reducing error by >60% in long-horizon manipulation tasks, and improving adaptation to unseen environments (*Deep RL at Scale paper*)

Nvidia — *Toronto, ON*

Research Scientist Intern — *February 2021 - May 2021*

- Sped up training of large GAN's (PixelGAN, BigGAN) on real-world datasets (FFHQ); supervised by **Dr. Sanja Fidler**
- Implemented and maintained custom optimizers and higher order gradient algorithms in a large **Pytorch** codebase

Nvidia — *Santa Clara, CA*

Computer Vision Engineer Intern — *June 2020 - September 2020*

- Reduced object detection post-processing time from **7ms** to **1.7ms** in **C++** production codebase for Tegra autonomous systems. Implemented novel **probabilistic voting** method with efficient **CUDA** kernels, replacing serial NMS
- Proposed a novel scale-invariant loss for poly-line detection, increasing F1 score by > **5%**

Uber Advanced Technologies Group (now Waabi) — *Toronto, ON*

Research Scientist Intern — *September 2019 - May 2020*

- First authored a paper accepted to **IEEE ICRA 2021** (*arxiv.org/abs/2011.01153*); supervised by **Dr. Raquel Urtasun**.
- Contributed to the research and development of a novel, end-to-end neural network for vehicle motion planning

Side Effects Software — *Toronto, ON*

Software Engineer Intern — *January. 2019 - April. 2019*

- Designed an interactive 3D terrain generation tool: *sidefx.com/tutorials/machine-learning-data-preparation/*
- Developed generative models (**pix2pix GAN**) to simulate erosion over **50,000×** faster than conventional methods

Projects

Flow

uWaterloo Course Ratings + Reviews

- **uwflow.com** is the go-to website for course reviews at uWaterloo with over **25,000** monthly active users
- Co-developed the backend infrastructure from the ground up with **Golang**, **Postgres**, and **Hasura** at the core

Storyview

- **storyview.me** is a platform I built for logging and sharing my adventures!

Education

University of Waterloo

Bachelor of Science in Computer Science (Dean's Honours, 3.95/4.0 GPA, 92%) — *September 2017 - April 2022*