# WEERAYUT BUAPHET

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

## SUMMARY

I am a Ph.D. student in the Natural Language Processing and Representation Learning Lab (NRL) at VISTEC, Thailand, supervised by Assoc. Prof. Sarana Nutanong and co-advised by Assoc. Prof. Attapol Rutherford.

**Ph.D. Thesis: Resource-Constrained Named Entity Recognition.** My research focuses on developing NER systems under limited-resource settings. I contributed a Thai fine-grained nested NER dataset and a bilingual (Thai–English) financial NER dataset, analyzed the generalization of encoder-based and large language model (LLM)-based NER methods to unseen entity types and domains, and studied the robustness of multilingual text normalization in informal and noisy text.

Currently working on LLM-based retrieval-augmented generation (RAG) systems for the medical domain, including RAG pipeline design, evaluation, and fine-tuning using supervised fine-tuning (SFT) and preference optimization methods (DPO, GRPO) for multi-turn medical question answering as part of the **ThaiLLM** project.

## EDUCATION

**Ph.D. in Information Science and Technology (5-year program, fully funded)** 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, Chiang Mai. 3.62/4.00 GPA (Ranked 1st out of 116)
Relevant coursework: Data Structures and Algorithms, Operating Systems, Software Engineering

## WORK EXPERIENCE & INTERNSHIPS

**ITU, Copenhagen, Denmark: PhD Internship (Supervisor: Assoc. Prof. Rob van der Goot)** Sep 2024 – Jul 2025

- Co-organized **The 10th Workshop on Noisy and User-generated Text (W-NUT)**, collocated with NAACL 2025.
- Conducted research on few-shot Named Entity Recognition, evaluating encoder-based and LLM-based models and analyzing their trade-offs in few-shot settings.
- Conducted research on multi-lexical normalization for Asian languages, collaborating with five international teams from Japan, Korea, Vietnam, Indonesia, and Thailand to create a multilingual Asian text normalization benchmark and introduce LLM-based text normalization methods.

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

- Conducted literature reviews and implemented a baseline Thai nested NER model.
- 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, LangChain, FastAPI, Streamlit, C++, HTML, PHP, JavaScript

## ACADEMIC PROJECTS

- **Thai Nested Named Entity Recognition Corpus (Thai N-NER).** 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.
- **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.
- **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.
- **Bilingual Named Entity Recognition for Finance (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.

## CONTRIBUTIONS AND ACTIVITIES

- **Reviewer**
  ARR-EMNLP 2024, 2025; ACL 2026
- **Mentor AI Builders (2022)**
  My colleague and I mentored five students—one per project—on question generation, fake news detection, and dataset and system development for a plant tissue laboratory. One project earned the Best Presentation Award.
- **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 teammate 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 teammate 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.