Hieu Vu (Ian)

✉ vmhieu17@gmail.com
in linkedin/vmhieu
📍 Hanoi, Vietnam

gscholar/Hieu M. Vu 🎓
github/lone17
leetcode/lone17 </>

Research Interests

My current research focues on 3 areas of Mechanistic Interpretability for Large Language Models (LLMs):

- **Activation Space Geometry (1):** Understanding how concepts and behaviors are represented in the activation space of LLMs. To what extent the Linear Representation Hypothesis and the Platonic Representation Hypothesis hold? [1]
- **Application of Mechanistic Interpretability (2):** Activation Steering, Model Editing, Model Fingerprinting, and other applications. [1, 3, 2]
- **Modular Decomposition (3):** Decomposing LLMs into smaller, interpretable modules that perform specific functions, and studying their interactions. [1, 4, 5]

Publications

*: co-first authorship

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

[2] Dung V Nguyen*, **Hieu M Vu***, Nhi Y Pham*, Lei Zhang, and Tan M Nguyen. *Activation Steering with a Feedback Controller*. Under review at ICLR 2026. 2025. arXiv: `2510.04309 [cs.LG]`.

[3] Ryan Lee*, Dung Viet Nguyen*, **Hieu M Vu***, Lei Zhang, Linh Duy Tran, and Tan Minh Nguyen. *Momentum Steering: Activation Steering Meets Optimization*. Under review at ICLR 2026. 2025.

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

[5] Tuan-Anh D Nguyen, **Hieu M Vu**, Nguyen Hong Son, and Minh-Tien Nguyen. "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. 2021, pp. 353–363.

Education

Bachelor Degree, Computer Science (Honours Programme) **2020**
UET - Vietnam National University, Hanoi *GPA: 3.83/4.00 (Rank: 1st/600+)*

- **Highest Ranking Graduate.**

Honours and Awards

FPT Smart Cloud Credits for Research, $2000 **2025**
FPT Smart Cloud

Best Paper Award **2021**
ICDAR 2021, Workshop on Document Images and Language
Paper title: A Span Extraction Approach for Information Extraction on Visually-Rich Documents

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

Certificate of Merit for Excellent Graduation **2020**
Vietnam National University
Awarded by the President of Vietnam National University to students with excellent academic performance and level of conduct during a 4-year undergraduate programme.

Certificate of Excellent Thesis Defence **2020**
UET - Vietnam National University
Awarded to the best thesis of each Undergraduate Thesis Defence Committee.
Thesis title: A Layout-aware key-value relation predicting model for document images.

Top 4 Zalo AI Challenge 2018 - Voice Track (Individual participant) **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.

Experience

Research Scholar **Jan 2026 — Present**
ML Alignment and Theory Scholars (MATS) *Bekerley, CA, USA*

Fully funded research scholar position under the ML Alignment and Theory Scholars (MATS) program, working on interpretability and alignment of large language models.

Visiting Scholar **Sept 2025 — Oct 2025**
National University of Singapore, Advisor: Prof. Tan Nguyen *Singapore*

Fully funded visit to NUS to collaborate on research with Prof. Tan Nguyen.

AI Research Engineer **July 2024 – Present**
Torilab *Hanoi, Vietnam*

- Research on controlling LLM behaviours via activation steering and related intervention methods.
- LLM fine-tuning (SFT, DPO), prompt engineering, and large-scale model serving (vLLM, sglang, LitServe).
- Developed core framework for customizable chatbots with memory, tool use, and multi-turn interactions.
- Additional work: engagement detection, diffusion-based image generation, multi-speaker diarization.

AI Research Engineer **Nov 2018 – May 2024**
Cinnamon AI *Hanoi, Vietnam*

- Co-creator of **kotaemon** (24.7k+ ★), an open-source platform for local RAG applications.
- Developed RAG-based and LLM-powered applications for industry use cases.
- Research in Document Image Understanding: information extraction, cross-lingual pre-training, and model development.
- Published peer-reviewed papers; one Best Paper Award at DIL-ICDAR'21.
- Led data standardization initiatives and improved internal data management systems.
- Mentored engineers and contributed to technical knowledge sharing across teams.

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.

Skills

Techincal Fields	**LLMs, ML Interpretability, RAG, Agentic AI**, Information Extraction, Document Understanding, NLP, Computer Vision, Image Processing, Time-series Analysis.
ML/AI Development	**Pytorch, Transformers, vLLM**, sglang, LangChain, Haystack, LlamaIndex, Tensorflow/Keras, Scikit-learn, OpenCV.
Software Development	**Git, Github Action**, Docker, CircleCI, DVC.
Mathematics	Linear Algebra, Probability, Statistics, Optimization, Control Theory.
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).

Professional Services

Teaching/Mentoring
Cinnamon AI Bootcamp 2020, 2022, 2023

- Mentored groups of 3-4 students.
- Designed syllabus, prepared entrance tests, interviewed candidates.
- Prepared materials and gave lectures on Language Modelling and Transformers.

Reviewer
ICLR 2026, MoFA @ ICML 2025, XLLM-Reason-Plan @ COLM 2025

Conference/Workshop Presentations

- Angular Steering: Behavior Control via Rotation in Activation Space. *NeurIPS 2025* (Spotlight Poster); *Mechanistic Interpretability Workshop at NeurIPS 2025* (Poster).
- A Span Extraction Approach for Information Extraction on Visually-Rich Documents. *ICDAR 2021, Workshop on Document Images and Language* (Oral).

Personal Projects

Kotaemon (24.7k+ ★) **- 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**.

References

Professor Tan M. Nguyen
Assistant Professor (Presidential Young Professor)
Department of Mathematics, National University of Singapore
Email: tanmn@nus.edu.sg

Diep Thi-Ngoc Nguyen, PhD
Deputy Head of Machine Learning Laboratory, Institute for Artificial Intelligence
University of Engineering and Technology, Vietnam National University, Hanoi
Email: ngocdiep@vnu.edu.vn

Dat Nguyen, PhD
Postdoctoral Fellow
Harvard University & Basis Research Institute
Email: datnguyen@seas.harvard.edu

Trung-Kien Nguyen, PhD
AI Manager
Cinnamon AI
Email:cain@cinnamon.is