KUAN-WEI LU

stanley860920@gmail.com · github.com/Stanleyluuuu · LinkedIn · Portfolio · (+886) 956-970-920

LLM/VLM Serving · ML Platform & Evaluation · AI Agents · Computer Vision · Python/C++

WORK EXPERIENCE

AI Software Engineer II @ Garmin Ltd. *November 2023 – present | Xindian, Taiwan*

Promoted from AI Software Engineer I, June 2026

- Drove company-wide GenAI adoption: evaluated 10+ LLMs/VLMs through structured POCs against internal benchmarks, leading to strategic model selection for **5 core production use cases**.
- Mentor engineering interns to independently ship production-grade services; serve as a **technical interviewer for 1+ year** (resulted in 2 full-time hires).

PROJECTS

Enterprise GenAI Platform — LLM/VLM Serving & Evaluation *April 2024 – present | Garmin*

Role: Platform Owner | Tech: vLLM, Kubernetes, Kong Gateway, Grafana, LLM/VLM, Python

- Built the company's **automated LLM/VLM evaluation framework** (LLM-as-judge, Ragas); optimized infrastructure via a 2nd-gen VLM, boosting inference speed by **3× with higher accuracy**. The platform scales to 30+ production use cases across 8 teams on Kubernetes, handling **2M+ requests/week**.
- Shipped reusable tool-calling configs and skill packages exposing platform APIs to internal AI coding agents, speeding up cross-team development.
- Designed the platform's API gateway layer via Kong across 16 APIs / 52 projects, sustaining **∼99.9% availability** with zero production outages; engineered Grafana dashboards for metric visualization that detected anomalies where Kong successfully blocked ∼4,000 unauthorized requests.

Real-time CV Surveillance & Analytics Platform *November 2023 – present | Garmin*

Role: End-to-end Owner & Maintainer | Tech: Python, Computer Vision, VLM, Apache Kafka, MySQL, MongoDB

- Built and operate an enterprise-grade CV pipeline across 5 sites, processing **63M+ telemetry events/month** via Kafka and fusing detection, tracking, and pose estimation to deliver **∼4,000 critical alerts/month**.
- Led a major system refactor that **cut per-model GPU memory by 70%** (from 2 GB to 600 MB) and tripled models served per GPU by deduplicating model instances across containers and implementing FP16 half-precision.
- Own code review and cross-team feature integration; introduced 10 new monitoring topic types and mentored junior team members on operations.

AI-Powered Document Inspection System *March 2024 – present | Garmin*

Role: End-to-end Owner | Tech: Python, VLM, OCR, Computer Vision, Automation

- Engineered an automated document-verification pipeline combining rule-based parsing with VLM-based data extraction, **saving 5,300+ manual hours annually** across 3 business units.
- Replaced legacy OCR with robust VLM extraction to resolve systemic text and barcode misinterpretation, enabling automated compliance validation across **hundreds of pages** per document.

Multimodal Tool-calling AI Agent *May 2026 – present | Garmin*

Role: End-to-end Owner | Tech: LangGraph, LangChain, Open-Source LLM

- Designing a **stateful tool-calling agent** on an internally hosted LLM that orchestrates multi-step workflows (detection, OCR, TTS/ASR) from natural language prompts.
- Formulating structured tool schemas and **deterministic error-recovery graphs** to guarantee agentic reliability across multi-step workflows.

TECHNICAL SKILLS

- **Languages:** Python (Primary); C++ for parallel & high-performance computing (OpenMP, MPI, CUDA)
- **LLM & GenAI:** LLM/VLM Serving (vLLM), AI Agents & Tool-Calling (LangGraph, LangChain), LLM Evaluation (Ragas, LLM-as-judge)
- **ML & CV:** PyTorch, OpenCV, Object Detection, Object Tracking, Pose Estimation, Segmentation, Scikit-Learn
- **Platform & Infra:** Docker, Kubernetes, Kong API Gateway, Apache Kafka, Grafana, MySQL, MongoDB, API Design, Git

EDUCATION

National Tsing Hua University (NTHU), Hsinchu, Taiwan *July 2020 – July 2022*

M.S. in Electrical Engineering, advised by Prof. Min Sun; Research: Semi-supervised learning for 2D object detection.

*Publication: "Robust 360-8PA: Redesigning the Normalized 8-point Algorithm for 360-FoV Images." B. Solarte, C.-H. Wu, **K.-W. Lu**, et al. ICRA 2021.*

National Chung Hsing University (NCHU), Taichung, Taiwan *September 2016 – June 2020*

B.S. in Bio-Industrial Mechatronics Engineering