Raymond Li

work@raymond.li · https://Raymond.Li · github.com/Raymo111 · linkedin.com/in/Raymo111

EXPERIENCE

Meta — Menlo Park, CA

Software Engineer, AI Infra - Experiences ML — **2025 – 2026**

- Built distributed training/inference analysis platform for GPU utilization and kernel optimization
- Re-architected monolithic analysis pipeline into highly-parallel **Prefect**-like DAG workflow
  - Reduced latency by **70%** (**7000 hours** of runtime saved **daily** across **150k** runs)
  - Eliminated OOMs, boosting overall execution success rate from **82%** to **99%**
  - Enabled partial retries on failures and improved observability (critical for SEV root-causing)
- Introduced **Protobuf** to PyTorch/Kineto for **3x** speedup and **85%** less memory than JSON traces
- **100%** AI-native agentic development since Oct 2025. Built Claude skills to solve domain-specific and general human-on-the-loop challenges, replacing manual work in various internal processes.

Production Engineer, PyTorch — **2024 – 2025**

- Drove **4%** speedup in **torch.compile**, saving **$10M** annual recurring infra costs, via optimization of node fusion in **TorchInductor**. Also improved PT2 compile observability, reliability and scalability.
- Built **torch.compile** efficiency analysis system and gleaned insights for top-10 models company-wide, providing data for org-level PyTorch compile engineering decision-making

Tesla — Palo Alto, CA

Software Engineering — **2024**

- Built multi-layered caching API for **SCons** build system to speed up builds by **70%** (**Python**, **Linux**)
- Automated tedious manual build-test-deploy process through a **Jenkins** CI pipeline and **Docker** containers, saving **3 days** per release cycle for core library used to build **200+** components
- Spearheaded upgrade of all (**9**) components built on **Ubuntu** 12.04 to 20.04. Full toolchain upgrades (e.g. **GCC** 4.X -> 10.5). Coordinated **16** Engineers in **5+** teams across **3** orgs.

Amazon Web Services (AWS) — Seattle, WA

Software Dev Engineering — **2023**

- Developed **Python** job to clean up **30K+** incommunicado (**96%** of all) DynamoDB Transaction-Coordinator nodes, significantly reducing EC2 usage costs and receiving SVP commendation

CharaChorder

Open Source — **2023 –**

- Led development of nexus, a cross-platform typing analysis desktop app, and CharaChorder Zero, a software-emulated version of fluid chorded-character entry hardware devices

PlayStation (Sony Interactive Entertainment) — Waterloo, ON

Software Dev Engineering — **2022**

- Independently concepted, designed, developed and documented **first ever PS5 Easter Egg**
- Migrated **50%** of custom **React Native** tile components in PS5 Control Center to SDK, simplifying codebase, reducing compilation time and improving maintainability

Software Dev Engineering in Test — **2021**

- First prize, PlayStation worldwide hackathon (game-recommendation app, out of **100+** teams)
- Developed Chrome Extension to compare different versions of test specifications, and scripts to automatically sort, version, label and template pytest test scripts, reducing time required by 90%

PROJECTS

i3lock-color 🔗 **C, X11** **1K+ stars, 130+ forks, 10K+ users**

- World's most popular non-default computer lockscreen

VoiceBox 🔗 **React.js, GCP**

- Text-to-speech webapp using **Google Cloud AI**, acquired by non-profit education org

EDUCATION

University of Waterloo **Bachelor of Computer Science, Dean's Honours**

- 3.9 GPA. Co-designed CS 136L **Linux** course (incl. Bash, Git, I/O Testing, Make, Debugging, IDEs)

UW Computer Science Club Sysadmin

- Led two teams (Systems Committee and Terminal Committee) of **20+** CS students to configure, upgrade and maintain **35+** servers, on-prem web- and email-hosting, **LDAP**, **Gitea**, **IRC/ZNC**, **NextCloud**, video-conferencing (BBB), **MediaWiki**, **CloudStack**, and **K8s**
- Responsible for the largest open-source software mirror in Canada
  - **100k+** HTTP requests/hour, **1.6 petabytes** (1600 TB) of data served annually

Undergraduate Research Assistant

- Built various ML models (KNN, SVM, random forest, etc) using **Python**, **pandas**, **NumPy** and **scikit-learn** to classify pilots' experience levels based on ALSIM simulator data of real pilot landings, creating world's-first system for AI-driven pilot experience evaluation.

SKILLS

Languages: Python, JavaScript (React, Node), Kotlin/Java, C/C++, HTML/CSS, Bash

Technologies: Git, Docker, Jenkins, GCP/AWS/Cloudflare, Apache, Arch Linux, WSL

AI Systems: Claude code, OpenAI Codex, Gemini, OpenCode, OpenClaw, multi-agent systems

Interests: Enjoys snowsports, horse riding, flying (GA), top-roping, chess. Plays violin. #btwIuseArch