Wei Cheng (Wayne) Chiu

waynehacking8@gmail.com | LinkedIn | GitHub | Website | PR Wall

Summary

Field Application Engineer working across customer-facing on-premises LLM deployment and the GPU inference stack — Linux / Kubernetes operations and troubleshooting on top, TensorRT-LLM / vLLM / SGLang serving in the middle, and CUDA / Tensor-Core kernels underneath. At Taiwan AILabs, I put on-premises LLM systems (FedGPT) into customer environments. Previously **sole developer** of an enterprise multi-agent AI platform on DGX-class hardware, taken from PoC to production at **two enterprise customers**. Active upstream — **17 merged/landed patches** across FlashInfer, vLLM, SGLang, PyTorch, Dynamo, CUTLASS, TensorRT-LLM and LMDeploy, concentrated on early consumer-Blackwell (**SM120**) and **NVFP4** enablement. Research background in AI/LLM security and privacy-preserving ML.

Technical Skills

LLM Serving & GPU	vLLM, TensorRT-LLM, SGLang, Triton, NVIDIA NIM, Dynamo, FlashInfer, NCCL, Flash Attention, KV-cache, Quantization (FP8 / INT4 / NVFP4); CUDA C/C++, WMMA & `mma.sync` Tensor Cores, CUTLASS, Nsight Systems & Compute
LLM Apps & Agents	Multi-Agent Orchestration (LangChain, LangGraph), Hybrid RAG (Vector + Keyword + Knowledge Graph), LLM Evaluation & Guardrails, Function Calling / Tool Use, MCP, Coding Agents
Infrastructure & Ops	Linux, Kubernetes, Docker, On-prem GPU Serving, Distributed Tracing; H100, DGX Spark (GB10), Blackwell workstation hardware
Privacy & Security ML	Federated Learning, Differential Privacy, Secure Multi-Party Computation (MP-SPDZ), Safety Alignment
Languages	Python, TypeScript, Rust, C/C++; PyTorch, FastAPI, NestJS; PostgreSQL, Milvus, Redis

Experience

Taiwan AILabs — Jul 2026 – Present
Field Application Engineer — *Taipei, Taiwan*

- Deploy on-premises LLM systems (**FedGPT**) into customer environments; translate site constraints into Linux / Kubernetes / GPU-serving plans, preflight checks, and acceptance criteria
- Troubleshoot cross-layer failures spanning containers, GPU runtime, storage, networking, certificates, and model-serving endpoints; explain performance, reliability, and operational trade-offs through handover

SYNCROBOTIC — Sep 2025 – Jun 2026
Machine Learning Engineer — Multi-Agent AI, RAG, LLM Serving — *New Taipei City, Taiwan*

- **Sole developer** of an enterprise multi-agent AI platform on DGX-class hardware — owned end-to-end from PoC through production and rolled out at **two enterprise customers**, partnering with each through deployment and troubleshooting
- Scoped ambiguous enterprise requirements into a working multi-agent system and **unified local and commercial LLM backends** under one routing layer, removing per-team model wrappers
- Designed a planning–execution–validation orchestration with topological scheduling and parallel agent dispatch for multi-step workflows
- Built a hybrid retrieval layer on LightRAG (vector + keyword + knowledge graph) with embedding-based semantic caching, cutting redundant LLM calls and keeping answers consistent on repeat queries
- Owned production reliability — prompt/pipeline regression tests, output validation across quality and safety, distributed tracing; added CUDA-accelerated hot paths on critical inference

Advantech — Jun 2025 – Aug 2025
AI Agent Engineer (Intern) — *Taipei, Taiwan*

- Built coding agents for internal developer tooling using **function calling / tool use** — automating code review, test generation, and static analysis across the SDLC; benchmarked tool-use reliability across commercial and local LLM backends and designed reusable prompt templates adopted by the team

TSMC (Taiwan Semiconductor Manufacturing Co.) — May 2021 – Oct 2022
R&D Alternative Military Service — Facility Engineer — *Tainan, Taiwan*

- Supported R&D fab facility operations at the world's leading semiconductor foundry under the national R&D alternative military service program

Open Source

SM120 / NVFP4 enablement across the LLM-inference stack — 17 merged/landed
FlashInfer · vLLM · SGLang · PyTorch · Dynamo · CUTLASS · TensorRT-LLM · LMDeploy — prs.wayne.is-a.dev

- Kernel-level support for consumer Blackwell (**SM120**) and **NVFP4** carried upward through kernels → engines → disaggregated serving
- Focus on *silent-correctness* bugs — the ones where tests pass and the answers are still wrong

Research & Publications

SelGrad: Selective Gradient Projection for Efficient Safety Alignment Against Harmful Fine-Tuning — first author, with Shao-Jui Wang. *IEEE Transactions on Dependable and Secure Computing (TDSC)*, under review, 2026. Preserves LLM safety alignment under harmful/adversarial fine-tuning at lower compute than full re-alignment.

Research at NTUST (advised by Prof. Shao-Jui Wang) in **AI/LLM security** and **privacy-enhancing ML** — federated learning, differential privacy, and secure multi-party computation.

Education

National Taiwan University of Science and Technology (NTUST) Aug 2024 – Apr 2026
M.S. in Computer Science and Information Engineering — GPA 4.09 *Taipei, Taiwan*

National Taiwan University (NTU) Sep 2014 – Jun 2020
M.S. in Structural Engineering (2020) | B.S. in Civil Engineering (2018) *Taipei, Taiwan*

Certifications & Awards

NVIDIA DLI — Accelerated Computing with CUDA C/C++ and CUDA Python (5 certificates) · NICS Privacy-Enhancing Technologies (PETs) Competition — Rank 3/11 (DP synthetic data via MP-SPDZ) · International ICT Innovative Services Awards, 2024