Kenneth Wang Yau LI

San Francisco Bay Area, CA | +1 510-938-9438 | liwangyau@gmail.com | https://kennethli319.github.io
Applied AI / Speech LLM Engineer | ASR/TTS Evaluation, LLM-as-Judge, Speech Agents | TN Visa eligible

PROFILE

Speech-focused ML engineer with experience building production-oriented ASR/TTS systems, speech LLM evaluation pipelines, and automated model-quality workflows. Strong background across streaming ASR, TTS, full-duplex speech-to-speech systems, LLM-as-judge evaluation, RAG, keyword/hotword boosting, and neural speech interfaces. Experienced in improving model quality, reducing evaluation cost and latency, and accelerating iteration across startup and large-scale production environments.

ROLE TARGET

Applied AI / Speech LLM / Audio ML roles focused on model training, evaluation systems, ASR/TTS workflows, LLM-as-judge systems, speech agents, multimodal model quality, and production data pipelines.

WORK EXPERIENCE

Jan 2025 – Present

Linguistic Engineer (Speech LLM / ASR-TTS Evaluation) @ Meta (Reality Labs Wearables)

LLM-as-Judge | Full-duplex Speech-to-Speech LLM | Agentic Feedback Loops | ASR/TTS Data Preparation

• Improved on-device and server-side ASR/TTS for voice assistants across wearable devices and Meta AI by leading data preparation and evaluation workflows for expressive speech and Llama 4 full-duplex speech LLMs.
• Designed metrics for multi-turn full-duplex speech LLMs across 31 locales, improving cost efficiency, quality tracking, iteration speed, and production readiness.
• Led the transition from manual human evaluation to automated LLM-as-judge workflows, reducing evaluation cost by 10x with private models / 1000x with open-source models and turnaround time by ~100x, from days to under 1 hour.
• Benchmarked internal models and leading external LLMs across multiple dimensions of audio-modality performance.
• Deployed chat-based agentic workflows to automate routine research and engineering support tasks, reduce repeated support requests, and enable self-serve access.
• Prototyped evaluation-driven feedback loops for autonomous agent self-improvement and workflow optimization.

Sep 2023 – Jan 2025

Founder & CEO @ InnerSpeech Canada / Hong Kong (Pre-seed, Brain-Computer Interface)

Founder Experience | EEG-to-Text | Large Scale Multi-modal Data Preparation

• Founded a non-invasive BCI speech startup focused on imagined/inner speech decoding using EEG and multimodal biosignals, with assistive communication as the initial target use case.
• Owned product framing, technical roadmap, partnership conversations, grant/startup programs and prototype development.
• Built and open-sourced the InnerSpeech biosignal speech recognition and synthesis toolkit covering EEG, EMG, HD-EEG, MEG, fNIRS, and invasive neural speech datasets.
• Developed Brain-to-Text Benchmark 2024 systems using RNN-Transformer modeling with language-model rescoring.
• Secured founder/startup support including HKSTP Ideation, HK Tech 300 seed approval, NVIDIA Inception, AWS Activate, Google for Startups, Microsoft Founder Hub, and Communitech Founder Program.

Apr 2024 – July 2024

Machine Learning Engineer @ Kea Cloud Inc. (Series A, Restaurant Voice AI)

ASR/TTS via External APIs | Keyword Boosting | Voice AI Pipeline with RAG

• Improved restaurant-ordering ASR through API integration, contextual biasing, and menu/entity recognition evaluation.
• Built LLM-augmented ordering-agent workflows using RAG-style ASR-to-NLU pipelines.
• Evaluated ASR boosting and keyword strategies for restaurant-specific entities, menu items, and conversational order flows.

Feb 2020 – Aug 2023

Speech Recognition Engineer III @ Dialpad Canada Inc. (Series F, Contact Center Voice AI)

In-house ASR Model Fine-tuning | ASR Data Augmentation | Contextual Biasing | WFST with Kaldi/NeMo/K2

• Built and improved production ASR systems using Kaldi, NeMo, and K2 across hybrid and end-to-end speech recognition stacks.
• Led ASR model updates and evaluation cycles, including analysis, data preparation, and deployment-oriented model comparisons.
• Developed and shipped contextual biasing methods, including n-gram and lattice boosting, to improve recognition of customer-specific vocabulary and recover failing business-call scenarios.
• Published and presented work on ASR boosting and G2P modeling through SANE 2022 and SIGMORPHON 2021.
• Co-supervised Master's thesis at University of Edinburgh and The University of British Columbia

Sep 2020 – May 2023 (part-time)

NLP Consultant @ The Hong Kong Polytechnic University

RAG-LLM workflow | Model Quantization | In-house Model Fine-tuning | Architect & Roadmap

• Advised on grammar error correction and RAG-style workflows, including retrieval-augmented model design and evaluation.
• Integrated rule-based and neural network-based methodologies to enhance system efficiency and performance.
• Investigated and implemented quantization, data augmentation, and model optimization to further elevate the system's capabilities.

EDUCATION

2018 – 2019 **Master of Science in Speech & Language Processing @ University of Edinburgh**

Thesis: Robust Word Recognition and Alignment of Child Speech Therapy Sessions using Audio and Ultrasound Imaging (PyTorch and Kaldi)

Coursework: Speech Synthesis (TTS), Automatic Speech Recognition, Natural Language Understanding, Generation, and Machine Translation, Reinforcement Learning, Neural Information Processing

- The Edinburgh Award (Enterprise)
- First place winner of the Business Ideas Competition by The University of Edinburgh
- Winner of Scottish Institute for Enterprise's Fresh Ideas Competition

2014 – 2018 **Bachelor of Arts in Linguistics and Language Applications (First Class Honours) @ City University of Hong Kong**
Major in Linguistics and Language Applications and Minor in Translations

Thesis: A Comparative Study of Interlingual vs. Neural Approach to Machine Translation of Numerical Expressions (with Tensorflow and Java)

PROJECTS & CERTIFICATIONS

2026 – Present **Open Audio Judge: Omni-LLM Evaluation for Voice-AI Systems** — Open-source contributions

- Built an open-source LLM-as-judge evaluation and monitoring toolkit for Voice-AI systems, using omni-model APIs and self-hosted models to assess ASR, TTS, and speech-agent outputs.

2025 – Present **Audio ML Course: Speech ML Systems Curriculum** — Open-source contributions

- Created and maintained an open-source course summarizing the development of speech ML, from audio representations and ASR/TTS foundations to modern speech agents, evaluation, serving, safety, and production reliability.

2023 – 2024 **NLP Consultant - UsherGPT** — University of Edinburgh, Usher Institute

- Guided development of UsherGPT, tailored for public health and medical data applications using Retrieval Augmented Generation (RAG) techniques.

2021 **Third Prize in MUCS 2021** — MUltilingual and Code-Switching ASR Challenges for Low Resource Indian Languages (MUCS)

- Team contributions to multilingual and low-resource ASR for Indian Languages. Benchmarking and open-sourcing various end-to-end methods

CORE SKILLS

Applied AI & Systems	Python, PyTorch, Docker, data pipelines, production debugging, evaluation automation, LLM-as-judge, internal tooling, LLM/RAG workflows, agentic development tooling
Speech ML	ASR, TTS, speech data preparation, ASR/TTS evaluation, hotword/keyword boosting, WER/MOS/CMOS analysis, streaming ASR, speech-to-speech workflows
Modeling & Frameworks	Kaldi, NVIDIA NeMo, K2, CTC, Transducer/RNN-T, Conformer, Thinker-Talker, n-gram rescoring
Research Areas	Speech recognition, speech synthesis, NLP, Non-invasive BCI, EEG/fNIRS speech decoding, multimodal speech interfaces

PUBLICATIONS

N-gram Boosting: Improving Contextual Biasing with Normalized N-gram Targets · *Poster at SANE 2022*

***Li, W. Y.**, Nadig, S., Chang, K., Mahmood, Z., Wang, R., Vandieken, S., Robertson, J. & Mailhot, F. (2023). N-gram Boosting: Improving Contextual Biasing with Normalized N-gram Targets. arXiv preprint arXiv:2308.02092.*

Avengers, Ensemble! Benefits of ensembling in grapheme-to-phoneme prediction · *Paper at 18th SIGMORPHON Workshop 2021*

*Gautam, V., **Li, W. Y.**, Mahmood, Z., Mailhot, F., Nadig, S., Wang, R., & Zhang, N. (2021, August). Avengers, Ensemble! Benefits of ensembling in grapheme-to-phoneme prediction. In Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology (pp. 141-147).*

Can linguistics help neural machine translation? - Evidence from a case study of interlingual vs neural machine translation of numerical expressions · *Presentation at AI and Linguistics Conference 2018 - Dr. Chunyu Kit (CityU) and **Wang Yau Li** (Edinburgh)*