Wei-Chung (Wells) Lu

wellslu.wc@gmail.com | 805 896 7712 | Website | LinkedIn | Github

Research Interest

Efficient Machine Learning and Systems for ML, with a primary focus on enhancing the robustness, reliability, and efficiency of Large Language Models (LLMs).

Education

University of California Santa Barbara, Master in Electrical and Computer Engineering — Sept. 2024 - Present

- GPA: 3.85/4.0

Soochow University, Bachelor in Data Science — Sept. 2017 - June 2021

- GPA: 3.74/4.0
- Dean's list: 2020, 2021

Research Experience

Research Intern, Cornell University — June 2025 - Present

- Assisted Prof.Saikat Dutta in conducting primary source research on the Omnicode, a benchmark that contains a broader and more diverse set of tasks for code-generated AI agents.
- Anonymous, **Wei-Chung Lu**, et al. "OmniCode: A Benchmark for Evaluating Software Development Agents," submitted to the International Conference on Learning Representations (ICLR), 2026. (Under double-blind review)

Research Assistance, University of California Santa Barbara — Sept. 2025 - Present

- Conducted independent research advised by Prof. Peng Li, focused on mitigating LLM uncertainty; proposed a method leveraging selective KV cache to optimize inference reliability.
- **Wei-Chung Lu**, Peng Li. "SKV (Semantic KV): Decoupling Uncertainty from Generation via Selective Caching," ongoing.

Research Assistance, Taipei Veterans General Hospital — Oct. 2020 - Aug. 2021

- Assisted Dr.Jong-Ling Fuh in conducting primary source research on the method for analyzing the risk of developing Alzheimer's disease.
- Implemented pre-processing and data mining procedures for MRI images and patient biological data.

Work Experience

AI Agent Engineer, NXP — April 2026 -

Artificial Intelligence Engineer, Kipt — Feb. 2024 - June 2024

- Spearheaded the development of a CTPN model to locate text detection, and developed a CNN-based OCR model with over 99% accuracy.

Artificial Intelligence Engineer, Open AI Fab — Mar. 2022 - Nov. 2023

- Independently developed a MobileNetV3-FCOS to identify children's development with over 80% mAP (mean Average Precision), and 25 fps in each test.

- Trained a MobileNetV3-SSD model to detect sock tops with over 95% mAP and inference less than 0.01 seconds.
- Converted model with Core ML and Snapdragon SDK for applying to edge or mobile devices.

Machine Learning Engineer, Merkle Aug. 2022 - May 2023

- Refined MLOps on GCP to automate model updates of customer tagging projects to reduce 80% working time.
- Enhanced prediction of collaborative filtering to improve sales performance uplift rate by 5% in A/B Testing.
- Mentored two interns to optimize program efficiency to reduce RAM usage and data load time by almost 50%.

Side Projects

Compare the Inference Speed of TVM and C++ on Edge Device Winter, 2025

- Cross-compiled AI models (take ResNet family as example) with TVM for implementing on Raspberry Pi 3 B+.
- Ran models with C++ source code and libtorch on Raspberry Pi 3 B+ to compare these three methods.
- Tools Used: Python, Pytorch, C++, TVM, libtorch

CP-Decomposition in CNN Fall, 2024

- Replaced Conv layer in LeNet model with CP-Decomposition Conv layer and trained it to reduce model weights.
- Tools Used: Python, Pytorch, C++, Cuda

Customer Tracking System Fall, 2020

- Developed a multi-model system for retail stores to identify new customers, customers who just came a few minutes ago, and members.
- Fine-tuned Joint Detection and Embedding (JDE) model to track and identify different customers with over 73% MOTA.
- Tools Used: Python, Pytorch, SQL, Linux, Parallel Computing

Skills

Programming: Python, C/C++, Linux, SQL, HTML, Javascript, Java, R, Verilog

SDE Skill: Git, Docker, Singularity/Apptainer, PyTorch, Tensorflow, GCP, AWS, Azure, CoreML, Snapdragon SDK

Language: English, Chinese, Japanese

Certification

Google Cloud Certified - Professional Data Engineer	Mar.2023
The AI training certification from Qualcomm	June 2022
Salesforce Certified Administrator	Feb. 2023