Ta Anh Khoa

✉ khoata301003@gmail.com ☎ (+84) 094 725 9391 Github Linkedin

Objective

Final-year IT student passionate about AI, with expertise in Python and deep learning. Proficient in LangChain, Hugging Face Transformers, and PyTorch. Experienced in building Retrieval-Augmented Generation (RAG) systems, prompt engineering, and fine-tuning large language models (LLMs). Strong background in vector databases, Docker, and API development. Eager to apply my skills to cutting-edge AI projects while collaborating in a dynamic team environment.

Skills

- **PROGRAMMING LANGUEGES**: Python, C++, SQL.
- **AI**:
  - Machine Learning, Deep Learning, NLP, Computer Vision (Basic), LLMs.
  - Framework: Scikit-learn, Keras, Tensorflow, PyTorch, Transformers, BigDL, Langchain, OpenCV.
  - Model Deployment: Streamlit, FastAPI.
- **DATA ENGINEERING & BIG DATA**:
  - Data Analysis: Matplotlib, Power BI.
  - Data Intergration: SSIS.
  - Data Analytics: SSAS.
  - Data Pipeline: Spark, Kafka, Dagster, DBT.
- **OTHER ENGINEERING SKILLS**: Git, Docker.
- **SOFT SKILLS**:
  - Fluent in English.
  - Presentation.
  - Problem-Solving.
  - Teamwork.

Education

University of Information Technology (UIT, VNU-HCM) 2021 – Now

- Major: Information Technology.
- Current GPA: 8/10.

Certificates

Google AI Essentials on Coursera Feb 24, 2025

Projects

VnExpress AI RAG Chatbot - Individual Project 12/2024 - 01/2025

- **Objective**: Develop a Retrieval-Augmented Generation (RAG) chatbot that answers user queries based on AI-related news articles from VnExpress.
- **Tools Used**: Python, FastAPI, Streamlit, Ollama, Together AI, LangChain, ChromaDB, Selenium.
- **Dataset**: AI news articles crawled from VnExpress, stored in a structured database for retrieval.
- **Personal Contributions**:
  - Implemented web scraping pipelines to collect and process AI news articles.
  - Built a FastAPI backend to handle chatbot queries and retrieve relevant documents.

  - Leveraged Sentence Transformers alongside BM25 and a cross-encoder re-ranker for efficient semantic search and ranking in a RAG pipeline.
  - Integrated RAG framework to generate informative responses from retrieved news articles.
  - Developed an interactive user interface using Streamlit.
  - Deployed the system with a scalable database for efficient retrieval.
- **Results**: The chatbot provides accurate responses based on real-time AI news, enhancing accessibility to AI-related information.
- ***Link to repo***

Detection of Spam Comments on Shopee E-Commerce Platform in Vietnam - Team Project 05/2024 - 07/2024

- **Objective**: Develop a system to detect spam comments on Shopee, one of Vietnam's leading e-commerce platforms, to ensure service quality and enhance user experience.
- **Tools Used**:Python, Pytorch, Transformer, XGBoost, Docker, MLFlow, FastAPI, Selenium.
- **Dataset**: Contains 5932 samples of product reviews from Shopee, labeled as Spam and Non-Spam.
- **Personal Contributions**:
  - Data analysis was performed and researched to identify spam patterns.
  - Labeled the data to create a high-quality dataset.
  - Performed data preprocessing and applied feature engineering techniques.
  - Trained, evaluate and compared ML, DL models including XGBoost, LSTM, GRU, CNN and PhoBERT with Accuracy, F1-Score, Recall and Precision.
  - Model Deployment with FastAPI.
- **Rerults**: The XGBoost model combined with PhoBERT feature extraction achieved 89% accuracy and an F1-score of 0.83, outperforming other models.
- ***Link to repo***

Classification of Negative Comments on Social Media - Team Project 04/2024 - 06/2024

- **Objective**: Develop a system to classify negative comments on social media, focusing on Vietnamese content from Facebook.
- **Tool Used**: Python, Pytorch, Transformer, Scikit-learn, Selenium, Jupyer notebook.
- **Dataset**: Contains 9388 samples of comments collected from various Facebook and TikTok posts, categorized into four types: non-negative, profanity-laden, personal attacks, and regional discrimination.
- **Personal Contributions**: Data preprocessing, optimizing the ML, DL models including the best Logistic Regression model, and error analysis.
- **Results**: The Logistic Regression model combined with Bag of Words achieved 76% accuracy and an F1-score of 0.77, improving negative comment detection performance.
- ***Link to repo***

Activities

Faculty of Information Science and Engineering, UIT 09/2024 - Present

- Role: Research student.
- As a research student, currently researching and developing a specialized Vietnamese language dataset for studies on Vietnamese dialects.