Nguyen Phung Bao Huy

☎ 080-7587-7443 ✉ nguyenphungbaohuy1@gmail.com linkedin.com/in/huy-nguyen-phung-bao github.com/baohuy11

Education

Institute of Science Tokyo (formerly Tokyo Institute of Technology) **Exp. Mar 2027**

B.Eng. in Transdisciplinary Science and Engineering *Tokyo, Japan*

- Concentration: Artificial Intelligence & Machine Learning
- Relevant Coursework: Statistics & Probability; Machine Learning (**Python**); Pattern Recognition; Data Structures & Algorithms (**C++**); Object-Oriented Programming (**Java**); System Programming (**Unix/Linux**)

Work Experience

Startup Company **Dec 2024 – Feb 2025**

AI Engineer Intern *Tokyo, Japan*

- Built a Python-based healthcare chatbot using **Large Language Models (LLMs)**, **Retriever-Augmented Generation (RAG)**, and electronic health records (EHR), improving diagnostic accuracy by **40%**.
- Developed a real-time patient Q&A flow with **React** and a **Python** backend, increasing user engagement by **25%**.
- Collaborated with clinical, product, and engineering teams in Agile (daily standups, sprint planning), strengthening cross-functional communication and end-to-end **AI product development skills**.

Projects

Transformer implementation | *Source Code (C)* | *Source Code (Python)* **C | Python**

- Developed an end-to-end English-Japanese translation system using **PyTorch**, implementing the complete **Transformer** architecture including **multi-head self-attention**, encoder-decoder structure, and leveraging pre-trained tokenizers
- Developed a high-performance Transformer encoder from scratch in **C**, focusing on computational efficiency by implementing core mechanisms like **scaled dot-product self-attention**, **positional encoding**, **feed-forward layers**, and **backpropagation** directly in C for optimized performance.

arXiV Research Paper classtification model | *Source Code* **Python | Docker**

- Developed an arXiv paper classification pipeline using **Scikit-learn** and **NLTK**; skills include NLP **(tokenization, lemmatization)**, feature engineering **(TF-IDF)**, and training/tuning a Gradient Boosting model **(GridSearchCV)**.
- Built and containerized a **RESTful API** using **FastAPI** and **Docke**r to deploy the classification model (via Pickle) and integrate PDF summarization functionality (using **PyPDF2, Hugging Face Models**).

AI Research Chatbot | *Source Code* **Python | Hugging Face | RAG | Langchain | Chroma**

- Developed an offline **Retrieval-Augmented Generation (RAG)** chatbot for intelligent Q&A over academic papers by ingesting daily publications, summarizing key insights, generating vector embeddings, and indexing them in **ChromaDB/FAISS** using **PyPDF**, **BeautifulSoup**, **LangChain** and **Hugging Face Models**.
- Architected a modular **FastAPI** backend with document ingestion pipelines, dynamic prompt configuration, and a production-ready deployment structure.

Technical Skills

- **Programming Languages:** Unix / Linux, Python, C / C++, Java
- **Frameworks & Tools:** PyTorch, TensorFlow, Git, NumPy, Pandas, Scikit-learn, React, Docker
- **Industry Knowledge:** Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Transformers, Generative AI, Machine Learning, Data structures and Algorithms, Object-Oriented Programming (OOP)

Other

- TOEIC 745 (2024), JLPT N1 (Exp. July 2025)
- Deep Generative Models Seminar, University of Tokyo (Feb–Mar 2025)
- AI Business Strategy Seminar, University of Tokyo (Jan–Mar 2025)
- Financial & Machine Learning Seminar, University of Tokyo (Feb–Mar 2025)
- ICPC Coding Competition (C++), Digital Creation Club traP at Institute of Science Tokyo (July 2024)
- Kaggle Contest, Digital Creation Club traP at Institute of Science Tokyo (Feb 2025)