

---

## SUMMARY

- Researcher/Engineer in AI/ML.
- Experienced in Document Image Understanding (Layout Analysis, OCR, Information Retrieval, Information Extraction, Language Modelling).
- Published [papers](#) at peer-reviewed conferences, including 1 Best Paper Award at DIL-ICDAR 2021.
- Bachelor of Computer Science (Honours Programme) - Highest Ranking Graduate (1st/600+).

---

## EXPERIENCE

### AI Research Engineer

Nov 2018 — Present

*Cinnamon AI*

*Hanoi, Vietnam*

- **Developing LLMs-powered applications for Insurance Domain.**
  - Developing an internal framework for Large Language Model (LLM) pipeline building: allowing users to build workflows that consists of LLM, Prompts, Vector Stores, Indexers and Retrievers, Agents, Output Parsers, etc.
  - Built demos for LLM workflows targeting the Insurance domain.
  - Related skills: Large Language Models, Prompt Engineering, Information Retrieval, Software Engineering, Business-oriented Development.
  - Technologies used: OpenAI, Cohere, LangChain, LangFlow, Github Actions, HuggingFace, LlamaIndex, Haystack.
- **Researched, developed, and implemented AI solutions for Document Image Understanding.**
  - Published [papers](#) at peer-reviewed conferences, including 1 Best Paper Award at DIL-ICDAR 2021.
  - Key Information Extraction on document images low-resource languages:
    - \* Implemented MVLM pre-training task for LayoutLM (and variants).
    - \* Adapted the English pre-trained weights to a low-resource language (Japanese).
    - \* Pre-trained LayoutLM-based models for the Japanese language and performed fine-tuning on several client data sets, increased the f1-score by 2% - 7%.
  - Document Image Classification, over 85% accuracy achieved on a client data set with 20+ classes.
  - Other: Document Segmentation; Document Object Detection (logos, stamps, check marks, etc.); Data Synthesis/Augmentation (Image Processing based); Text Segmentation; printed/handwriting OCR.
  - Related skills: Research, Training/fine-tuning, Language Model Pre-training, Image Processing, Computer Vision, Natural Language Processing.
  - Technologies used: Python, Pytorch, Tensorflow/Keras, Transformers (Hugging Face), OpenCV, Scikit-learn, L<sup>A</sup>T<sub>E</sub>X, DVC, Neptune, CircleCI, Docker.
- **Developed data-driven products and processes.**
  - Worked on Data Management CLI tool: Synchronization and local version control, used by AI Engineers and Researchers to query data from a central database and manage the local copy.
  - Roadmap planning for data-related objectives: lead discussions, identify issues, propose solutions, decide action items for data centralization, data management, data integrity, labeling UI improvement, etc.
  - Initiated data standardization: defined and implemented processes regarding data life cycle and organization, enabling datasets from different client to be re-useable collectively.
  - Related skills: Data Management, Label Schema Design, Data-driven Development, Software Engineering.
- Other tasks: Supporting cross-department alignment; Conducting/Facilitating technical sharing sessions; Training/Mentoring.

### Undergraduate Research Assistant

Aug 2017 — Sep 2019

*IOT Lab, UET - Vietnam National University, Hanoi*

*Hanoi, Vietnam*

- **Developed the Machine Learning Toolkit of an online wellbore data interpretation and management platform.**
  - Built predictive models for geophysical data using machine learning and statistical methods.
  - Worked in conjunction with domain experts and FE/BE engineers to ensure requirements are met.
  - Problem worked on: Facies/rock type classification; Time-series Analysis; Permeability Regression; Integrated Prediction Error Filter Analysis (INPEFA) curve calculation; Cumulative and Federated Learning.
- Related skills: Data Science, Machine Learning, Time-series Analysis.
- Technologies used: Python, Keras/Tensorflow, OpenCV, Scikit-learn, XGBoost, Javascript.

## SKILLS

<b>Technical Fields</b>	<b>Information Extraction, Document Understanding</b> , Data Science, Natural Language Processing, Computer Vision.
<b>ML/AI Development</b>	<b>Pytorch, Transformers, RAG</b> , LangChain, LlamaIndex, Tensorflow/Keras, Scikit-learn, OpenCV.
<b>Software Development</b>	<b>Git, Github Action</b> , CircleCI, DVC, Docker.
<b>Programming Languages</b>	<b>Python</b> , C/C++, Java, Shell Script.
<b>Industrial Domains</b>	Insurance, Manufacturing.
<b>Environments</b>	GCP, AWS, Linux, Windows.
<b>Natural Languages</b>	Vietnamese (native), English (fluent), Japanese (JLPT N4).
<b>Misc</b>	Problem Solving ( <a href="#">Leetcode</a> ), Attentive to detail, Presentation, Communication, Research ( <a href="#">Google Scholar</a> ).

## PUBLICATIONS

- 1 Nguyen, Dat, **Hieu M Vu**, Cong-Thanh Le, Bach Le, David Lo, and Corina Pasareanu (2024). “Inferring Properties of Graph Neural Networks”. In: *arXiv preprint arXiv:2401.03790*.
- 2 Nguyen, Bao-Sinh, Dung Tien Le, **Hieu M Vu**, Tuan-Anh D Nguyen, Minh-Tien Nguyen, and Hung Le (2022). “Improving Document Image Understanding with Reinforcement Finetuning”. In: *International Conference on Neural Information Processing*. **Oral presentation**. Springer, pp. 51–63.
- 3 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 presentation**. IEEE, pp. 1–8.
- 4 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.
- 5 **Vu, Hieu M** and Diep Thi-Ngoc Nguyen (2020). “Revising FUNSD dataset for key-value detection in document images”. In: *arXiv preprint arXiv:2010.05322*.

## EDUCATION

<b>Bachelor Degree, Computer Science (Honours Programme)</b> <i>UET - Vietnam National University, Hanoi</i>	<b>GPA: 3.83/4.00</b> <i>Rank: 1st/600+</i>
<ul style="list-style-type: none"><li>• <b>Highest Ranking Graduate.</b></li><li>• Merit for Excellent Graduation Certificate recipient.</li><li>• Excellent Thesis Defence Certificate recipient.</li><li>• Excellent Student Certificate recipient (3 times).</li><li>• Academic Encouragement Scholarship recipient (4 times).</li></ul>	

## HONOURS AND AWARDS

<b>Best Paper Award</b> <i>ICDAR 2021, Workshop on Document Images and Language</i> Paper title: <a href="#">A Span Extraction Approach for Information Extraction on Visually-Rich Documents</a> Accepted for oral presentation and awarded the Best Paper Award at <a href="#">Workshop on Document Images and Language, ICDAR 2021</a> .	<b>Sep 2021</b>
<b>Top 4 Zalo AI Challenge 2018 - Voice Track (Individual participant)</b> <i>Zalo, VNG Corporation</i> Finished at 4th place on the Private Leaderboard of the Voice Gender/Accent Classification challenge. <a href="#">Zalo AI Challenge</a> 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.	<b>Sep 2018</b>
<b>Certificate of Highest Ranking Graduate</b> <i>UET - Vietnam National University</i> Awarded to students who graduated with the highest GPA amongst the graduating class.	

## PERSONAL PROJECTS

---

### Question Answering on PDF documents

Jan 2024 — Present

*Open-source project*



- A Retrieval Augmented Generation (RAG) application for question answering on PDF documents.
- Complete pipeline from PDF parsing to indexing, retrieval and generation.

### 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](#).
- Conducted experiments on improving data utility of the DECAF method using alternating graph during synthesis while still achieving similar level of fairness.
- Gave discussion and suggestions on the choice of data utility metrics.

### Channel-invariant Deformable Convolution

2020

*A part of my Undergrade Thesis*



- A modified version of Deformable Convolution where the convolution offsets stay the same for all channels.
- Sped up the Deformable Convolution operation by an order of magnitude while still achieving similar performance.

### Gender/Accent Classification for Vietnamese short voice recordings

Aug 2018 — Sep 2018

*Zalo AI Challenge 2018*



- Problem description: Classify the speaker's voice in a recording (typically under 3 seconds) by gender (male/female) and regional accent (northern/central/southern).
- **4th place** on the Private Leaderboard, achieved **79.208% accuracy** within 10 days as an individual participant.
- About the competition: 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.

### Electric Meter OCR

Oct 2019 — Nov 2019

*University Coursework Project*



- Extract the value on the dial from images of electric meters. The solution is meant to be used in embedded hardware.
- Achieved **0.08 on edit distance** with total code size **under 10MB** and processing time **under 0.3s/image**.

## OUTREACH

---

### Cinnamon AI Bootcamp 2020, 2022, 2023

*Teaching/Mentoring*

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