Khoi Le

minhkhoile2001nsg@gmail.com tokisakikurumi2001.github.io Google Scholar

Research interests

My research interests are in large language models and their applications, including efficient data sampling, cross-lingual transfer learning and low-resource adaptation.

Education

University of Technology, VNUHCM — **HCMC, Vietnam**
Bachelor of Computer Science (English Program) — *Aug 2019 – Nov 2023*

- **GPA**: 3.81/4.00 (Rank: 3/343)
- **Thesis:** LAMPAT: Low-rank Adaptation for Multilingual Paraphrasing using Adversarial Training
- **Thesis Advisor**: Assoc. Prof. Quan Thanh Tho

Work experience

AI Resident — **HCMC, Vietnam**
VinAI Research — *Jul 2022 – Jul 2025*

- **Supervisor**: Assistant Prof. Luu Anh Tuan
- **Research Topics**: Multilingual pretrained language models, low-resource adaptation

Publications

UniBridge: A Unified Approach to Cross-Lingual Transfer Learning for Low-Resource Languages — May 2024

Khoi Le*, Trinh Pham*, and Anh-Tuan Luu
Proceedings of ACL 2024

LAMPAT: Low-Rank Adaption for Multilingual Paraphrasing Using Adversarial Training — Dec 2023

Khoi Le*, Trinh Pham*, Tho Quan and Anh-Tuan Luu
Proceedings of AAAI 2024

Projects

Cross-lingual transfer learning — *Github*

- Develop a system for low-resource languages modeling that leverages the knowledge encoded in multilingual pre-trained language models.
- The system can strategically find a suitable vocabulary size for each language and initialize the new embedding from both semantic and lexical knowledge.
- The system can incorporate different languages at inference time to fully benefit from the knowledge shared between group of similar languages.

Unsupervised multilingual paraphrasing system — *Github*

- Develop a system that could paraphrase sentences in multi-language settings.
- The system is developed without the aid of parallel corpora, only take advantages of the rich and resourceful monolingual corpora.
- The system enhances the quality of paraphrase by using virtual label training - a branch of adversarial training - to generate paraphrase with various linguistic features.

Technologies

Languages: Python, JavaScript, Ruby

Technologies: Pytorch, Huggingface, Numpy, Pandas, FastAPI