WEERAYUT BUAPHET

• weerayut.b_s20@vistec.ac.th • linkedin.com/in/weerayutbu • weerayutbu.github.io

SUMMARY

I'm a five-year Ph.D. scholarship student in the Natural Language Processing and Representation Learning Lab (NRL) at VISTEC, Thailand, supervised by Associate Professor Prof. Dr. Sarana.

PhD Thesis: Resource-Constrained Named Entity Recognition — Contributed a Thai-language dataset for fine-grained nested NER and a bilingual financial NER dataset for the stock market; analyzed the generalization of encoder-based and LLM-based NER models to unseen entity types and new domains; and addressed multilingual text normalization challenges in informal language.

Currently, I am focused on developing and evaluating an LLM-based retrieval-augmented generation (RAG) system for the medical domain, using supervised fine-tuning (SFT) and reinforcement learning methods (e.g., DPO, GRPO) to train models for multi-turn medical question answering (ThaiLLM).

EDUCATION

Ph.D. in Information Science and Technology Aug 2020 - Present
Vidyasirimedhi Institute of Science and Technology (VISTEC) 4.00/4.00 GPA
Relevant coursework: Natural Language Processing, Computational Machine Intelligence and Applications

B.Eng. in Computer Engineering Mar 2016 - Mar 2020
Rajamangala University of Technology Lanna (RMUTL-CM) 3.62/4.00 GPA (Top 1)
Relevant coursework: Data Structures and Algorithms, Operating Systems, Software Engineering

INTERNSHIP

ITU, Copenhagen, Denmark: PhD Internship Sep 2024 – Jul 2025

- Co-organizing WNUT 2025.
- Conducting research on few-shot Named Entity Recognition (FS-NER).
- Conducting research on multi-lexical normalization.

VISTEC, Rayong, Thailand: Researcher Assistant Nov 2019 – Aug 2020

- Conducted literature reviews and implemented a baseline Thai Nested NER model using Python and Torch.
- Performed quality control and error analysis to ensure the accuracy and reliability of the Thai N-NER model.

TECHNICAL SKILLS

- **Languages:** Thai (Native), English (TOEFL ITP 550)
- **Tools:** Python, PyTorch, Docker, SQL, C++, HTML, PHP, JavaScript
- **Learning:** LangChain, LangGraph, The Model Context Protocol (MCP)

ACADEMIC PROJECTS

- **Thai Nested Named Entity Recognition Corpus (Thai N-NER).** Aug 2019 – May 2022
  Weerayut Buaphet, Can Udomcharoenchaikit, Peerat Limkonchotiwat, Attapol Rutherford, and Sarana Nutanong. 2022. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1473–1486, Dublin, Ireland. Association for Computational Linguistics.
- **LLM-Augmented Prototype Representation for Few-Shot Named-Entity Recognition** Nov, 2025
  Weerayut Buaphet, Peerat Limkonchotiwat, Attapol Rutherford, Can Udomcharoenchaikit, and Sarana Nutanong. 2025. IEEE Access.
- **MultiLexNorm++: A Unified Benchmark and a Generative Model for Lexical Normalization for Asian Languages** Accepted to TALLIP, in production
  Weerayut Buaphet, Thanh-Nhi Nguyen, Risa Kondo, Tomoyuki Kajiwara, ..., Rob van der Goot. 2025. ACM Transactions on Asian and Low-Resource Language Information Processing.

- **Bi-lingual Named Entity Recognition for financing (Fin-NER).** Ongoing project
  Creating a finance-NER dataset composed of two languages, Thai and English. This project aims to study the knowledge transfer from high-resource to low-resource languages in the financial domain.
- **Mitigating Spurious Correlation in Natural Language Understanding with Counterfactual Inference.** Dec 2022
  Can Udomcharoenchaikit, Wuttikorn Ponwitayarat, Patomporn Payoungkhamdee, Kanruethai Masuk, Weerayut Buaphet, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11308–11321, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- **Cross-Lingual Data Augmentation For Thai Question-Answering.** Dec 2023
  Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2023. In Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, pages 193–203, Singapore. Association for Computational Linguistics.

CONTRIBUTIONS AND ACTIVITIES

- **Reviewer**
  ARR-EMNLP 2024, 2025
- **Mentor AI Builders (2022)**
  This program aims to develop knowledge in Data Science and Artificial Intelligence (AI) for middle and high school students interested in practical applications. My friend and I are responsible for five students working on various projects, such as question generation, fake news detection, and creating a dataset and system to support a plant tissue laboratory.
- **Internet of Things (11th 9 RMUT competition 2019) - RMUTSB**
  We got 2nd place in the RMUT group IoT competition in Thailand. We used an ESP20 to read sensor data and send it via MQTT to a Raspberry Pi server, which visualized the data on a web interface. I programmed the visualization and configured the ESP20 while my friend handled the hardware connections.
- **The Robotic Design Contest (RDC 2018)**
  This program selects national representatives for the International Design Contest RoBoCon (IDC RoBoCon). All teams are required to design a robot to solve a provided problem. It promotes equality with mixed teams, equal resources, and collaboration, including training sessions at all levels. My friend and I achieved the following:
  - 2nd place in the Northern region at Chiang Mai University.
  - 1st place in Thailand at Chulalongkorn University.
  - 3rd place internationally at the Tokyo Institute of Technology.