Nguyen Phu Truong

Natural Language Processing Engineer | nguyenphutruong2707@gmail.com | (+84)981-285-376

Ha Noi, Viet Nam | linkedin.com/in/nguyentruong2707 | github.com/HKAB

Summary

AI Engineer with **3+ years** of experience in Natural Language Processing. Passionate about building scalable machine learning systems, optimizing inference efficiency, and applying large language models in real-world applications. Strong background in algorithms, model deployment, and production-ready AI pipelines.

Skills

Languages: English (TOEIC 960/990, Speaking & Writing 310/400), Vietnamese (native)

Programming: Python, React, Node.js, FastAPI, Docker

AI & Data Science: Mathematics, Data Structures & Algorithms, Machine Learning, Deep Learning

Frameworks & Tools: PyTorch, NeMo, Transformers, Triton Inference Server, vLLM

Cloud: AWS

Work Experience

AI Engineer, GHTK – Ha Noi Apr. 2025 – Present

Project: Order Information Extraction from Conversations

- Fine-tuned Qwen3-8B with QLoRA to extract customer details and order data (name, phone, product, price).
- Deployed the model with vLLM for efficient, low-latency serving.
- Technologies: LLMs, vLLM

Project: Speech Emotion Recognition on phone call

- Improved negative recall by **50%** using Cleanlab + Label Studio data refinement.
- Reduced inference cost with model pruning.
- Technologies: PyTorch, Triton, Cleanlab, Label Studio

AI Engineer, FTECH – Ha Noi Jul. 2023 – Apr. 2025

Project: AI-powered Pronunciation Evaluation

- Delivered system serving **4000 requests/sec** with **10x** runtime improvement and **95%** lower CPU cost.
- Built end-to-end ASR with advanced post-processing alignment algorithm, improved accuracy and significantly reduced runtime vs. hybrid system.
- Integrated Large Language Models (LLMs) for automatic IELTS Speaking assessments, achieving a low mean squared error (MSE) of **0.3** in predicting IELTS Speaking scores on a private dataset.
- Technologies: PyTorch, CUDA, FastAPI, LLMs

Project: Callbot for FQA Customer Service

- Developed Speech-to-Text module based on Whisper, optimize for phone call domain and short audio.
- Technologies: Triton Inference Server, Whisper

Project: Streaming ASR for Wake-up Call Task

- Achieved **96%** accuracy, deployed on Android/iOS devices.
- Technologies: TensorFlow Lite, Mobile SDKs

AI Engineer Intern, VinBigdata – Ha Noi Apr. 2023 – Nov. 2023

- Cleaned and preprocessed speech data for ASR training.
- Improved language model for rescoring of an Kaldi-based ASR system.

AI Trainee, Vingroup AI Engineer Training Program – Ha Noi Jul. 2022 – Apr. 2023

- Projects: Traffic forecasting, Age prediction, Whisper fine-tuning for Vietnamese, Tone restoration.

Lab Assistant, Data Science and Knowledge Technology Lab – Ha Noi 2019 – 2022

- Researched quantization& pruning for efficient inference.
- Shared and discussed academic knowledge in NLP field every week.

Education

University of Engineering and Technology, BS in Information Technology Aug. 2018 – Nov. 2022

- GPA: 3.61/4.0 (Excellent)
- Thesis: Faster inference on Branchynet with entropy-based layer elimination (An algorithmically approach for faster inference with BERT)

Recent projects

N8N Pipeline – Etsy Auto Upload An Etsy Shop

- Designed an AI-driven automated pipeline using N8N to generate product titles, descriptions, and tags with LLMs directly from product images.
- Integrated Etsy API interaction to auto-upload listings, reducing manual effort.
- Technologies: N8N, OpenAI API, Google Cloud Storage

Virtual Idol – 24/7 YouTube Livestream Private Blockchain Team

- Developed an AI-powered cost-efficient virtual idol system capable of real-time audience interaction using Gemini for dialogue and ElevenLabs for speech synthesis.
- Orchestrated behavior control via LLM-driven responses, AutoIT automation, and 3tene avatar animation, enabling continuous 24/7 streaming.
- Technologies: OpenAI API, ElevenLabs, AutoIT, 3tene

Vietnamese Streaming RNN-T vietnamese-rnnt-tutorial

- Crawled, cleaned over **10,000 hours** of Vietnamese speech data sourced from the Internet.
- Implemented a custom RNN-T from scratch. Modified and improved Whisper-small encoder architecture, specially for streaming purpose.
- Archiving a major reduction in Word Error Rate (WER) of **36% on VIVOS** and **55% on Common Voice 17 (Vietnamese)** compare to Whisper-small.
- Sucessfully exported and demonstrated local serving on Chrome using ONNX Runtime Web.
- Publish a complete tutorial and deploy the quantized model on Huggingface Space, achieving **RTF < 1** even with Free tier CPU.
- Technologies: PyTorch, NeMo, ONNX, Huggingface Spaces

Activities and Achievements

ProCon Vietnam 2020 2020

Qualified for the final group stage (32 teams).

Provincial Informatics Competition for 12th Grade Excellent Students 2017

Consolation Prize.