Sheng-Tse Lin

+1 012-345-6789 · hunterl@andrew.cmu.edu · sheng-tse.github.io
LinkedIn · GitHub · Google Scholar

Education

Carnegie Mellon University *Aug 2025 – Dec 2026*
M.S. in Information Systems Management *Pittsburgh, PA*

- **Relevant Coursework:** Introduction to Deep Learning (PhD), Distributed Systems, Generative AI, LLM Systems, AI Agents
- **Honors & Awards:** **3rd/238** Automatic Speech Recognition, **5th/266** Face Verification

Fu Jen Catholic University *Sep 2019 – Jun 2022*
B.B.A. in Finance and International Business *New Taipei City, Taiwan*

Skills

Programming	Python, C++, CUDA, Rust, Java, SQL
Frameworks	PyTorch, JAX, Flax, Optax, Hugging Face, PEFT/LoRA, DPO/GRPO, DeepSpeed, LangGraph
Tools	vLLM, SGLang, FastAPI, Docker, Kubernetes, Slurm, AWS, GCP, W&B, Git

Work Experience

Forward Deployed Engineer *Aug 2026 – Present*
Smile Line AI *San Francisco, CA*

- Scoped the AI engine for Smile Line's flagship Smile Lite MDP app with the founding team, from model and data choices to EU and US privacy rules, ahead of deploying it to the company's device and app
- Benchmarked detection models and quantized the chosen **YOLO** detector to run chairside on hardware the company already ships, so it checks whether each photo is framed correctly while the clinician shoots

Teaching Assistant, 11-785 Introduction to Deep Learning *Apr 2026 – Present*
Carnegie Mellon University, Language Technologies Institute *Pittsburgh, PA*

- Ported course homeworks to JAX, Flax, and Optax for TPU training with Google, for **300+** students, and authored an **LLM** pretraining assignment to build a Llama-style model from scratch with RoPE, SwiGLU, and RMSNorm

Research Intern *Jul 2026 – Aug 2026*
Carnegie Mellon University *Pittsburgh, PA*

- Designed a fail-closed policy engine that keeps a governed **LLM** agent from answering off-corpus and an evidence gate that withheld **296** of **300** claims under speaker overlap with full coverage on clean audio

Machine Learning Engineer *Jun 2024 – Jun 2025*
Future Educators *Washington, D.C.*

- Built an **OpenAI**-backed **RAG** advising assistant over **1.2M+** GPT-labeled federal charter school records and deployed it on AWS as a FastAPI service at p95 under 2 s for **300+** daily sessions, **47%** faster than manual lookup
- Shipped an **LLM** document-processing pipeline on Claude 3.5 Sonnet that reached **93%** extraction accuracy on a human-labeled eval set and cut manual review time by **70%**

Machine Learning Engineer *Jun 2023 – Jun 2024*
The Heroes Garage *Boston, MA*

- Reduced bid error against manual appraisal by **37%** with a gradient-boosted model over Carfax history and past sale prices that set the buying team's live bids at Manheim and ACV auctions
- Deployed a repair cost model trained on past repair orders that cut dealer estimate error (MAPE) from **22%** to **7%**

Publications

AcoustiClaim: A Numeric Claim Benchmark with Instrument Ground Truth [Paper] [Code]
Sheng-Tse Lin, Siyuan Zhai*, Chien-Liang Kuo*, Massa Baali, Bhiksha Raj *Under review, ICASSP 2027*

Open Source

Argonauts Code: One Orchestrator for Every Coding Agent [Code] *Oct 2026 – Present*
Creator & Maintainer *1k+ Downloads*

- Wrote a Rust terminal app where coding agents such as Claude Code and Codex build in parallel and review each other's work

jupynvim: Jupyter Notebooks and Remote Kernels in Neovim [Code] *May 2026 – Present*
Creator & Maintainer *180+ Stars · 3k+ Downloads*

- Brought Jupyter notebooks natively into the terminal and over SSH via a Rust backend on ZeroMQ and a Lua UI

FlashInfer: Kernel Library for LLM Serving [PR] *Oct 2026 – Oct 2026*
Contributor *6.8k+ Stars*

- Fixed FlashInfer's Mamba selective state update CUDA kernels so head dims that are not multiples of 64 compile and run correctly