




廖介任(Rick)

- LLM Application Engineer
- Deep Learning Engineer
- Speech Processing Engineer

 Zhonghe District, New Taipei City,
Taiwan

- Experienced LLM Finetune and Prompt Engineering. Expertise in LLM Application Developments and Integrated WebUI Design, Especially in whisper and code generation.
- Skilled End2End ASR Developer: 7 Years STT Engine Building with Kaldi and End-To-End(Deep Learning) Solutions.
- Skilled at Deep Learning Models Pruning and Quantization

- email: xrickliao@gmail.com

- phone: +886 958846585

- github: <https://www.github.com/xrick>

- cake resume: <https://www.cake.me/me/rick-liao-f2b2c3>

- linkedin: <https://www.linkedin.com/in/rick-liao/>

Work Experience



AI-LLM Engineer • 新加坡楓葉資訊科技

二月 2025 - Present

- **Responsibility**
 - LLM RAG Web Apps Development.
 - LLM Fine-tuning.
 - Dsign of Aiagent + n8n automation workflow software.
- **Projects**
 - Responsible for fine-tuning DeepSeek-R1:7b with company internal data to increase the search accuracy about 2% using LoRA.
 - DQE RAG Application: Increasing the design fault finding efficiency about 10%~15%.
github: https://github.com/xrick/DQE_RAG_APP
 - SalesRAG: Increasing the efficiency of searching products information about 30%~40%.
github: <https://github.com/xrick/SalesRAG>
- My expertise in deep learning model compression and optimization is transferable to edge devices and quality control within semiconductor production environments.
- Experiences such as Design Quality Estimation RAG can be applied to data-driven decision-making and prediction, directly enhancing semiconductor process optimization.
- My LLM skills in document analysis and RAG address manufacturing knowledge management and support semiconductor design optimization.



Senior AI工程師 • 聯億通股份有限公司

九月 2023 - 九月 2024

- **Responsibility**
 - On-Chips Deep-Learning Models Development.
 - Deep Learning Model Compression.
 - RAG Applications Development.
- **Projects**

- Fire and Smoke Alarm Detection Module for Smart Sockets:
github: <https://github.com/xrick/uec-ai-dev>
 - An audio event classification model on extremely resource-constrained environments.
 - Achieves an accuracy of 96.8%.
 - Compressed Size: from 18.9MB down to under 100KB.
 - Developed a customized AutoML system using the Microsoft FLAML framework.
 - Speeds up model training and adapts to various customer requirements.
- Online Customer Q&A Web Application:
 - web application for online customer service.
 - Increase customers' understanding of products and provides answers to their questions.
 - Utilizes LLMs to analyze customers' questions.
- High-accuracy fire and smoke detection experience meets semiconductor manufacturing's need for efficient anomaly detection in resource-limited settings.



Digital Speech and Audio Algorithm Engineer • 台灣歌爾泰克股份有限公司

十二月 2019 - 三月 2022

Data Analysis Web Application Development:

- Developed a web application for voice spectrum analysis.

Edge Device Deep Learning Model Development:

- Developed a deep learning voice command model on chips, achieving accuracy of 96% and compression rate of over 80%.
- Developed an anomaly detection model for fan sound abnormalities, achieving an accuracy rate of 95%.



NLP and Deep Learning Technology Assistant Manager • 威剛-隆辰星

十二月 2017 - 十二月 2018

Web Services Development:

- Developed a conversational engine Restful Web API using Python.
- Built a distributed crawler to collect textual data.

Chatbot Engine Construction:

- Developed a Rule-Based Chatbot Engine with machine learning.
- Established a natural language processing pipeline to accelerate corpus processing and debugging.
- Building a Kaldi-Based ASR (Automatic Speech Recognition) Engine.

Deep Learning Team Building:

- Established a deep learning algorithm team.
- Planned the team's short-term development roadmap for 1 to 3 years.
- Experience in leading deep learning teams and managing projects benefits collaborative, multidisciplinary work in advanced manufacturing.

Education



2018 - 2027



2003 - 2006

Ph.D. Candidate.
Computer Science and Information
Engineering.
National Taiwan Normal University

Research Areas:

- Application of Large Language Models (LLMs) in speech and text processing
- Pronunciation assessment
- Automated speech and text scoring
- Speech enhancement
- Few-shot deep learning

Master’s Degree in Computer Science,
National Taiwan Normal University

Research Areas:

- Meta-Heuristic Optimization Algorithms.
- Software Design Patterns and Architectures.

Skills

- Deep Learning Algorithms
- Deep Learning Models Pruning and Quantization
- LLM RAG
- LLM fine-tuning using LoRA
- LangChain and LangGraph
- PostgreSQL
- NoSQL: Redis, MongoDB
- Milvus vector database
- Docker and Kubernete

Language

- English - Advanced
- TOEFL : 90
- GRE : 301