Yen-Hsiang Wang

📞 +886 928-666-218 ✉ heliart_sho@hotmail.com 🔗 https://heliart.me/ ⭘ github.com/kenny75557

Summary

NLP researcher specializing in LLM-powered retrieval, query expansion, and low-resource language modeling. Experienced in both academic research and industrial AI applications, with hands-on experience in deploying LLM-based search and retrieval systems, as well as fine-tuning LLMs using parameter-efficient techniques such as LoRA to adapt models for various industry-specific tasks.

Currently passionate about improving the robustness of LLM reasoning models and developing AI agents with enhanced reliability and interpretability.

Publications

Conference/Journal Papers   († indicates equal contribution.)

1 **Yen-Hsiang Wang**, Feng-Dian Su, Tzu-Yu Yeh, Yao-Chung Fan. A Cross-Lingual Statutory Article Retrieval Dataset for Taiwan Legal Studies *In ROCLING 2024. (Best Paper Nominee)*

2 You-Cheng Liao, Chen-Jui Yu, Chi-Yi Lin, He-Feng Yun, **Yen-Hsiang Wang**, Hsiao-Min Li, Yao-Chung Fan. Learning-From-Mistakes Prompting for Indigenous Language Translation *In ACL LoResMT 2024.*

3 Xingdi Yuan, Tong Wang, **Yen-Hsiang Wang**†, Emery Fine, Rania Abdelghani, Pauline Lucas, Hélène Sauzéon, Pierre-Yves Oudeyer. Selecting better samples from pre-trained LLMs: a case study on question generation *In Findings of ACL 2023.*

4 Rania Abdelghani, **Yen-Hsiang Wang**, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, Pierre-Yves Oudeyer. GPT-3-driven pedagogical agents to train children's curious question-asking skills *In International Journal of Artificial Intelligence in Education.*

Tutorial

- ROCLING 2024 "Mastering Retrieval-Augmented Generation: Unlocking Advanced Retrieval Techniques," [Website and Slide]

Selected Project

PTT Social Analyze Engine | *Scikit-learn, ElasticSearch, FastAPI, Docker*   **Sep. 2024 – Dec. 2024**

- **Social Opinion Clustering:** Applied the K-means algorithm to cluster topics within a specified time frame on PTT (web forum) and employed LLMs to generate concise summaries and trend-based headlines.[Demo]
- **PTT Social Analyze System:** Utilized BGE-M3 as a text embedding model to retrieve similar articles based on user queries and applied In-Context Learning with LLMs to produce analysis of sentiment and stance.[Demo]
- [Github Link]

PTT Search Engine | *Vue.js, Flask, ElasticSearch, Docker*   **Oct. 2020 – Jan. 2021**

- Built a search engine including all articles and commands crawled from popular boards on PTT.
- Implemented BM25 with Elastic Search for retrieving comments from the database and processed the data using Object-Oriented Programming techniques.
- [Github Link] / [Site]

Work Experience

National Chung Hsing University Natural Language Processing Lab   **Jan. 2024 – Dec. 2024**
Research Assistant   *Taichung, Taiwan*

- Exploring how LLMs can enhance translation for low-resource or unseen languages by leveraging Chain-of-Thought reasoning and providing relevant patterns through sentence retrieval. [LoResMT ACL'24]
- Proposed a scalable method for generating synthetic legal QA dataset and validating the performance of Sparse embedding, Dense embedding, and LLM query expansion in legal domain retrieval. [Dataset][Paper]
- Building a Retrieval Augmentation Generation application for agricultural domain question-answering, focusing on retrieval and re-ranking methods to enhance accuracy. [Shennong-TAIDE].

EmotiBot   **Jun. 2023 – Dec. 2023**
Machine Learning Engineer   *Taipei, Taiwan*

- Implemented LLM-powered intelligent customer support systems for Apple and KGI Securities, optimizing RAG-based retrieval models to reduce response latency and improve user satisfaction.

Inria, National Institute for Research in Digital Science and Technology   **Mar. 2022 – Sep. 2022**
Research Engineer Intern   *France*
- Developed an LLM-driven educational framework to generate curiosity-driven questions for elementary students, reducing manual effort and enhancing the quality of educational materials. [IJAIED]
- Researching on Question Generation Sampling and Evaluation of large language models. [ACL'23]

Industrial Technology Research Institute   **Jul. 2021 – Aug. 2021**
Software Engineer Intern   *Hsinchu, Taiwan*
- Fine-tuned a Transformer-based model for Legal Judgement Question-Answering, enhancing legal text comprehension and improving answer accuracy.
- Developed a full-stack web-based Manual Annotation System for Data Labeling, streamlining the annotation process and improving efficiency.

Education

National Chung Hsing University   **Sep. 2020 – Sep. 2022**
M.S. in Computer Science. Advisor: Yao-Chung, Fan   *Taichung City, Taiwan*

Fu Jen Catholic University   **Sep. 2016 – Jun. 2020**
B.S. Computer Science and Information Engineering.   *New Taipei City, Taiwan*

Awards

Best Paper Nominee | *ACLCLP*   **Nov. 2024**
- Construct a Cross-Lingual Legal Retrieval Dataset for AI/NLP Research and Provide Multiple Effective Baselines.

Skills

Programming Languages
- C/C^{++}, **Python**, JavaScript

Tools
- Frameworks: React, Flask, FastAPI
- Databases: MySQL, **Elasticsearch**, MongoDB
- ML & Big Data: PyTorch, Spark, Hadoop
- NLP: Hugging Face Transformers, PEFT(e.g., LoRA), TRL (Transformers Reinforcement Learning), TGI
- LLM Toolkits: LangChain, LlamaIndex, Ollama

Language Proficiency

Chinese | Native

English | TOEFL 80
- Reading 21
- Listening 21
- Speaking 19
- Writing 19