



fit@hcmus

UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY

THESIS PROPOSAL

INTEGRATING RASA FRAMEWORK AND RETRIEVAL-AUGMENTED FINE-TUNING FOR VSTEP EXAM ADVISING CHATBOT

1 GENERAL INFORMATION

Supervisor:

– Ph.D Trần Duy Hoàng (Faculty of Information Technology)

[Group] Students:

1. Đinh Dương Hải Nam (Student ID: 21127113)
2. Hồ Thanh Nhân (Student ID: 21127122)

Type of thesis: Research

Implementation period: From 09/2024 to 03/2025

2 RESEARCH CONTENT

2.1 Introduction

2.1.1 Problem Statement

Current Large Language Models (LLMs) are limited by their training data and may struggle with new information or knowledge domains beyond their scope. To address this, methods such as Retrieval-Augmented Generation (RAG) expand knowledge by retrieving real-time information from external sources, while function calling allows LLMs to interact with the real world through predefined functions.

However, implementing function calling in real-world applications presents certain challenges. LLMs require a significant number of parameters to understand and select functions reliably, leading to high computational demands for both training and deployment. This results in a cost issue: current function-calling solutions, based on large models, can be expensive and inaccessible for many applications, especially those with limited resources.

From these challenges, this thesis proposes a new approach to enable function calling for smaller language models, optimizing deployment costs. We will build a chatbot system that integrates the RASA framework-a popular chatbot development platform before LLMs-with Retrieval-Augmented Fine-Tuning (RAFT). This solution aims to provide a viable and cost-effective alternative to traditional function-calling solutions based on large LLMs.

To demonstrate the flexibility and adaptability of this chatbot system in various application domains, this thesis presents a real-world case study: developing a chatbot to provide information support for the **VSTEP** (Vietnamese Standardized Test of English Proficiency) exam. The goal of this application is to automate the consultation process, reduce the workload on support staff, and provide quick and accurate information to candidates.

2.2 Objectives

This research aims to develop a chatbot system using a novel approach to address three main issues: Queries requiring database retrieval, Frequently asked questions and Out-of-domain questions.

2.3 Scope of the Thesis

2.3.1 Research Subjects

The primary research subjects include technologies and techniques such as RASA Framework, RAG, and Retrieval-Augmented Fine-Tuning (RAFT).

Research data: historical chatbot conversations and information about VSTEP candidates over the years, data from real-world surveys, exam regulations, and related sources.

2.3.2 Research Limitations

This thesis will focus on a new approach to building an education advisory system, specifically for the VSTEP exam:

- Data limitations: The primary available data consists of historical conversations, inquiries, and concerns from candidates/parents with Saigon University, as well as related VSTEP exam websites. The data lacks diversity and volume.
- Resource limitations: Due to budget constraints, only free and trial versions of resources will be used, limiting access to and utilization of large-scale resources.

2.4 Proposed Approach

The chatbot system will be built based on three main processing layers:

- Many studies have applied the RASA Framework to university admission advisory systems. Examples include Building a Chatbot for Supporting the Admission of Universities [1], NEU-chatbot-Chatbot for admission of National Economics University [2], Xay dung khung ung dung AI chatbot trong linh vuc quy che dao

tao [3]. RASA is efficient for predefined conversational intents and responses, offering fast replies ideal for handling frequently repeated questions.

- Despite its advantages in response speed, RASA has limitations. It requires manually defining intents and stories, with fixed responses, making it rigid in knowledge adaptation. To address these limitations, we apply Fine-tuning LLMs with RAFT based on RAFT: Adapting Language Model to Domain Specific RAG [4]. This method fine-tunes LLMs to learn from question datasets, improving nuanced query handling, contextual awareness, and enabling retrieval of relevant information. Combining this with Chain-of-Thought (CoT) reasoning enhances response accuracy.

- The outcome of fine-tuning an LLM is a new model version adjusted for a specific domain. However, standard fine-tuning has limitations: knowledge remains static (only updated until training completion), poor handling of out-of-domain queries, and limited reasoning ability. To supplement the LLM with additional knowledge, this thesis incorporates Retrieval-Augmented Generation (RAG), inspired by Retrieval-Augmented Generation for Large Language Models: A Survey [5]. Integrating RAFT fine-tuning with RAG enhances response quality.

This thesis combines the mentioned technologies and techniques to create a chatbot system with more effective response capabilities.

2.5 Expected Outcomes

A chatbot system that meets user needs with higher accuracy:

- A chatbot integrating Rasa, Fine-Tuned LLM with RAFT, and RAG for flexible, precise responses.
- Deployment on messaging platforms (Zalo, Messenger, other chat platforms, or relevant websites).
- Accuracy evaluation using a real-world question dataset.


2.6 Implementation Plan

No.	Task Description	Timeline
1	Research theories, collect and process data	09/2024 - 11/2024
2	Develop chatbot using RASA Framework	11/2024 - 12/2024
3	Fine-tune LLM using RAFT	12/2024 - 01/2025
4	Integrate and optimize RAG system	01/2025 - 02/2025
5	Finalize model, deployment, and thesis writing	02/2025 - 03/2025

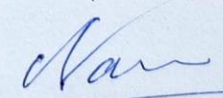
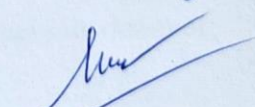
References

- [1] M.-T. Nguyen, M. Tran-Tien, A. P. Viet, H.-T. Vu, and V.-H. Nguyen, "Building a chatbot for supporting the admission of universities," *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, vol. 1, pp. 1-6, 11 2021.
- [2] T. T. Nguyen, A. D. Le, H. T. Hoang, and T. Nguyen, "Neu-chatbot: Chatbot for admission of national economics university," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100036, 2021.
- [3] Đoàn Thị Hồng Phước, L. V. T. Lân, and N. V. Trung, "Xây dựng khung ứng dụng ai chatbot trong lĩnh vực quy chế đào tạo," *Hue University Journal of Science: Techniques and Technology*, vol. 131, pp. 39-52, 06 2023.
- [4] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez, "Raft: Adapting language model to domain specific rag," 2024.
- [5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024.
- [6] M. R. J, K. VM, H. Warriar, and Y. Gupta, "Fine tuning llm for enterprise: Practical guidelines and recommendations," 2024.

CONFIRMATION
ADVISOR'S SIGNATURE
(Full Name)


Trần Duy Hoàng

Ho Chi Minh City, November 21, 2024
STUDENTS' SIGNATURES
(Full Names)


Đinh Diễm Hà Nân

Hồ Thanh Nhân