vmhieu17@gmail.com
linkedin/vmhieu
Hanoi, Vietnam

Hieu Vu (Ian)

AI/ML Research Engineer

gscholar/Hieu M. Vu
github/lone17
leetcode/lone17

Publications

(making it seems more impressive than it is because it's all I got)

1 **Vu**, **Hieu M** and Tan Minh Nguyen (2025). "Angular Steering: Behavior Control via Rotation in Activation Space". In: *Advances in Neural Information Processing Systems*. **Spotlight (top 3.5%)**.

2 Son, Nguyen Hong, **Hieu M Vu**, Tuan-Anh D Nguyen, and Minh-Tien Nguyen (2022). "Jointly Learning Span Extraction and Sequence Labeling for Information Extraction from Business Documents". In: *2022 International Joint Conference on Neural Networks (IJCNN)*. **Oral**. IEEE, pp. 1–8.

3 Nguyen, Tuan-Anh D, **Hieu M Vu**, Nguyen Hong Son, and Minh-Tien Nguyen (2021). "A Span Extraction Approach for Information Extraction on Visually-Rich Documents". In: *Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*. **Best Paper Award**. Springer, pp. 353–363.

Education

(grade is just a number and I got lucky)

Bachelor Degree, Computer Science (Honours Programme) **2020**
UET - Vietnam National University, Hanoi *GPA: 3.83/4.00 (Rank: 1st)*

- **Highest Ranking Graduate.**

Experience

(trying to up-sell myself)

AI Research Engineer **July 2024 — Present**
Torilab

- **Building AI avatars.**
  - Conducting research on controlling LLMs' behaviours without prompting (i.e. activation steering).
  - LLMs fine-tuning (SFT, DPO, classification), prompt engineering, LLMs serving (vLLM, sglang, LitServe) with dynamic batching.
  - Developing the core framework for the customizable chatbots with memory, tool use, dynamic prompts and multi-turn interactions.
  - Other random tasks: User's engagement detection with head-pose tracking; Image Generation with Diffusion-based models and LoRAs; multi-speaker diartization.

AI Research Engineer **Nov 2018 — May 2024**
Cinnamon AI

- **Developed RAG-based applications.**
  - Co-creator of **kotaemon** (24.5k+ ★): An open-source tool for local RAG application built for both end users and developers.
  - Built demos for LLM-powered applications targeting the Insurance domain.
- **Researched, developed, and implemented AI solutions for Document Image Understanding.**
  - Information Extraction and Cross-lingual adaptive pre-training for a low-resource language (Japanese).
  - Created new technical assets by introducing new Information Extraction models that became the new standard for client projects. Increased the f1-score by 2% - 7%.
  - Publish papers on Information Extraction at peer-reviewed conferences, 1 Best Paper Award at DIL-ICDAR'21.
- **Developed data-driven products and processes.**
  - Implement synchronization and local version control for the internal data management system.
  - Initiated data standardization, defined and implemented processes regarding data life cycle and organization.
- Other tasks: Support cross-department alignment; Conduct/Facilitate technical sharing sessions; Training/Mentoring.

Undergraduate Research Assistant **Aug 2017 — Sep 2019**
IOT Lab, UET - Vietnam National University

- **Predictive models for wellbore data using machine learning and statistical methods.**
  - Facies/rock type classification; Time-series Analysis; Permeability Regression; Integrated Prediction Error Filter Analysis (INPEFA) curve calculation; Cumulative and Federated Learning.

Honours and Awards

(more uncomfortably showing off)

Best Paper Award **Sep 2021**
ICDAR 2021, Workshop on Document Images and Language
Paper title: A Span Extraction Approach for Information Extraction on Visually-Rich Documents
Accepted for oral presentation and awarded the Best Paper Award at Workshop on Document Images and Language, ICDAR 2021.

Certificate of Highest Ranking Graduate **Aug 2020**
UET - Vietnam National University
Awarded to students who graduated with the highest GPA amongst the graduating class.

Top 4 Zalo AI Challenge 2018 - Voice Track (Individual participant) **Sep 2018**
Zalo, VNG Corporation
Finished at 4th place on the Private Leaderboard of the Voice Gender/Accent Classification challenge.
Zalo AI Challenge is an annual Kaggle-like competition hosted by Zalo - one of the biggest tech companies in Vietnam. In 2018, the competition attracted over 700 teams competed in 3 challenges.

Personal Projects

(probably no one cares but can't waste emtpy space)

Kotaemon (24.5k+ ★) **- An open-source tool for local RAG application.** **Jan 2024 — May 2024**
Open-source project, Co-creator

- A local RAG-based tool for chatting with your documents. Built with both end users and developers in mind.
- For end users: A local Question Answering UI for RAG-based QA.
- For developers: A framework for building your own RAG-based QA pipeline.

Data Utility Improvement Experiment for DECAF **Oct 2022 — Nov 2022**
Personal research

- A personal research on Causal Inference, Algorithmic Fairness and specifically the paper DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative networks.
- Improved data utility of the DECAF method using alternating graph during synthesis while still achieving similar level of fairness.

Gender/Accent Classification for Vietnamese short voice recordings **Aug 2018 — Sep 2018**
Zalo AI Challenge 2018

- Classify the speaker's voice in a recording (typically under 3 seconds) by gender and regional accent.
- **4th place** on the Private Leaderboard, achieved **79.208% accuracy** in 10 days as an **individual participant**.

Skills

(for those bots that love keywords)

Techincal Fields	**LLMs, ML Interpretability**, Information Extraction, Document Understanding, NLP, Computer Vision, Image Processing, Time-series Analysis.
ML/AI Development	**Pytorch, Transformers, RAG**, LangChain, Haystack, LlamaIndex, Tensorflow/Keras, Scikit-learn, OpenCV.
Software Development	**Git, Github Action**, Docker, CircleCI, DVC.
Programming Languages	**Python**, C/C++, Java, Shell Script.
Industrial Domains	Insurance, Manufacturing, Virtual Companion.
Environments	GCP, AWS, Linux, Windows/WSL.
Natural Languages	Vietnamese (native), English (IELTS 7.5), Japanese (JLPT N4).
Other	Research (Google Scholar), Problem Solving (Leetcode).