Minh Tran Duc

tranducminh503@gmail.com | github.com/tdm503 | Phone: +84 942 476 222

Overview

Recent computer science graduate with one year of hands-on experience in the AI industry. I have worked extensively with machine learning algorithms and developed end-to-end solutions involving data preprocessing, model training, and evaluation. My experience includes building LLM-based agents, fine-tuning and optimizing large language models to solve domain-specific problems, and integrating them into production systems. I have experience fine-tuning embedding models and implementing logic for scoring, risk detection, and semantic matching. I've also designed custom evaluation methods for text-related problems to ensure model performance aligns with business objectives.

Education

Hanoi University of Science and Technology Hanoi, Vietnam
Bachelor in Computer Science, Troy Program.
CPA: 2.9 2021 – 2025

Experience

Technica Industrial AI Hanoi, Vietnam
Position: AI Engineer 02/2025 – Present

- Researched and applied machine learning algorithms to solve industry-specific problems.
- Built and optimized data pipelines for large-scale industrial datasets.
- Collected, cleaned, and engineered data to improve model reliability and performance.
- Fine-tuned embedding models for domain-specific semantic understanding.
- Developed evaluation methods to measure the effectiveness of NLP and retrieval tasks.
- Deployed a Japanese-market insurance chatbot, integrating LLMs and memory modules (e.g., Mem0,Cognee) to enhance contextual dialogue and response quality.
- Evaluated and refined models to ensure robustness and scalability in production environments.

Data Science and Knowledge Technology Laboratory Hanoi, Vietnam
Position: Research Collaborator 02/2025 – Present

- Collaborated on research projects focused on continual learning methodologies.
- Designed and executed experiments to evaluate and enhance model adaptability.

Ownego Hanoi, Vietnam
Position: AI Engineering Intern 9/2024 – 02/2025

- Implemented AI-powered background removal pipeline, deployed to production.
- Automated image processing workflows by integrating deep learning models.

Gameloft Hanoi Studio Hanoi, Vietnam
Position: Game Developer Intern 03/2024 – 06/2024

- Developed Brotato game using SDL2 library.
- Built game prototypes to validate core mechanics and user interactions.

Projects

Dialog Clustering

- Successfully clustered over 5,000 customer support dialog records into coherent topic-based groups, enabling better understanding of recurring customer concerns and automation opportunities.
- Developed a custom dialog clustering pipeline combining embedding generation, UMAP for dimensionality reduction, HDBSCAN for unsupervised clustering, and c-TF-IDF for topic representation.
- Fine-tuned embedding models to improve representation quality for short, noisy conversational text, leading to more coherent clustering results.
- Designed and implemented evaluation methods (e.g., cluster purity, silhouette score, intra-/inter-cluster distance) to assess clustering performance and compare embedding strategies.

- Built a benchmarking framework to test various embedding models (e.g., multilingual SBERT, fine-tuned variants) and select the most effective one for dialog clustering.
- Modified and optimized the HDBSCAN implementation to better handle the nuances of domain-specific conversational data.
- Conducted in-depth analysis of dataset characteristics and tailored clustering strategies accordingly to maximize topic coherence and coverage.

Insurance Callbot

- Built an insurance-focused callbot system powered by LLM agents to handle voice-based customer support tasks such as claim status, policy queries, and document requests.
- Leveraged `openai` SDK to build dynamic reasoning agents with function-calling capabilities for structured task execution and real-time API integrations (e.g., CRM, policy DB).
- Integrated `mem0` for long-term memory management and context retention across multi-turn conversations, enabling better follow-up and continuity in user interactions.
- Designed a modular multi-agent framework where sub-agents (intent classifier, knowledge retriever, form filler) communicate via an agent router using ReAct-style prompting.

Background Removal

- Fine-tuned IS-Net model to optimize inference performance for target hardware and meet company requirements.
- Adjusted hyperparameters and streamlined model architecture to improve segmentation accuracy and processing speed.

KLDA with RFSF Embedding Enhancement - Research project

github.com/tdm503/klda_with_rfsf

- Employed Random Fourier Features (RFF) to project embedding vectors into a higher-dimensional space, boosting classification performance.
- Integrated Random Feature Subspace Feature selection (RFSF) algorithm to refine RFF mappings and further enhance model accuracy.

GMM Model Improvement - Research project

github.com/tdm503/GMM

- Enhanced the original GMM model by applying continual learning techniques to improve adaptability and mitigate catastrophic forgetting.
- Integrated state-of-the-art continual learning strategies to extend the model's capability for sequential tasks.
- Designed and conducted experiments to benchmark performance improvements over the baseline GMM implementation.

Skills

Programing Language: C++, Python, Java
Framework: Pytorch, Tensorflow, FastAPI, Pandas, NumPy, OpenaiSDK, mem0, LangChain, SGLang,Fast MCP
Languages: English (IELTS 6.5) (2021)