Ho Ko

Bay Area, CA

hoko.xyz | ho.woody.ko@gmail.com +1.626.731.7537

EDUCATION

University of Southern California Los Angeles, California
Masters of Science; Computer Science *August 2021 – May 2023*
GPA 3.95 / 4.0

University of Waterloo Waterloo, Ontario
Masters of Engineering; Electrical and Computer Engineering *September 2020 – August 2021*
Specialization in Artificial Intelligence and Machine Learning *GPA 3.9 / 4.0*

National Central University Taoyuan, Taiwan
Bachelors of Science; Computer Science and Information Engineering *July 2015 – June 2019*
GPA: 3.75 / 4.0

WORK EXPERIENCE

Machine Learning Engineer | Google / Lens | Mountain View, California *Feb 2026 – Current*

- **Powered Google Lens, AI Mode in Google Search, and AI Overviews** by building advanced **retrieval models** to ensure precise **information grounding.**
- **Developed deep learning ranking models** to optimize output quality while **distilling Gemini** into highly compressed, efficient models for the **Google Search stack.**
- **Collaborated with Google Health and Fitbit** to utilize **foundational PPG model** for representation extraction and build end-to-end models predicting **hypertension** from raw smartwatch signals.

Software Engineer | Google / Google AI Edge | Sunnyvale, California *Aug 2023 – Feb 2026*

- **Engineered a next-generation, on-device inference engine** by spearheading core optimizations that delivered a **200-300% performance increase** for LLMs like **Gemini Nano** and **Gemma**, enabling secure, private, and offline functionality across mobile, web, and desktop platforms.
- **Collaborated with Google DeepMind** on a research project to develop **Sparse RAG**, a novel system for optimizing Retrieval-Augmented Generation. Our joint effort, which resulted in a co-authored paper at **ICLR**, established an efficient method to improve model quality and performance by integrating **reranking and inference.**
- Architected and developed the engine's core components to natively support real-time, **multi-modal** inputs, including **streaming vision** and **audio.**

Software Engineer | Mycarelinq Inc | Santa Cruz, California *Feb 2023 – Mar 2023*

- **Led the development of a secure RAG system** and **function calling framework**, defining data ingestion strategies and engineering **natural language pipelines** that enabled accurate, context-aware information retrieval and automated, real-time database management.

Software Engineer Intern | Google / TensorFlow Lite | Mountain View, California *May 2022 – Aug 2022*

- Developed core **TensorFlow Lite** developer tools to fully automate **C++ and Python inference code generation**, accelerating ML deployment workflows through **100% coverage** of all common **Task Library** model use cases.

AI Engineer and Researcher | Ubiik | Taipei, Taiwan *Jan 2020 – Jun 2020*

- **Pioneered an end-to-end solar forecasting pipeline** by engineering a novel **noise-reduction data fusion** technique that unified web-crawled production data with satellite imagery, and architected a modified U-Net model to predict cloud movement, achieving **>99% average solar power production prediction accuracy** across Taiwan.

Research Assistant | Caltech | Pasadena, California *Jun 2018 – Aug 2018*

- **Co-authored research in *The Astrophysical Journal Letters* on Type IIb supernovae** by analyzing large-scale ZTF imaging data, coordinating multi-wavelength observational campaigns (**Keck, SWIFT, P200**), and simulating **computational physics models** of double-peak light curves.