Giang Son Ba

Nghia Do, Cau Giay, Viet Nam
(+84) 976943504 sobagi.104@gmail.com sonba Jester6136

Education

Hung Yen University of Technology and Education **July 2019 – July 2023**
Engineer of Information Technology *My Hao, Hung Yen*

Experience

Alpha Lab **Oct 2022 – Present**
NLP Researcher *Remote*

- Research in Vision-Language Models (VLM) to address Vietnamese Named Entity Recognition (NER) challenges.
- Developed three multimodal NER datasets in Vietnamese based on versions of the VLSP challenges.
- Enhanced Vietnamese multimodal named entity recognition using CrossAttention, argument PixelCNN, contrastive learning, ResNet, and PhoBERT.
- Developed an architecture for more fine-grained hidden representation in NER for biomedical text.
- Fine-tuned summarization models ViT5 and BARDPho for Vietnamese text.

AI Academy Vietnam **May 2023 – Present**
NLP Engineer *Cau Giay, Ha Noi*

- Developing Datalens - AI Stograge Platform.
- Developed Smartchat: "Accelerate Your Business with an AI Assistant" (see Projects for details).
- Leading the development of FinScope, a comprehensive market analysis and prediction platform designed to support investor decision-making with AI-driven insights.
- Enhanced event extraction methods, achieving an 80% improvement in accuracy through optimized machine learning models for event detection and machine reading comprehension.
- Contributed to a Speech-to-Text CoreNLP project, improving speech recognition accuracy and developing an auto-labeling system for audio data, increasing labeling efficiency by 700% compared to manual annotation.
- Built a question-answering demo for VN-Economy using large language models (LLMs) and the LangChain framework, significantly enhancing the system's QA capabilities.
- Developed a keyword extraction module to streamline data processing and information retrieval.
- Implemented module news translation: from Russian, English, and Chinese to Vietnamese.
- Built an HS-code search system, improving classification accuracy and retrieval speed.

AI Academy Vietnam **Aug 2022 – Oct 2022**
AI Engineer Intern *Cau Giay, Ha Noi*

- Building chatbot demo for Nam Dinh Gov.
- Enhancing RASA chatbot Core with FastText Embeddings.
- Integrating RASA chatbot with Telegram and Web Applications.

Projects

Datalens | *GraphRAG, Weaviate, Apache Iceberg, MinIO, Data Lakehouse* **June 2025 – Present**

- Leading the design and initial implementation of an enterprise data lakehouse platform to power AI-driven applications, integrating GraphRAG, MinIO, Weaviate, and Apache Iceberg as core components.
- Built a scalable ingestion pipeline with Airbyte to load data from databases and Google Drive into Apache Iceberg, enabling schema evolution and ACID compliance.
- Developed a modified MinerU pipeline for Vietnamese OCR on scanned PDFs, enhancing text extraction accuracy for downstream NLP tasks.
- Deploying Weaviate and Neo4j as a hybrid vector-graph database ecosystem with MinIO, indexing dense embeddings in Weaviate and entity-relation graphs in Neo4j for semantic search, multi-hop reasoning, and zero-copy document access.

SmartChat | *Chatbot, RAG, LLMs, Indexing, Information Retrieval, VectorDB, AI Agent* **August 2024 - June 2025**

- Engineered a multi-format chatbot system capable of processing and indexing both structured (Excel, QA pairs) and unstructured data (docx, txt, PDFs) for comprehensive knowledge base creation.
- Automated data collection with a custom chatbot brain, seamlessly connecting to external systems.
- Supported 5,000+ concurrent users with autoscaling, load-balanced APIs.
- Improved chatbot accuracy and response time with optimized query transformation and retrieval.

- Leveraged advanced retrieval-augmented generation (RAG) techniques to optimize document-based responses by utilizing a vector database (VectorDB) for improved data access and storage.
- Implemented speculative RAG approaches to reduce latency in generating responses, improving user experience and system efficiency.
- Developed AI agents for multi-document summarization, automated report generation, and intelligent QA pair creation from documentation.

FinScope | *Generative AI, LLMs, Prompting, Instruction FT, Time Series Forecasting* **March 2024 - September 2024**

- Leading a cross-functional team in the development of the FinScope product, guiding both technical and strategic aspects of the project.
- Proposed innovative product ideas for FinScope, outlining technical requirements and adaptations necessary for seamless integration into the platform.
- Designed FinScope as a comprehensive platform for market data analysis, aimed at enhancing investor decision-making and supporting brokers with AI-driven insights.
- Architecting and developing CoreNLP for FinScope, transforming unstructured news and market trends into structured data for advanced analytics and market predictions.
- Developing CoreIndices, a framework that integrates market indices, news trends, and macroeconomic data to provide enhanced investment insights and predictive modeling.

V-OsINT | *Generative AI, Machine Learning, Deep Learning, NLP* **September 2023 - May 2024**

- Developed a system for detecting and extracting key events from news articles.
- Implemented multiple news summarization models to generate concise and informative summaries.
- Built a machine learning-based keyword extraction model for Vietnamese news.
- Implemented a multilingual news translation module, translating Russian, English, and Chinese news into Vietnamese.
- Improved event extraction accuracy by 80% by integrating machine reading comprehension, question-answering, and event detection models.

Speech-to-Text | *Acoustic Modeling, Language Modeling, Decoding, Feature Extraction, BERT* **February 2024**

- Developed preprocessing speech data (removing noise sound, bad-end sound, normalizing speed voice, breaking non-vocal).
- Developed a web-based speech-to-text labeling tool.
- Developed an automatic labeling module that improves performance by at least 700% compared to manual labeling by humans using perplexity and limited WER.
- Proposed post-processing methods for integration (auto-revise bad words generated, auto-generate punctualization).

Publications

An Architecture for More Fine-Grained Hidden Representation in Named Entity Recognition for Biomedical Texts.
ICTA 2023 - Dec 13, 2023

Skills

AI/NLP: LLMs, RAG, NER, Summarization, Vietnamese Language Processing, Multimodal Models
MLOps/Data: Qdrant, Weaviate, Apache Iceberg, MinIO, LangChain, Docker, FastAPI
Programming: Python, C#
Tools: Git, GitHub, Docker, VS Code
Languages: Vietnamese (Native), English - 780 TOEIC (Professional Reading & Writing)