Le Khanh Duy

duyvt6663 | duy.le1201@hcmut.edu.vn | (+84) 901136210 | duyvt6663.github.io

EDUCATION

Ho Chi Minh City University of Technology | HCMC, VietNam

Bachelor of Computer Science - *Senior year* - GPA: 3.8/4.0 | Oct. 2020 - Present

Awards: 3rd Prize in *Vietnam Mathematical Olympiad* 2019; 1st Prize in *Blockchain Computing and Mathematics contest* VIASM 2022

Coursework: Data Mining, Algorithms and Data Structure, Computer Vision, Database System, Calculus, Linear Algebra, Statistics and Probability, Natural Language Processing, Machine Learning.

PROFESSIONAL EXPERIENCE

KAIST INTERACTION LAB | **Daejeon, Korea**
Visiting Student Researcher | Jun. 2023 - Oct. 2023

- Proposed and implemented a multi-staged pipeline for data exploration focusing on text highlights, recommendations, and visualization using **Large Language Models** (LLM)
- Grounded LLM's visual reasoning on data using **Text-to-SQL** and **Semantic Similarity**. Recommendations were ranked with **pair-wise ranking prompting**
- Achieved **80.5%** match accuracy; **98%** relevance and **4.07/5** interestingness for recommendation

PUBLICATIONS

[p.1] DataDive: Supporting Readers' Contextualization of Statistical Statements with Data Exploration
Hyunwoo Kim, **Khanh Duy Le**, Gionnieve Lim, Dae Hyun Kim, Yoo Jin Hong, Juho Kim
IUI 2024: Annual Conference on Intelligent User Interfaces (to appear)

[p.2] Semi-supervised Semantic Segmentation using Redesigned Self-Training for White Blood Cell
Quoc-Vinh Luu, **Khanh-Duy Le**, Thanh-Huy Nguyen, Thanh-Minh Nguyen, Tien-Thinh Nguyen, Quang-Vinh Dinh
ICISN 2024: International Conference on Intelligent Systems & Networks (to appear)

PROJECTS (More detail at duyvt6663.github.io)

Kalapa Medical QA | *contest*

- Developed a Vietnamese Medical QA system leveraging a medical corpus comprising 600 articles
- Integrated **Retrieval Augmented Generation** (RAG) by ensembling **SimeCSE_Vietnamese** and **BM25**, with a layer of keyword search using **Ahocorasick** on top
- Finetuned a 3B reader **Vietcunna** with **LoRA** on **GPT-4** generated data
- Achieved **59.6**% accuracy, **0.7** score, **Top 10** out of 300 teams

ZaloAI Math Solver | *contest*

- Developed a Vietnamese Math Solver System leveraging **Llama-7B** for vi2en translation plus custom parser and **Tool-integrated Reasoning Agent** (ToRA)
- Finetuned a baseline **Vi-DeBERTA**, and experimented with various of SOTA mathword models like MetaMath, GPT4 COT
- Achieved **75**% accuracy

Zone-h Captcha Recognition

- Contribute a dataset of **50,000** captcha images with MNIST-like characters with lines, circles, rotation noises
- Train a model composed of GRU + CNN + CTC-loss on the data and obtain **86%** accuracy

Hierarchical Stock Clustering | *research*

- Performed hierarchical clustering on S&P 500 data using **Hierarchical Agglomerative** Linkage algorithms and **Directed Bubble Hierarchical Tree** (DBHT), and benchmarked them against GICS
- Tailored DBHT with ward linkage to improve its Adjusted Rand Index by **20%**
- Constructed a portfolio out of S&P 500 by applying **Hierarchical Risk Parity** on the clusters

TECHNICAL SKILLS

English Proficiency	IELTS 7.5 Overall
Programming Languages	Python, C++, C, SQL, JavaScript
Tech/Tools	Langchain, OpenAI, Pandas, Numpy, Transformers, PyTorch, Django