



NGOC-MINH, TA

Data & AI Engineer

 ngocminhta.id.vn

 +84 942 942 968

 Hanoi, Vietnam

 ngocminhta.nmt@gmail.com

 ngocminhta

 ngocminh-ta

OBJECTIVES

Utilizing the skills and knowledge acquired during my university major, I want to pursue a career as a technopreneur specializing in big data and AI. Pursuing higher education helps me get more knowledge and practical experience, which gives me an advancement in my career.

SKILLS

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

Technical: Statistics, Optimization & algorithms, Big Data Architecture, Machine Learning, Deep Learning, Natural Language Processing.

Tech-stack: Python, Java, Scala, SQL, R, Hadoop, Spark, Kubernetes.

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

EDUCATION

| | | |
|-----------------|---|----------|
| Sep. 2021 - Now | Hanoi University of Science and Technology Bachelor of Science in Data Science and Artificial Intelligence | 3.50/4.0 |
| May - Jun. 2024 | Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) Exchange research resident of NLP track with the acceptance rate of 4% | Exchange |

RELEVANT EXPERIENCE

| | | |
|-----------------|-------------------------------------|--|
| May - Jun. 2024 | Research internship resident | Mohamed bin Zayed University of Artificial Intelligence I joined the Undergraduate Research Internship program and currently working on the project of <i>Machine-generated text detection</i> under the supervision of professor Preslav Nakov. |
| Apr. 2023 - Now | Data & AI Engineer | Viettel Group Headquarters I joined Viettel Digital Talent program (selection rate of 10%), top 2 best projects of IT division. I'm the only selected interns in AI to work at the headquarters. <ul style="list-style-type: none">• Building new components and features for data science and data lakehouse project.• Investigate to develop AI components in new big data platforms and products.• Current project: Build LLM Assistant for Data Lakehouse. |
| Jun. 2023 - Now | Lecturer | Datapot Analytics Group <ul style="list-style-type: none">• Jointly contribute to building curriculum of Data Science Foundation course.• Teach practical parts and give lab instructions of courses in SQL, Azure, Python. |
| Feb. 2022 - Now | Research Assistant | The Intl. Research Center for AI - Foundation Models Lab <ul style="list-style-type: none">• My research mainly focus on Vietnamese large language model training techniques.• Prior: Optimization and Operation Research lab. |

HONORS AND AWARDS

| | |
|-------------|---|
| 2024 | Gold Prize in HUST-SolCT Creative Idea Challenge 2024. |
| 2022 & 2023 | Certificate of recognition from University's president for excellent academic performance and remarkable contributions to students' activities. |
| 2022 | University's prize for Student with Five Good Merits. |

| | |
|------|--|
| 2022 | Champion of Data Science Talent Competition 2022. |
| 2022 | Silver Prize in HUST-SolCT Creative Idea Challenge 2022. |
| 2022 | Bronze Prize in AOF Young Startup 2022. |

CERTIFICATIONS

| | |
|------|--|
| 2023 | IBM Data Engineering and Machine Learning using Spark. |
| 2023 | IBM Data Engineering Professional Certificate. |
| 2023 | Google Data Analytics Professional Certificate. |
| 2020 | Intermediate Tensorflow in Google Cloud. |

REMARKABLE PROJECTS

May 2024 – Now **Machine-generated text detection** **Research project**

| | |
|----------------------------|--|
| Project description | This projects aim to classify an input text is 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 | <ul style="list-style-type: none"> 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. |

Mar. 2024 – Now **LLM's Explainability Towards Explainable Temporal Reasoning** **Research project**

| | |
|----------------------------|--|
| Project description | This projects aim to clarify the explainability of LLM, enclosing the gap between LLM and Explainable AI – which make LLM's decision understandable and clearer. |
| Fields | Explainable AI, Deep Neural Networks, LLM |
| Team size | 2 |
| Progress | <ul style="list-style-type: none"> Researching about the method of understanding LLM's decisions in various aspects. Implementing the training code to make LLM has the ability of temporal reasoning. |

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 tables in data lakehouse. |
| Used technologies | Python, Tensorflow, Transformer, QLoRA, SparkAI, Vector DB |
| Team size | 1 |
| Results | <ul style="list-style-type: none"> Complete training models for code generation and question-answering in English. Partially complete the extensions for Visual Studio Code and JetBrains's IDE (IntelliJ, PyCharm, etc.) along with the library for LLM integration with connection to data lakehouse. |

| | |
|----------------------------|---|
| 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 | <ul style="list-style-type: none"> • Successfully implement LongLoRA Techniques for LLM. • Successfully proposed a method for collecting Vietnamese LongAlpaca Dataset, building in progress. |

| | |
|----------------------------|---|
| Project description | This project aims to create a multilingual chatbot assistant for data science and engineering knowledges, skills and tools for enterprise. |
| Used technologies | Python, Tensorflow, Transformer, QLoRA, SFT |
| Team size | 2 |
| Results | <ul style="list-style-type: none"> • The models has the accuracy of 76% in the collected dataset from IT experts, also has the similar benchmark to LLaMA-2 7B. • A chatbot has been established in Viettel Group IT Division in experiment phase and get many positive feedbacks. • It can understand and answer fluently in English; however, some new contexts (written in English) in Vietnamese question make it confused (answer in English instead of Vietnamese) |

EXTRACURRICULAR EXPERIENCES

I am the only student on the HUST-SOICT Ho Chi Minh Youth Union Executive Committee in charge of organizing student activities, particularly assisting students in engaging more in studying, scientific research, and acquiring more valuable skills for their future career paths.

Sponsored by Google for Developers, GDSC-HUST aims to help students work with Google's latest technology. Besides, GDSC-HUST has been recognized as the most outstanding GDSC in Vietnam.

REFERENCES

Dr. Sang Dinh Viet

Deputy Director, The Intl. Research Center for AI,
Head of Foundation Models Lab,
Hanoi University of Science and Technology
sangdv@soict.hust.edu.vn

Prof. Preslav Nakov, PhD

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

Mr. Thanh Nguyen Chi

Former head of Technology R&D Department,
IT Division, Viettel Group HQ
thanhnc29@vpbank.com.vn