Dr. Peerat Limkonchotiwat

LinkedIn: https://www.linkedin.com/in/peerat-limkonchotiwat/ Github: https://github.com/mrpeerat
Email: peerat.limkonchotiwat@gmail.com Website: https://mrpeerat.github.io/ Tel. +66 90620262
Publications: https://scholar.google.com/citations?user=T-rvPZ4AAAAJ&hl=th

I am an AI engineer in AI Singapore (AISG) working on large language model fields, such as training algorithms, data attribution, and evaluation and benchmarking, to improve Southeast Asian LLMs. Before joining AISG, I was a Ph.D. student in information science and technology (IST) at VISTEC, Thailand. My PhD research focused on large language models, dense retrievals, semantic understanding, multilingual learning, question answering, and entity linking. Moreover, I am interested in applying research in real-world scenarios such as information retrieval (search and rank), question-answering, and chatbot systems.

EDUCATION

Vidyasirimedhi Institute of Science and Technology (VISTEC) **Rayong, TH**
Ph.D. in Information Science and Technology (5 years program) *Aug 2019–July 2024*
GPA: 3.84/4.00

- Worked under Assoc. Prof. Dr. Sarana Nutanong and Dr. Ekapol Chuangsuwanich at the Natural Language and Representation Learning Lab (NRL). Full scholarship: monthly stipend, annual research grant, and conference travel grant.
- Developed domain adaptation and out-of-domain handling techniques for word segmentation in 4 languages: Thai, Chinese, Japanese, and Urdu. [EMNLP'21, ACL'21]
- Developed a monolingual and multilingual sentence embedding based on a pre-trained language model for semantic textual similarity and text mining tasks. [EMNLP'21-'22, TACL'23, ACL'24]
- Developed a novel monolingual and multilingual text embedding for a retrieval question-answering and web-search framework. [NAACL'22, ACL'23, EMNLP'24]
- Supervised 8 graduated students and 2 research assistants. Their works were published in ACL (NER and entity linking) and EMNLP (knowledge distillation for LLMs).

Rajamangala University of Technology Lanna Chiang Mai (RMUTL) **Chiang Mai, TH**
Bachelor of Science in Computer Engineering *Sept 2015–Mar 2019*
GPA: 3.25/4.0 (Class rank: 1st)

RESEARCH AND TECHNICAL EXPERIENCES

AI Engineer **Singapore, SG**
AI Singapore (AISG) *September 2024–Present*

- Developing a Southeast Asian large language model called SEA-LION. My main responsibilities are pre-training data, fine-tuning algorithms, and evaluation and benchmarking. Launched models: SEA-LIONv3, WangchanLIONv2.
- Thai main contributor. I helped my team develop data preprocessing for Thai, such as removing PII, data cleaning, data filtering, data collection, and misspellings.
- Formulating cultural and safety alignment data for Southeast Asian countries. The dataset aims to detect harmful and inappropriate texts in SEA languages.
- As an individual contributor, I do research with many external collaborators to bring knowledge and experience about developing LLMs, benchmarks, and data to the team. I collaborate with Google, Sony, Thoughtworks, and Cohere, where my works have been published at top NLP conferences.
- Working with Prof. Trevor Cohn and Prof. Partha Talukdar from Google Deepmind to help Southeast Asian NLP not to be low-resource languages through a program called SEALD. My main responsibilities are data collection (SFT and alignment data) for Thai and designing an algorithm for SEA LLMs.

Invited Researcher **Bangkok, TH**
Faculty of Arts, Chulalongkorn University *January 2025– Present*

- Research funded by Chulalongkorn University (150,000 Baht)
- Publish research papers at top NLP conferences as a main author.
- Work closely with Chulalongkorn Professor (Prof. Ekapol Chuangsuwanich and Prof. Attapol Thamrongrattanarit) about NLP research topics.

Applied Science Internship **Cambridge, GBR**
Alexa Knowledge, Amazon. *October 2022– April 2023*

- 6 months internship at Alexa Knowledge, Cambridge, GBR, working with Weiwei Cheng (mentor), Christos Christodoulopoulos, Amir Saffari, Jens Lehmann, and Daniel Masato (PM).
- Implemented a novel multilingual end-to-end entity linking system, mReFinED. The system achieved superior performance and runtime across various benchmark datasets [EMNLP 2023].
- During the internship, I learned about research and development (R&D), Amazon's business model, leadership principles, and software engineering.

Subject Matter Expert in NLP **Bangkok, TH**
Artificial Intelligence Research Institute (AIResearch, Thailand). *Aug 2019– June 2024*

- Launched two Thai large language models, such as WangchanGLM and WangchanLion. Both models outperform ChatGPT and existing Thai LLMs in Machine Reading Comprehension benchmarks.
- Launched a multi-domain and multi-task Thai instruction dataset, including legal, medical, finance, and retail domains. The dataset consists of 40,000 instructions with high-quality and diverse tasks, such as closeQA, openQA, summarization, creative writing, and brainstorming.
- Launched a toolkit for fine-tuning and evaluating Thai LLMs called WangchanX. The toolkit supports various LLMs, such as LlaMa3, SEA-LION, SeaLLMs, and PolyLM. Moreover, we provide an evaluation tool for Thai LLMs. The tool evaluates the response of LLMs in terms of correctness, robustness, and hallucination.

SKILLS

- **Programming languages**: Python (Proficient), SQL (Proficient)
- **Deep learning frameworks**: Keras, Tensorflow, PyTorch, and Transformer
- **Languages**: Thai (Native), English (Proficient)

SOCIAL OUTREACH

- **AI Builder 2021-2024 Program**. Mentored high-school students in creating AI projects such as Cross-Lingual Data Augmentation For Thai Question-Answering (GenBench@EMNLP'23) and Self-instruction for Thai LLMs (ACL-SRW'24). Website: https://vistec-ai.github.io/ai-builders/.
- **Thai-Sentence-Vector-Benchmark Project**. We formulate the first Thai text embedding benchmark, which consists of four tasks: semantic textual similarity (STS), text classification, text pair classification, and retrieval QA. We also experiment with various baseline and existing models on our benchmark. (Github: https://github.com/mrpeerat/Thai-Sentence-Vector-Benchmark)
- **The advisory board of SIGSEA.** As a board member, I contribute to helping SEA not to be low-resource languages throughout **SEACrowd and SEA VL programs.** My responsibility is to review Thai datasets and provide an overview of the project. I also helped them experiment with Thai LLMs and MLLMs.
- **Reviewers**. ACL, EMNLP, EACL, AACL, and *CL's workshop.
- **Talks**. Southeast Asia LLMs: SEA-LION and Wangchan-LION at NLP-OSS@EMNLP'23
- **Special Professor**. From March to May 2024, I was a special lecturer for an NLP class at the Faculty of Arts, Chulalongkorn University, where I taught about NLP and its applications for 8 consecutive weeks.

SELECTED PUBLICATIONS (*equal contribution)

- Peerat Limkonchotiwat, Raheem Sawar, Wannaphong Phatthiyaphaibun, Ekapol Chuangsuwanich, Sarana Nutanong., “**Domain Adaptation of Thai Word Segmentation Models using Stacked Ensemble**” EMNLP 2020
- Nattapol Trijakwanich, Peerat Limkonchotiwat**,** Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, Sarana Nutanong**.,** “**Robust Fragment-Based Framework for Cross-lingual Sentence Retrieval**” EMNLP 2021
- Peerat Limkonchotiwat, Raheem Sawar, Wannaphong Phatthiyaphaibun, Ekapol Chuangsuwanich, Sarana Nutanong., “**Handling Cross- and Out-of-Domain Samples in Thai Word Segmentation**” ACL 2021
- Peerat Limkonchotiwat**,** Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong**.,** “**CL-ReLKT: Cross-lingual Language Knowledge Transfer for Multilingual Retrieval Question Answering**” NAACL 2022
- Peerat Limkonchotiwat**,** Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong**.,** “**ConGen: Unsupervised Control and Generalization Distillation For Sentence Representation**” EMNLP 2022
- Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong**.,** “**Typo-Robust Representation Learning for Dense Retrieval**” ACL 2023
- Peerat Limkonchotiwat**,** Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong**.,** “**An Efficient Self-Supervised Cross-View Training For Sentence Embedding**” TACL 2023
- Peerat Limkonchotiwat, Weiwei Cheng, Christos Christodoulopoulos, Amir Saffari, Jens Lehmann**.,** “**mReFinED: An Efficient End-to-End Multilingual Entity Linking System**” EMNLP 2023
- Peerat Limkonchotiwat***,** Wuttikorn Ponwitayarat*, Ekapol Chuangsuwanich, Sarana Nutanong**.,** “**Space Decomposition for Sentence Embedding**” ACL 2024
- Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Jinheon Baek, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong., “**An Empirical Study of Multilingual Reasoning Distillation for Question Answering**” EMNLP2024
- Panuthep Tasawong, Peerat Limkonchotiwat, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong., “**Efficient Overshadowed Entity Disambiguation by Mitigating Shortcut Learning**” EMNLP 2024
- Peerat Limkonchotiwat***,** Wuttikorn Ponwitayarat*, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, Sarana Nutanong**.,** “**MrCrolin: Multi-consistency Cross-lingual Training for Retrieval Question Answering**” EMNLP 2024
- Holy Lovenia*, Rahmad Mahendar*, ..., Peerat Limkonchotiwat***, …,** Samuel Cahyawijaya* “**SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages**” EMNLP 2024
- Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, ..., Peerat Limkonchotiwat, …., Chong-Wah Ngo., “**WorldCuisines: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines**” NAACL 2025