

MINH N. TA

GenAI/NLP Researcher

Last updated: June 2, 2025

 tnminh.com
 +84 942 942 968
 Hanoi, Vietnam

 minh@tnminh.com
 ngocminhta
 ngocminh-ta

OBJECTIVES

In the short term, I seek to apply my expertise in Natural Language Processing and Generative AI to solve real-world problems through impactful research and engineering. In the long term, my goal is to become a leading technologist and innovator in the field of AI, contributing to the development of responsible, human-centric technologies and building solutions that benefit society.

SKILLS

Technical: Statistics, Optimization & algorithms, Big Data Architecture, Machine Learning, Deep Learning, NLP, LLM.

Tech-stack: Python, Java, Scala, SQL, Hadoop, Spark, Kubernetes, PyTorch, TensorFlow, Langchain.

Interpersonal: Leadership, Logical & Critical thinking, Teamwork, Problem-solving.

Languages: Vietnamese (Native), English (Advanced).

EDUCATION

2025 – 2029 **Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)**
Doctor of Philosophy in Natural Language Processing
Supervisor: Professor Preslav Nakov

2021 – 2025 **Hanoi University of Science and Technology** **Grade: Excellent**
Bachelor's in Data Science, specialized in Artificial Intelligence
Thesis: ScholarSleuth: A System for Fine-grained AI-Generated Content Detection

RELEVANT EXPERIENCE

2024.5 to Now **Visiting Researcher** **Mohamed bin Zayed University of Artificial Intelligence**
I joined the Research Internship program and currently working on some projects of *AI-generated text detection* under the supervision of professor Preslav Nakov.

2023.8 to Now **Teaching Assistant** **School of ICT, Hanoi University of Science and Technology**
I'm a teaching assistant in some undergraduate courses of Database, Deep learning, and NLP. I got certificates of recognition for outstanding performance in this activity.

2022.2 to Now **Research Assistant** **The International Research Center for AI – Foundation Models Lab**
My research mainly focus on applying state-of-the-art LLM techniques to handle NLP tasks in Vietnamese under supervision of Dr. Dinh Viet Sang.

2023.4 to 2025.3 **Data & AI Engineer** **Viettel Group Headquarters**
I joined Viettel Digital Talent program (selection rate: 10%), top 2 best projects of IT division. I'm the only selected intern in Data & AI engineering to be promoted to work at the headquarters as an official employee.

- Building new components and features for data science and data lakehouse project.
- Develop new AI components in new big data platforms and products.

2023.6 to 2025.2 **Tutor** **Datapot Analytics Group & Foreign Trade University**

- Jointly contribute to building curriculum of Data Science Foundation course.
- Teach all practical parts of courses in Data Science in Economics.

HONORS AND AWARDS

2025 Excellence Academic Performance Scholarship from University President

2024 Certificate of Merit from Hanoi's Youth Union for excellency in contributing to student activities.

2024 Provincial-level prize for Student with Five Good Merits (Academic Performance, Ethics, Physical Fitness, Volunteerism, and Integration). (top 0.1% students in Hanoi)

2024 Best Presentation Award in MBZUAI UGRIP Program

2024 Gold Prize in HUST-SolCT Creative Idea Challenge 2024.

2023 – 2024	Certificates of Recognition from the dean for outstanding performance while serving as a teaching assistant.
2022 – 2024	Certificates of recognition from University's president for excellent academic performance and remarkable contributions to students' activities.
2023	Certificate of recognition as an excellence mentor during the Digital Talent Development program of Google and Vietnam National Innovation Center.
2022	Champion of Data Science Talent Competition 2022.
2022	Silver Prize in HUST-SolCT Creative Idea Challenge 2022.

CERTIFICATIONS

2024	University of Illinois Urbana-Champaign's Certificate of Strategic Technology Management.
2023	IBM Data Engineering Professional Certificate.
2023	Google Data Analytics Professional Certificate.

PUBLICATIONS

For an updated list, please refer to my [Google Scholar](#) profile.

2025	GenAI Content Detection Task 1: English and Multilingual Machine-Generated Text Detection: AI vs. Human PDF Proceedings of the 1st Workshop on GenAI Content Detection, COLING 2025.
2024	LLM-DetectAlve: a Tool for Fine-Grained Machine-Generated Text Detection. PDF Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (<i>Co-first author</i>).

REMARKABLE PROJECTS

Oct. – Feb. 2024	Sensitive data detection Private project for Enterprise								
	<table border="1"> <tr> <td>Project description</td><td>This project aims to create a tool to detect some sensitive data defined in Decree No. 13/2023/ND-CP of the Government on Personal Data Protection such as basic identification information, and some specific sensitive data such as political or religious attitude, etc. in data streaming flow or big batch of data files in Viettel Lakehouse Platform.</td></tr> <tr> <td>Used technologies</td><td>PyTorch, LLM, Passkey retrieval, Kubernetes.</td></tr> <tr> <td>Team size</td><td>1</td></tr> <tr> <td>Results</td><td> - Generated a simulated dataset for sensitive data using the predefined contents in the decree. - Developed a simple model to detect sensitive data based on patterns and handled some simple cases and a model with passkey retrieval techniques to identify some small sensitive data in a very long input. - Deployed the model to Kubernetes and integrated it to a data lakehouse system. </td></tr> </table>	Project description	This project aims to create a tool to detect some sensitive data defined in Decree No. 13/2023/ND-CP of the Government on Personal Data Protection such as basic identification information, and some specific sensitive data such as political or religious attitude, etc. in data streaming flow or big batch of data files in Viettel Lakehouse Platform.	Used technologies	PyTorch, LLM, Passkey retrieval, Kubernetes.	Team size	1	Results	- Generated a simulated dataset for sensitive data using the predefined contents in the decree. - Developed a simple model to detect sensitive data based on patterns and handled some simple cases and a model with passkey retrieval techniques to identify some small sensitive data in a very long input. - Deployed the model to Kubernetes and integrated it to a data lakehouse system.
Project description	This project aims to create a tool to detect some sensitive data defined in Decree No. 13/2023/ND-CP of the Government on Personal Data Protection such as basic identification information, and some specific sensitive data such as political or religious attitude, etc. in data streaming flow or big batch of data files in Viettel Lakehouse Platform.								
Used technologies	PyTorch, LLM, Passkey retrieval, Kubernetes.								
Team size	1								
Results	- Generated a simulated dataset for sensitive data using the predefined contents in the decree. - Developed a simple model to detect sensitive data based on patterns and handled some simple cases and a model with passkey retrieval techniques to identify some small sensitive data in a very long input. - Deployed the model to Kubernetes and integrated it to a data lakehouse system.								
May – Aug. 2024	Machine-generated text detection Research project								
	<table border="1"> <tr> <td>Project description</td><td>This projects submitted a paper to EMNLP 2024 - System Demo. as a classification tool of an input text into one of the four types: human-written, machine-generated, machine-polished and machine-humanized text.</td></tr> <tr> <td>Used technologies</td><td>PyTorch, Llama-3, RoBERTa, Doman-adversarial NN, GAN.</td></tr> <tr> <td>Team size</td><td>6</td></tr> <tr> <td>My involvement</td><td> - Completed building dataset and model for training process. - Research to implement in two ways: 1, Using RoBERTa with DANN architecture for simple detector; 2, Using Llama-3 with the GAN architecture for explainable detector. </td></tr> </table>	Project description	This projects submitted a paper to EMNLP 2024 - System Demo. as a classification tool of an input text into one of the four types: human-written, machine-generated, machine-polished and machine-humanized text.	Used technologies	PyTorch, Llama-3, RoBERTa, Doman-adversarial NN, GAN.	Team size	6	My involvement	- Completed building dataset and model for training process. - Research to implement in two ways: 1, Using RoBERTa with DANN architecture for simple detector; 2, Using Llama-3 with the GAN architecture for explainable detector.
Project description	This projects submitted a paper to EMNLP 2024 - System Demo. as a classification tool of an input text into one of the four types: human-written, machine-generated, machine-polished and machine-humanized text.								
Used technologies	PyTorch, Llama-3, RoBERTa, Doman-adversarial NN, GAN.								
Team size	6								
My involvement	- Completed building dataset and model for training process. - Research to implement in two ways: 1, Using RoBERTa with DANN architecture for simple detector; 2, Using Llama-3 with the GAN architecture for explainable detector.								

Dec. 2023
– Mar. 2024

Integrated LLM Coding Partner for Data Lakehouse

Private project for Enterprise

Project description	This projects aim to create an assistant which can: i , give informations about data lakehouse and user manual; ii , generate SQL, Python and Scala code to extracting information from data lakehouse.
Used technologies	Python, Tensorflow, Transformer, QLoRA, SparkAI, RAG
Team size	1
Results	• Complete training models for code generation and question-answering in English. • Complete the extensions for Visual Studio Code and JetBrains's IDE along with the library for LLM integration with connection to data lakehouse.

Oct. – Dec. 2023

LongLoRA Technique and Vietnamese LongAlpaca Dataset

Research Project

Project description	This projects investigates in the technique to make the LLM have the capacity of understanding a whole long book; along with create a new dataset for training using this technique
Used technologies	Python, Tensorflow, Transformer, LoRA
Team size	1
Results	• Successfully implement LongLoRA Techniques for LLM. • Successfully proposed a method for collecting Vietnamese LongAlpaca Dataset, building in progress.

EXTRACURRICULAR EXPERIENCES

2024.6 to Now

Deputy Secretary

Ho Chi Minh's Youth Union - SOICT

I am the only student on the HUST-SOICT Ho Chi Minh Youth Union Executive Committee (where the secretary is the professor).

I'm in charge of organizing student activities (workshop, competitions, conferences, etc.), particularly assisting students in engaging more in studying, scientific research, and acquiring more valuable skills for their future career paths.

Prior to this appointment, I served as the head of Study, Research and Orientation department.

2022.8 to 2024.8

President & Leader

Google Developer Group on Campus - Hanoi University of Science and Technology

I served as the leader after the group joined as an official club of the university. During these two years, I led several workshops and competitions, which engaged 3,000+ students and collaborated with 10+ companies in Vietnam.

Besides, GDG-HUST has been recognized as one of the most outstanding Google Developer Groups in South-East Asia.

REFERENCES

Prof. Preslav Nakov, PhD.

Head of Natural Language Processing department,
Professor of Natural Language Processing
Mohamed bin Zayed University of AI
preslav.nakov@mbzuai.ac.ae

Dr. Sang Viet Dinh

Vice Dean of School of ICT,
Director of The Intl. Research Center for AI
Hanoi University of Science and Technology
sangdv@soict.hust.edu.vn

Mr. Thanh Chi Nguyen

Head of Customer 360 Department, VPBank
Former Head of Technology R&D Department,
IT Division, Viettel Group HQ
thanhnc29@vpbank.com.vn

Dr. Tuan Cong Truong

Director of Innovation and Entrepreneurship Center,
Secretary of University's Youth Union
Hanoi University of Science and Technology
tuan.truongcong@hust.edu.vn