VU VIET ANH

vietanhresearcher@gmail.com | github.com/VuVietAnh2309
linkedin.com/in/anh-vu-researcher | orcid.org/0009-0006-6977-5922



OBJECTIVES

Motivated AI Engineer with a solid foundation in Machine Learning, Deep Learning, and Large Language Models (LLMs). Passionate about developing and deploying innovative AI solutions while contributing to cutting-edge research. Eager to bridge the gap between academic research and practical applications in a dynamic environment, with a long-term vision of advancing AI technology and pursuing higher-level studies.

EDUCATION

Master of Computer Science – *Hanoi University of Science and Technology*, Hanoi, Vietnam — Jun 2025 – Present
Current CPA: 3.71/4.0

Bachelor of Computer Science – *Hanoi University of Science and Technology*, Hanoi, Vietnam — Oct 2020 – Jul 2024
CPA: 3.48/4.0

WORK EXPERIENCE

Prep Education, *AI Engineer* | Hanoi — Oct 2025 – Present

Project: Automated Essay Scoring System (AES) for Writing and Speaking

- Developed an AI-powered Automated Essay Scoring (AES) system for evaluating IELTS Writing Task 1 & 2 and Speaking performance with high accuracy.
- AES FOR SPEAKING: Designed and implemented the Speaking scoring system using Machine Learning models integrated with three specialized subsystems: APA (Automated Pronunciation Assessment) for pronunciation accuracy and fluency evaluation, GEC (Grammatical Error Correction) for detecting and correcting grammatical errors, and Dify for analyzing vocabulary richness and lexical diversity. Achieved performance metrics: Accuracy $\pm 0.5 \sim 82\%$, QWK and PCC $\sim 90\%$.
- AES FOR WRITING: Designed and trained deep learning models for cross-prompt essay scoring: PANN (Cross-Prompt Model base Deberta-V3) with 189M parameters, and quantized T5 model optimized to 120M parameters. Achieved strong performance: Accuracy $\pm 0.5 \sim 85\%$ (Wt1) and $\sim 87\%$ (Wt2); Accuracy $\pm 1 \sim 98\%$.
- Implemented end-to-end MLOps pipeline: model training $\rightarrow$ Docker containerization $\rightarrow$ GitLab CI/CD automation $\rightarrow$ ArgoCD deployment to production Kubernetes cluster.
- Optimized model serving with dynamic batching techniques and request queuing to handle high-throughput inference for thousands of concurrent student assessments.
- Deployed the system into production, enabling real-time automated scoring with low latency and high availability.

Project: English Lexical Grammar (ELG) Correction System

- Developed a hybrid English grammar correction system combining rule-based approaches with Large Language Models (LLMs) for enhanced accuracy.
- Implemented comprehensive error detection and correction for multiple linguistic aspects: grammatical errors, vocabulary misuse, spelling mistakes, collocation errors, idiomatic expressions abuse or misuse, lexical resource appropriateness, linking devices, and phrase usage errors, helping students identify and correct writing mistakes effectively.
- Designed rule-based modules for deterministic pattern matching alongside LLM-powered inference for complex grammatical constructs.
- Achieved F0.5 score of approximately 60% on private test sets, balancing precision and recall with emphasis on precision to minimize false corrections.
- Designed an LLM-as-judge evaluation framework with an explicit rubric of semantic criteria — grammatical correctness, meaning preservation, minimal-edit fidelity, error-type accuracy, and naturalness of the corrected output — using an evaluator LLM to score each criterion per sample. This produced semantically grounded quality signals beyond what string-overlap metrics (F0.5, BLEU) can capture, surfacing failure modes such as over-correction, paraphrasing drift, and false positives, and directly guided prompt-tuning and rule-refinement iterations.
- Built production-grade serving infrastructure following MLOps best practices: vLLM as the LLM inference backend (continuous batching, PagedAttention) for high-throughput, low-latency serving; Docker containerization for model packaging; GitLab CI/CD for automated testing and deployment; ArgoCD for Kubernetes orchestration.
- Layered application-level request batching and caching on top of vLLM to further reduce tail latency and compute cost while maintaining high throughput under bursty student traffic.
- Deployed the system to support thousands of students daily in improving their writing skills through real-time grammar and vocabulary correction.

Project: Chinese Vocabulary Extraction Pipeline

- Built an end-to-end AI service that transforms images of Chinese text into structured vocabulary data with full linguistic annotations (simplified/traditional, pinyin, POS, bilingual definitions), supporting flashcard generation and language-learning content.
- Designed a 5-stage pipeline: PP-OCRv5 (text detection + recognition) → custom cleaner (reading order, line merging, confidence filtering) → linguistic analysis (jieba segmentation, pypinyin, opencc simplified↔traditional, POS tagging) → self-hosted Qwen3.5-4B-Instruct-2507 enrichment (served via vLLM with continuous batching and PagedAttention) → schema validator.
- Benchmarked a single-stage VLM baseline (Qwen3.6-35B-A3B receiving the raw image directly) against the two-stage OCR + LLM pipeline (PP-OCRv5 followed by Qwen3.5-4B-Instruct-2507). The two-stage approach delivered higher recognition accuracy on dense Chinese text, lower end-to-end latency, substantially reduced inference cost, and made each stage independently testable and replaceable.
- Adopted a hybrid LLM strategy: libraries handle deterministic work (segmentation, pinyin, traditional conversion); the LLM only resolves what libraries cannot do without context — polyphonic pinyin, ambiguous 1-to-many traditional mappings, and bilingual definitions — reducing token usage and improving reliability.
- Encoded explicit Chinese linguistic rules in the LLM prompt (numeral-and-measure splits, particle separation, erhua handling, fixed-phrase preservation, place-name conventions) to ensure consistent re-segmentation of compound entries.
- Optimized GPU memory on PaddlePaddle by capping the VRAM allocation fraction and resizing inputs to max 1920px, lowering idle VRAM from ∼8.7 GB to ∼600 MB and peak usage to ∼2.4 GB on a 24 GB card.
- Implemented an async OCR batcher over an asyncio queue that collects concurrent requests into GPU batches (up to 8 images per forward pass with a 50 ms collection window), maximizing throughput without blocking the event loop.
- Exposed the system as a FastAPI service (predict, pipeline, and health endpoints with Swagger UI) supporting multi-lingual definitions (vi, en, zh, ja, ko) and a Streamlit dev UI; packaged with Docker + NVIDIA Container Toolkit for reproducible GPU deployment.

Vietsens Technology Group, *AI Engineer* | Hanoi — Oct 2024 – Sep 2025

Project: AI Healthcare Assistant for Hospital Management

- Developed an AI-integrated solution for healthcare, deployed within the National Hospital Management Software, particularly adopted by Bach Mai Hospital.
- Utilized advanced Machine Learning and Natural Language Processing (NLP) techniques, including Large Language Models (LLMs), for real-time medical support.
- Designed and implemented modules for patient data analysis, case management assistance, and initial diagnosis support.
- Built interactive AI agents to assist doctors and medical staff with medical history inquiry, prescription, and diagnostics.
- Ensured seamless integration with existing hospital information systems (HIS) for improved operational efficiency and patient care.

Project: Prescription Recommendation and Conflict Detection System

- Developed a predictive system to support doctors in prescribing medications based on patient information.
- Utilized data analysis techniques to preprocess and extract features from electronic health records.
- Applied traditional machine learning classification models (Decision Trees, Random Forests, SVM) and experimented with advanced classifiers for improved accuracy.
- Integrated Large Language Models (LLMs) to evaluate prescriptions, detect duplicate active ingredients, and identify potential drug-drug interactions.
- Delivered a decision-support tool that helps healthcare professionals reduce prescription errors and improve patient safety.

FPT Information System Company, *AI Engineer Intern* | Hanoi — Aug 2023 – May 2024

Project: Image Captioning

- Developed an image captioning system using advanced deep learning models.
- Processed image data to generate descriptive captions that accurately reflect the content of the images.
- Specifically utilized state-of-the-art models such as Transformers and Convolutional Neural Networks (CNNs) integrated with attention mechanisms to enhance caption generation.
- This project aims to create an effective solution for automated image description, leveraging cutting-edge techniques in the field of computer vision and natural language processing.

Project: OCR for Financial Report Recognition

- Developed an OCR system to recognize and extract key information from financial reports of banks.
- Applied state-of-the-art OCR models such as PaddleOCR, Tesseract, and TrOCR (Transformer-based OCR) integrated with NLP for data structuring.
- Utilized LayoutLM / LayoutLMv3 for document layout understanding to improve extraction of tables, figures, and key-value pairs.
- Implemented pre-processing for document images (deskewing, binarization, noise reduction) to improve text detection and recognition accuracy.
- Built post-processing pipeline with Regex + Named Entity Recognition (NER) to normalize extracted information.
- Deployed the model as part of an AI pipeline to assist financial data analysts.

Hanoi University of Science and Technology, *Computer Science Student* | Hanoi Sep 2022 – Sep 2024

Project: Adapting to climate change by improving extreme weather forecasts (WIDS Datathon 2023 - Kaggle)

- This project involves developing a climate change prediction model based on machine learning models.
- Processed data in a time-series format and applied ensemble techniques to enhance the performance of the machine learning models.
- Specifically utilized the latest prediction models such as Gradient Boosting, Light Gradient Boosting, and CatBoost, and combined them.
- This project was part of a competition held at the end of 2022 on the Kaggle platform aimed at predicting climate change.

Project: Vietnamese Speaker Verification (VLSP 2021 competition)

- This project involves developing a Vietnamese speaker verification system using deep learning models.
- Processed voice samples to create speaker-specific acoustic features for enrollment and verification.
- Specifically utilized models such as ECAPA_TDNN, ECAPA_CNN_TDNN, and RawNet3, and experimented with loss functions like triplet, softmax, and proto.
- This project is part of the VLSP competition focusing on speaker verification, aiming to accurately determine if two voice samples are from the same person.

Project: Aspect-Based Sentiment Analysis (ABSA - NLP LAB)

- Built a Vietnamese Aspect-Based Sentiment Analysis system using SA-Transformer.
- Combined NLP techniques such as POS tagging and Biaffine Parsing to extract aspect-sentiment pairs.
- Collected and annotated domain-specific data to train and evaluate the model.
- Achieved improved accuracy in aspect-level sentiment classification compared to baseline methods.

RESEARCH EXPERIENCE

Student Research with Dr. Kate Han and Dr. Thanh Nguyen – March 2023 to Present — **Machine Learning & Ensemble Learning**

VISTA: A Variable Length Genetic Algorithm and LSTM-Based Surrogate Assisted Ensemble Selection algorithm in Multiple Layers Ensemble System – *2024 IEEE Congress on Evolutionary Computation (CEC)* (Published, June 2024)

- This project focused on developing the VISTA method, which integrates a Variable-Length Genetic Algorithm (VLGA) with an LSTM-based surrogate model for optimizing ensemble selection in Multiple Layers Ensemble Systems (MLES).
- Researched existing ensemble learning methods and surrogate-assisted evolutionary algorithms (SAEA).
- Proposed a new method combining VLGA with LSTM to transform variable-length encoding into fixed-size representations for fitness prediction.
- Implemented a multi-layer ensemble system where each layer's ensemble of classifiers (EoC) is trained on both original data and predictions from the previous layer.
- Conducted experiments on 15 popular datasets, demonstrating that VISTA outperforms benchmark algorithms.
- Successfully developed and validated the VISTA method, highlighting its efficacy in improving the performance of MLES.

Imputation Surveys

- Surveyed imputation (SimpleImputer, KNN, MICE, GINN) on 15 UCI datasets for downstream classification.

Student Research with Assoc.Prof. Pham Van Hai & Dr. Pham Danh Tuyen – Sep 2023 to Present — **Computer Vision & Deep Learning**

Enhancing Facial Expression Recognition with Lightweight Attention-Based CNN –
***Journal of Ambient Intelligence and Humanized Computing* (Published, August 2026)**

- Designed a lightweight deep learning model for facial expression recognition optimized for deployment on edge devices with limited computational resources.
- Utilized a truncated MobileNetV2 backbone for efficient feature extraction from facial images.
- Proposed a Patch Extraction Block to divide the feature maps into non-overlapping regions, enabling the model to focus on local facial features even under occlusion or head pose variations.
- Integrated an Attention Classifier with self-attention mechanism to enhance feature discrimination and classification accuracy.
- Achieved competitive results on standard datasets (RAF-DB, FER2013, FERPlus) and a custom real-world challenge dataset with occlusion and pose variations.

Student Research with Dr. Dinh Phu Hung – 2026 to Present **Multimodal Medical Image Fusion**

CoRe-Fusion: Training-Free Multimodal Medical Image Fusion (Under review)

- Training-free PET/SPECT–MRI/CT fusion; curated co-registered brain datasets and an external PET–CT cohort.
- Benchmarked 11 state-of-the-art methods on shared objective metrics, released as the fusion-metrics library.

OPEN-SOURCE SOFTWARE

fusion-metrics – MATLAB-verified image-fusion metrics for Python (Author, 2026)

- Python library of 35 image-fusion quality metrics ($Q^{AB/F}$, MI, SSIM, SD, FMI, FSIM, VIF/VIFF, …), ported line by line from reference MATLAB code and verified against the original .m files to round-off level ($\leq$ 1e-12).
- On PyPI (`pip install fusion-metrics`) with CLI, CI and Zenodo DOI. github.com/VuVietAnh2309/fusion-metrics

TECHNICAL SKILLS

Programming Languages	Python, Bash, C++
Machine Learning Frameworks	Keras, PyTorch
Computer Vision	Object detection, Segmentation, Image processing algorithms
Natural Language Processing	Large Language Models (LLMs), NLP techniques, Question Answering systems
OCR & Document Processing	PaddleOCR, Tesseract, TrOCR, LayoutLM, LayoutLMv3
MLOps & Model Deployment	Docker, Kubernetes, GitLab CI/CD, ArgoCD, vLLM (continuous batching, PagedAttention), Model serving optimization, Dynamic batching, Caching strategies, Production monitoring
Areas of Expertise	Deep Learning, Computer Vision, Large Language Models, Healthcare AI applications, Education

ACHIEVEMENTS

- **2nd Place – VLSP 2026 Competition Award (Top 1 Private Test Score)** (VLSP, 2026)

  Numerical Reasoning Question Answering Shared Task (ViTNumChart) – VLM-based Financial Question Answering (FinQA)

- **2nd Place – WIDS DATATHON 2023** (Kaggle, 2023)

  Adapting to climate change by improving extreme weather forecasts

LANGUAGES

English	VSTEP B2 (March 25, 2025). Strong ability to read and comprehend English documents. Basic communication skills.

PUBLICATIONS

- **Viet Anh Vu**, Gia Khanh Pham, Quy Nam Tran, Son Tung Do, Danh Tuyen Pham & Van Hai Pham. 2026. *"Enhancing facial expression recognition with lightweight attention-based CNN" – Journal of Ambient Intelligence and Humanized Computing*, Springer, published online 10 August 2026. DOI: 10.1007/s12652-026-05116-0.
- Kate Han, Truong Thanh Nguyen, **Viet Anh Vu**, Alan Wee-Chung Liew, Truong Dang & Tien Thanh Nguyen. 2024. *"VISTA: A Variable Length Genetic Algorithm and LSTM-Based Surrogate Assisted Ensemble Selection algorithm in Multiple Layers Ensemble System" – 2024 IEEE Congress on Evolutionary Computation (CEC)*, Yokohama, Japan, 30 June – 5 July 2024, pp. 1–9. DOI: 10.1109/CEC60901.2024.10612029.

CERTIFICATIONS

IBM Python for Data Science	Certification completed
Kaggle – Intro to Machine Learning	Certification completed
Kaggle – Computer Vision	Certification completed