A Large Language Model-Based Question Answering System for Online Public Administrative Services

Dinh-Dien La[1], Tuan-Anh Nguyen[2], Duc-Huy Mai[2]†, Tuan-Anh Nguyen[2]†, Thi-Thanh Ha[2], Trung-Nghia Phung[2], and Van-Khanh Tran[2]

[1] Department of Information and Communications, Ha Giang Province, Viet Nam.

[2] Institute of Applied Science and Technology - IAST, University of Information and Communication Technology, Thai Nguyen University, Vietnam.

Correspondence: Van-Khanh Tran, tvkhanh@ictu.edu.vn

Received: 02/07/2024, revised: 03/09/2024, accepted: 13/09/2024

Digital Object Identifier: 10.32913/mic-ict-research-vn.v2024.n3.1290

Abstract: Public administrative services are one of the key concerns for citizens; however, there are still numerous procedures and issues that citizens encounter and need clarification on. This study aims to alleviate some of this burden on officials by developing a Question Answering System (QnA) to address inquiries about public administrative services. We present a framework for utilizing Large Language Models to develop and implement a QnA system on the Online Public Administrative Service Portal in Ha Giang Province. We also provide extensive experimental analysis to demonstrate the effectiveness of each component in the QnA system.

Keywords: *Question answering system, public administrative services, large language model*

Tiêu đề: **Hệ thống hỏi đáp dịch vụ công trực tuyến dựa trên mô hình ngôn ngữ lớn**

Tóm tắt: Dịch vụ hành chính công là một trong những mối quan tâm chính của người dân, tuy nhiên vẫn còn nhiều thủ tục và vấn đề mà người dân gặp phải và cần được giải đáp. Nghiên cứu này nhằm giảm bớt một phần gánh nặng cho các cán bộ bằng cách phát triển Hệ thống Hỏi đáp để giải quyết các thắc mắc về dịch vụ hành chính công. Chúng tôi trình bày một nền tảng sử dụng các mô hình ngôn ngữ lớn nhằm phát triển và triển khai hệ thống Hỏi đáp trên cổng Dịch vụ hành chính công trực tuyến tại tỉnh Hà Giang. Chúng tôi cũng cung cấp phân tích thực nghiệm chuyên sâu để chứng minh tính hiệu quả của từng thành phần trong hệ thống Hỏi đáp.

Từ khóa: *Question answering system, public administrative services, large language model*

I. INTRODUCTION

Operating with a commitment to transparency and accessibility, the National Public Service Portal[1] serves as a centralized platform dedicated to assisting individuals and businesses, connecting and providing information on administrative procedures and online public services. It supports the implementation, monitoring, and evaluation of administrative procedure resolution, and online public services, as well as receiving, and handling feedback, and suggestions from individuals and organizations nationwide. The most common bottleneck in the resolution of administrative procedures often lies in the documentation guidance and preparation. Many applications are redundant or lack necessary documentation as per regulations, leading to submissions to the wrong authority for resolution. If this stage is effectively managed, the time required to complete administrative procedures following the law would be significantly reduced, alleviating the burden on both citizens and the officials responsible for receiving and processing applications. In addition, the primary channels for information exchange between the government and citizens are through direct interactions, telephone calls, or social networks. This often results in delays for citizens and repetitive responses from officials.

To provide additional interactive channels between individuals, organizations, and the public services offered

†First batch students in IAST Young Talent Program, 2024

[1]https://dichvucong.gov.vn

by the National Public Service Portal in a quick, timely, 24/7 manner, with personalized 1-1 assistance, and to enhance satisfaction among citizens and businesses with transparency, an artificial intelligence assistant system (a.k.a. QnA [1], or Chatbot [2]) needs to be implemented. This QnA system facilitates information retrieval from the service portal and automatically answers user queries related to many areas such as legal document searches, online public services, and administrative procedures.

Large Language Models (LLMs) [3] represent the cutting edge of natural language processing [4]. The evolution of LLMs [3] has demonstrated promising opportunities for enhancing the efficiency and accessibility of online public administrative services. To the best of our knowledge, this paper is the first work that presents a detailed experimental analysis of LLMs' application in online public administrative services, with a focus on QnA's capabilities in Ha Giang province[2], Vietnam. Our main contributions are threefold:

- Development of public administration QnA datasets: We present a comprehensive set of QnA datasets that cover a wide range of queries related to public administrative services in the National public service portal, Ha Giang portal, and three other provinces' portal.
- Framework for implementing QnA system: We propose a framework for integrating LLMs into QnA system for public administrative services.
- Experimental analysis for each stage in the framework: We provide extensive experiments and evaluations for the QnA system.

II. RELATED WORKS

QnA systems in public administration are designed to handle citizen inquiries and provide timely, accurate information. Several cities in Vietnam are exploring the integration of LLM technology, i.e., ChatGPT[3] into their public administration systems to enhance service delivery and efficiency, such as Ho Chi Minh, Hanoi, Quang Nam, and Da Nang, etc. Authors in [5] present a model to retrieve public administrative information and answer questions related to Vietnamese legal documents. To our knowledge, this is the first work in the literature that provides detailed descriptions of public administrative processes and presents a systematic approach to QnA systems in this context.

a) Online Public Service Portal in Ha Giang

To date, the Ha Giang province Online Public Service Portal has registered 33,434 accounts and received more than 399,440 visits this year. On average, the portal handles approximately 950 applications daily, with an online submission rate of around 90%. This work provides a comprehensive dataset including a wide range of public administrative questions and answers, as well as detailed information on administrative procedures.

b) Large Language Models

LLMs leverage deep learning techniques and massive computational resources to process and generate language, making them capable of a wide range of tasks, from simple question answering to complex text generation and language translation. The rapid advancement of LLMs has opened new avenues for enhancing public administrative services, particularly in the context of online platforms. However, research publications on the application of LLMs in the context of Vietnam's public administrative services remain underexplored. In this work, we experiment with some Vietnamese pre-trained LLMs such as Vinallama-7b-chat[4], Vistral-7b-chat[5], and phoGPT-4b-chat[6], running on local devices. Using Vietnamese Large Language Models, which have been further pre-trained on a substantial volume of relevant Vietnamese data, significantly enhances their ability to understand and process specific laws, regulations, and public services. This leads to better performance in retrieval, reasoning, and answer generation specific to the Vietnamese context. Furthermore, we fine-tune the model using the LoRA method [6] on the training dataset to enhance model's capability to better adapt to the public administration domain.

III. METHODS

1. Datasets

a) QnA Dataset

We provide a comprehensive collection of Q&A datasets covering a wide array of queries related to online public administrative services. These datasets are sourced from the Ha Giang Portal with 1,627 data samples, the National Public Service Portal and additional portals for the provinces of Bac Ninh, Quang Ninh, and Bac Giang in total of 11,334 QnA pairs. We manually selected 202 QnA samples to serve as the test set. Question types of the dataset are in seven categories, including Reasoning Questions, Factoid Questions, Yes/No Questions, Multiple-choice Questions, and Questions involving multiple agencies, multiple relevant documents, and multiple relevant articles. There is a mix of question types, including reasoning-based, factual, yes/no,

[2]https://dichvucong.hagiang.gov.vn
[3]https://openai.com/
[4]https://huggingface.co/vilm/vinallama-7b-chat
[5]https://huggingface.co/Viet-Mistral/Vistral-7B-Chat
[6]https://huggingface.co/vinai/PhoGPT-4B-Chat

and multiple-choice questions, reflecting a diverse set of queries that may require different approaches to answer.

b) Administrative Procedure Dataset in Ha Giang

In compliance with the directives from the Chairman of the People's Committee of Ha Giang Province, we have compiled a comprehensive dataset of 1,943 administrative procedures[7]. Each entry in this dataset includes the following details: ID, link, name, status, sequence of execution, target subjects, methods of implementation, required documents, expected outcomes, resolution time, and the responsible agency, among other relevant information.

The longest procedure in our dataset contains 7,490 words, while the shortest has only 132 words. Due to this variability, some procedures exceed the input limits of embedding and retrieval models. This significant range in word counts underscores the necessity for thorough preprocessing and segmentation to ensure consistent model performance across varying document sizes. Additionally, the third quartile (Q3) value indicates that most samples are considerably shorter than the maximum length, highlighting the presence of a few outliers with exceptionally high word counts. On average, questions in the dataset are about 36 words long and feature numerous key vocabulary terms that enhance retrieval effectiveness. Table III shows statistics of datasets.

By compiling the areas that frequently receive inquiries from citizens to the 1022 [8] hotline of the province, we identified several priority areas for analysis of the questions, including border gate economic zone management, industry and trade, education, and training, among others. Corresponding to these priority areas, we manually extracted 202 QnA pairs from a total of 1,943 administrative procedures. Table I and II summarize a classification of 202 questions from a test set based on different categories, including:

- Reasoning Questions (RQ): These are questions that require logical reasoning to answer, with a total of 14 questions in this category.
- Factoid Questions (FQ): Factoid questions are those that can be answered with specific, extractable information from texts, such as factual data or details from documents. There are 173 such questions, which make up the majority of the dataset.
- Yes/No Questions (YN): These are simple questions where the answer is either "yes" or "no." There are 14 questions in this category.
- Multiple Choice Questions (MC): These questions provide multiple answer options, from which the correct answer must be chosen. There are 20 multiple-choice questions.
- Number of Relevant Documents (ND): This represents the number of legal documents related to each query. There are 187 queries related to 206 legal documents.
- Number of Relevant Articles (NA): This refers to the number of legal articles relevant to each query. There are 92 queries related to 138 legal articles.
- Questions Involving Multiple Agencies (MA): These questions require input or resolution from multiple government agencies. There are 57 such questions.

While Factoid Questions (FQ) dominate the dataset, comprising 85.64% of the total questions, a significant number of questions (187 out of 202) relate to legal documents, and 92 of them specifically relate to articles within those documents. This classification highlights the heavy emphasis on fact-based questions and legal document references, suggesting that the dataset may be designed for scenarios where retrieving and applying legal or procedural information is critical.

c) Administrative Documents Segmentation

The deep learning model follows input constraints by dividing administrative documents into smaller segments aligned with specific administrative procedure fields. This segmentation strategy maintains semantic coherence within each document segment, ensuring the integrity of information retrieval for improved accuracy and contextual relevance of outcomes. Moreover, administrative procedure identifiers are prefixed to segmented documents, providing essential context, especially for concise fields such as fees, charges, agency levels, and other specifics. Figure 1 depicts the length distribution of administrative data before and after segmentation.

2. Administrative Procedure Retrieval

Retrieving documents based on a query involves searching for and selecting appropriate documents to fulfill specific information needs. This typically includes employing models and algorithms to assess the relevance of documents to the query. BM25 [7] relies solely on keyword matching and word frequency within documents, without capturing semantic relationships between words. While it performs well with straightforward queries and clear keywords, BM25 struggles with complex queries that involve multiple layers of meaning or contextual nuances, potentially leading to the omission of relevant documents. Additionally, BM25 may be less effective with documents that have irregular styles, spelling errors, or word variations, unless supplemented by additional preprocessing techniques.

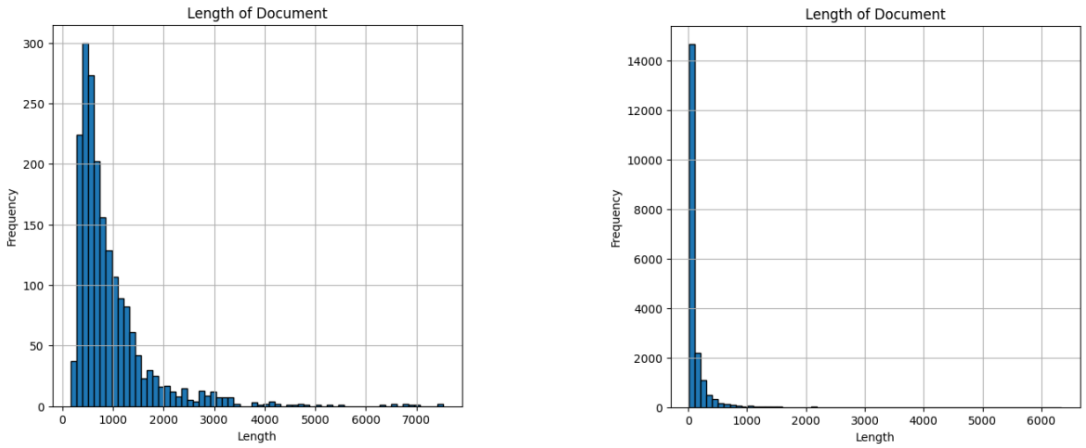
[7]https://dichvucong.hagiang.gov.vn/vi/procedure/search?keyword=
[8]https://1022.hagiang.gov.vn/vi/

Table I
STATISTICS OF QUESTION TYPES

No.	Question Type	Symbol	Quantity
1	Reasoning Questions: Questions involving reasoning	RQ	14
2	Factoid Questions: Questions with extractable answers	FQ	173
3	Yes/No Questions	YN	14
4	Multiple Choice Questions	MC	20
5	Number of relevant documents for each query	ND	187
6	Number of relevant articles for each query	NA	92
7	Questions involving multiple agencies	MA	57

Table II
QUESTION CLASSIFICATION BY GROUP

No.	Question Group	Quantity
1	Factoid Questions (FQ)	112
2	Factoid Questions involving multiple agencies (FQ-MA)	45
3	Factoid-Multiple Choice-Multiple Agencies (FQ-MC-MA)	11
4	Factoid-Multiple Choice Questions (FQ-MC)	3
5	Factoid-Yes/No Questions (FQ-YN)	1
6	Factoid Questions-Multiple Agencies (FQ-MA)	1
7	Multiple Choice Questions (MC)	5
8	Reasoning Questions (RQ)	10
9	Reasoning-Multiple Choice Questions (RQ-MC)	1
10	Reasoning-Yes/No Questions (RQ-YN)	3
11	Yes/No Questions (YN)	10



(a) before pre-processing.

(b) after pre-processing.

Figure 1. Document length (a) before and (b) after pre-processing

Table III
STATISTICS OF ADMINISTRATIVE PUBLIC DATASET

File	Count	Words per sample			
		Min	Max	Avg	Q3
No. of administrative procedures	1,943	132	7490	911	1081
Question (train)	11,334	5	464	30	36
Answer (train)	11,334	1	4,909	174	218
Question (test)	202	7	129	36	41
Answer (test)	202	1	979	131	163

On the other hand, BERT's [8] performance heavily depends on the quality of its training data and fine-tuning. If the data is not representative or properly tailored to the application context, BERT may yield inaccurate results. While BERT processes sentences word by word, SBERT [9] considers entire sentences, providing a more comprehensive understanding of text semantics. This allows search engines to better interpret and categorize content, leading to more in-depth content indexing and more precise search results.

Combining BM25 and SBERT can create a more robust evaluation system by integrating semantic understanding with keyword matching. This is achieved by normalizing the scores from each method and aggregating them using an integrated formula. In this work, we set up an ensemble

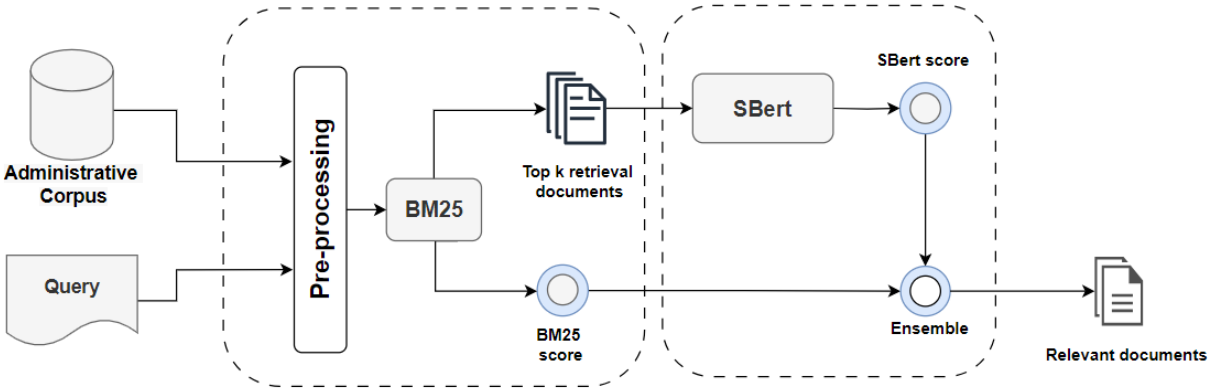


Figure 2. Architecture of the Administrative Procedure Retrieval

method that integrates lexical search BM25 with a semantic search SBERT model. Figure 2 shows the retrieval architecture while experimental results are presented in section IV.1.

a) Text Ranking

Keyword Ranking: The utilization of Okapi BM25 [7] for keyword ranking is an established lexicon-based approach aimed at prioritizing administrative documents. This algorithm assesses administrative procedure relevance by analyzing the frequency of vocabulary terms present in queries. By leveraging commonly used keywords and specialized terminology found in inquiries and related literature, BM25 enhances the model's ability to retrieve pertinent information while optimizing computational efficiency. However, its effectiveness may be constrained when confronted with queries containing low-value vocabulary or ambiguous content, such as "Chi phí đền bù" (*"Compensation"* in English).

Semantic Ranking: Semantic ranking focuses on assessing semantic relevance. The process entails breaking down documents into coherent segments or sentences, which are then transformed into multidimensional vectors. This transformation is also applied to the query, enabling ranking based on semantic similarity. We use Vietnamese-Bi-Encoder model [10], a specific SBERT model for the Vietnamese language, for this task and further fine-tune the model on public administrative datasets for enhancing embedding and semantic search.

b) Ensemble

At this stage, we use min-max scaling algorithm to normalize the output scores x for each query q as:

$$\text{norm}(x) = \frac{x - m_q + 0.01}{M_q - m_q + 0.01} \quad (1)$$

where $M_q = \max(\text{BM25})$ and $m_q = \min(\text{BM25})$. The constant 0.01 is added to prevent division by zero when $M_q = m_q$. The final score is computed as below:

$$\text{Ensemble-Score} = \alpha \cdot w_{\text{BM25}} + (1 - \alpha) \cdot w_{\text{SBERT}} \quad (2)$$

We then employ grid search to determine the optimal weight for parameter α, constrained within the range of 0 to 1.

3. Administrative Procedure Answer Generation

The process of generating answers from retrieved candidate articles aims to gather pertinent information that fulfills the user's information requirements efficiently. This involves identifying key passages within the articles and synthesizing them to construct a coherent and comprehensive response. After the text retrieval stage, the relevant documents undergo preprocessing before being fed into the large language model (LLM). Specifically, synthesize related documents one after another to form a document with enough information. The context within documents is improved by adding additional information from the corpus while removing information with low similarity scores. This process is illustrated in Figure 3.

a) Prompt

A prompt [11] is a piece of text or statement that provides context or specific requirements, and instructions for the model to understand and produce appropriate results. We use a simple prompt that asks the model to get the contextual information provided that the documents are retrieved to answer the question.

b) Chosing LLM for Generating Task

For a QnA task focused on Vietnamese, especially in specialized domains such as law and governance, a Vietnamese large language model offers significant advantages

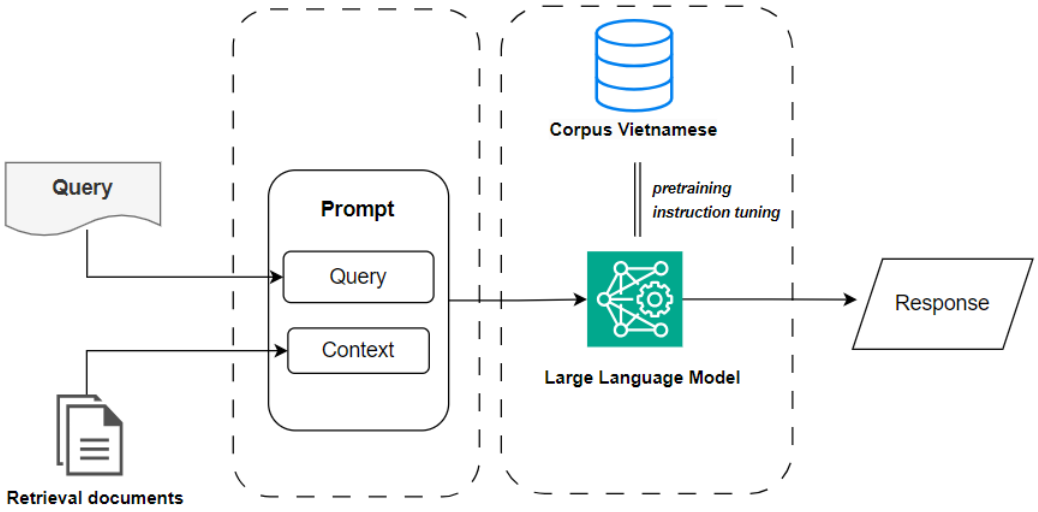


Figure 3. Architecture of the Administrative Procedure Answer Generation

in terms of language proficiency, domain specialization, performance efficiency, and alignment with local needs. Vietnamese LLM candidates to chose were Vinallama-7b-chat[9], Vistral-7b-chat[10], and phoGPT-4b-chat[11]. By leveraging methods like LoRA for fine-tuning, the model can be further optimized to handle specific types of queries effectively, ensuring that it provides the most accurate and relevant responses for Vietnamese users.

c) *Fine-tuning LLMs*

LLMs demonstrated advanced performances in specific domains, but their effectiveness in the public administrative domain is unknown. This provides an opportunity to study whether our chosen LLMs (fine-tuned or not) have high performance on administrative document understanding. The fine-tuning process for LLMs involves several steps to optimize them for specific applications. First, a Vietnamese LLM, such as Vistral 7b chat, is selected and reduced in size using 4-bit quantization, which decreases memory requirements while maintaining high performance. Next, a training dataset consisting of 11,334 QnA pairs about public administrative procedures is prepared. This dataset is formatted by adding special tokens like <s>, </s>, [INST], and [/INST] for training purposes. The model is then fine-tuned using this data, with a specific LoRA configuration designed to enhance processing efficiency and reduce computational costs.

During the fine-tuning process, an A10G GPU with 24GB of RAM was utilized, completing the process in approximately 8 hours over 3 epochs, for a total of 4,249 steps. The LoRA configuration included a lora_alpha of 16, a dropout rate of 0.1, and a rank of 64, targeting specific modules such as q_proj, k_proj, v_proj, o_proj, and gate_proj. The training setup featured a per-device batch size of 8, gradient accumulation steps set to 8, and the paged_adamw_32bit optimizer, ensuring efficient processing throughout the tuning process.

d) *Evaluating Answer Generation Task*

To evaluate the quality of answer generation for administrative procedures and the performance of our QnA system (detailed in Section IV.3), we employ RAGAS (Retrieval Augmented Generation Assessment) as outlined by [12]. RAGAS is a comprehensive framework for assessing Retrieval Augmented Generation (RAG) systems, offering a variety of useful metrics for thorough evaluation. The main RAGAS metrics for this task are Faithfulness, Answer Similarity, Answer Correctness, and Answer Relevancy, details as below:

- Faithfulness measures the factual consistency of the generated response against the retrieved context: if all the claims that are made in the answer can be inferred from the given context then the response is considered "faithful" to the provided context.
- Answer similarity measures the semantic resemblance between the generated answer and the ground truth answer. This is done via cosine similarity between the embedding vectors of the ground truth answer and the generated answer.

[9]https://huggingface.co/vilm/vinallama-7b-chat
[10]https://huggingface.co/Viet-Mistral/Vistral-7B-Chat
[11]https://huggingface.co/vinai/PhoGPT-4B-Chat

- Answer relevancy measures how pertinent the generated question is to the given question. This is computed by generating a number of artificial questions based on the answer and measuring the similarity between the original question and those artificial questions.
- Answer correctness measures how accurate the generated answer is relative to a "golden" answer that is deemed to be the correct answer. It is based on a weighted sum of factual consistency and the semantic similarity between the ground-truth answer and the generated response.

Utilizing GPT-4 to assess the output of chosen LLMs within the RAGAS metrics framework allows for a more reliable and comprehensive evaluation of LLMs, ensuring they meet the standards required for specialized tasks, such as legal and procedural QnA in Vietnamese.

IV. EXPERIMENT RESULTS

1. Administrative Procedure Retrieval

In this task, we use precision, recall, and F2-macro [13] to assess retrieval performance. The metrics are computed as follows:

$$\text{Precision}_i = \frac{\text{Number of correctly retrieved articles of question } i}{\text{Number of retrieved articles of question } i}$$

$$\text{Recall}_i = \frac{\text{Number of correctly retrieved articles of question } i}{\text{Number of relevant articles of question } i}$$

$$F2_i = \frac{5 \times \text{Precision}_i \times \text{Recall}_i}{4 \times \text{Precision}_i + \text{Recall}_i}; \text{and } F2 = \text{Average of } (F2_i)$$

Pre-processed data is first fed to the BM25 component and top-k documents with best BM25 scores are retrieved. Table V shows recall performances for a variety of chosen top-k. As expected, the recall is highest for larger top-k values. For $k = 500$, the recall reaches 99%, indicating that nearly all relevant documents are retrieved when considering a large set of results. This shows that BM25 is effective at retrieving relevant documents when allowed to return a large number of candidates. In our experiments, we select 100 articles with the highest BM25 score for this task.

Combining lexical-based search BM25 with semantic search SBERT has demonstrated significantly better performance compared to using BM25 alone. In the ensemble stage (Section III.2.b), we use grid search to find the best weight α for BM25 and SBERT. The grid search results in table IV shows the F2-score for different values of α across various top-k settings. The highest F2-scores are observed for $\alpha = 0.3$, with the top scores across multiple top-k settings, particularly for $k = 3, 4, 5, 6$, where the F2-score reaches its peak at 0.7787 and 0.7793. This suggests that $\alpha = 0.3$ is the optimal choice for maximizing the F2-score.

Table VI presents the results of administrative procedure retrieval using two methods, BM25, and an ensemble approach, across different top-k values. BM25 performs reasonably well at lower top-k values but starts to decline as more results are included. Even though BM25 achieves relatively good recall at higher top-k values, its precision drops too much, impacting the overall F2-score, especially at $k = 10$. The ensemble method achieves better F2-scores across all top-k values, with the highest F2-score at $k = 6$ (0.7793). This indicates that the ensemble method is more balanced, giving greater emphasis to recall, which aligns with the F2-score's focus on recall.

2. Administrative Procedure Answer Generation

We conducted an evaluation for this task using common metrics such as ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum, and BLEU. Results are shown in Table VII. Vinallama-7b-chat leads in most ROUGE metrics, particularly in unigram, bigram, and sequence overlap, indicating it generates answers closely resembling the reference texts in terms of word and sequence similarity, while Vistral-7b-chat performs best in BLEU score, suggesting it generates high-quality, fluent answers with good n-gram precision.

We utilized GPT-4 to evaluate the quality of answer generation for administrative procedures and the performance of our QnA system within the RAGAS metrics framework. Results are shown in Table VIII. Vistral-7b-chat leads with the highest correctness score (0.8009), followed by Vinallama-7b-chat (0.7648), and phoGPT-4b-chat lags behind (0.6341). This suggests that Vistral-7b-chat provides the most accurate answers, while phoGPT-4b-chat needs improvement in generating correct responses.

3. Administrative Procedure Question Answering

In this task, we use RAGAS[12] with 4 evaluation indicators for each component of the QnA system, such as Context Recall, Context Precision, Faithfulness, and Answer Relevance. For each question in the 202 questions in the test set, we retrieve top-k = 6 articles as input into the large language models to generate answers. Results are shown in Table IX.

Overall, Vistral-7b-chat outperforms the other models across all metrics since it has the highest context precision (0.76538), context recall (0.72835), faithfulness (0.7509), and answer relevance (0.72483). Faithfulness, which measures how accurately the answers reflect the information in the context, is highest in Vistral-7b-chat (0.7509), indicating its answers are the most reliable. The model also leads with the highest answer relevance (0.72483), showing that it provides the most pertinent and appropriate responses to questions.

Table IV
F2-SCORE RESULTS FOR DIFFERENT α AND TOP-K

α	$top-k$									
	1	2	3	4	5	6	7	8	9	10
1.0	0.6364	0.6512	0.6582	0.6334	0.6448	0.6420	0.6329	0.6319	0.6186	0.6029
0.9	0.6463	0.6644	0.6587	0.6651	0.6617	0.6551	0.6547	0.6437	0.6343	0.6219
0.8	0.6661	0.6719	0.6933	0.6931	0.6931	0.6859	0.6789	0.6662	0.6562	0.6463
0.7	0.7027	0.7181	0.7196	0.7139	0.7047	0.7045	0.6942	0.6827	0.6731	0.6616
0.6	0.7197	0.7628	0.7433	0.7405	0.7300	0.7241	0.7159	0.7095	0.6967	0.6810
0.5	0.7365	0.7546	0.7612	0.7577	0.7636	0.7590	0.7393	0.7256	0.7152	0.6951
0.4	0.7315	0.7628	0.7785	0.7781	0.7782	0.7785	0.7656	0.7446	0.7333	0.7132
0.3	0.7315	**0.7735**	**0.7787**	**0.7787**	**0.7785**	**0.7793**	**0.7729**	**0.7543**	**0.7365**	**0.7184**
0.2	0.7513	0.7710	0.7667	0.7673	0.7782	0.7783	0.7687	0.7500	0.7313	0.7180
0.1	**0.7563**	0.7660	0.7692	0.7673	0.7628	0.7620	0.7591	0.7406	0.7309	0.7133
0.0	0.7414	0.7503	0.7580	0.7639	0.7585	0.7550	0.7408	0.7335	0.7244	0.7075

Table V
RECALL PERFORMANCES FOR DIFFERENT TOP-K VALUES OF BM25.

Top k	500	100	50	10	3	1
Score %	99	97	92	83	75	63

Table VI
RESULTS OF ADMINISTRATIVE PROCEDURE RETRIEVAL

Method (top-k)	Precision	Recall	F2 score
BM25 (k=1)	0.6386	0.6386	0.6386
BM25 (k=2)	0.542	0.7037	0.6641
BM25 (k=3)	0.4991	0.7549	0.6847
BM25 (k=5)	0.4624	0.7871	0.6902
BM25 (k=10)	0.3809	0.83	0.6716
Ensemble (k=1)	**0.7376**	0.7310	0.7323
Ensemble (k=2)	0.7029	0.8102	0.7734
Ensemble (k=3)	0.674	0.8432	0.7787
Ensemble (k=4)	0.6592	0.8605	0.7784
Ensemble (k=5)	0.6481	0.8778	0.7784
Ensemble (k=6)	0.6367	0.8968	**0.7793**
Ensemble (k=7)	0.6178	0.9067	0.7729
Ensemble (k=8)	0.5785	0.9084	0.7543
Ensemble (k=9)	0.5355	0.915	0.7365
Ensemble (k=10)	0.4972	**0.9199**	0.7184

Table VII
RESULTS OF ANSWER GENERATION TASK

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	BLEU
Vinallama-7b-chat	0.5653	0.3910	0.4216	0.4889	0.2513
Vistral-7b-chat	0.546	0.3835	0.4087	0.4803	0.2533
phoGPT-4b-chat	0.4655	0.312	0.3511	0.404	0.1864

Table VIII
RESULTS OF ANSWER GENERATION TASK

Model	Answer similarity	Answer correctness	Faithfulness	Answer Relevance
Vinallama-7b-chat	0.828988	0.7648	0.743966	0.758991
Vistral-7b-chat	0.917168	0.8009	0.740286	0.726984
phoGPT-4b-chat	0.677705	0.6341	0.6802	0.836788

Table IX
THE RESULT OF END-TO-END QNA SYSTEM

Model	Context Precision	Context Recall	Faithfulness	Answer Relevance
Vinallama-7b-chat	0.72759	0.685632	0.67094	0.692839
Vistral-7b-chat	0.76538	0.72835	0.7509	0.72483
phoGPT-4b-chat	0.67919	0.64319	0.61489	0.56775

V. CONCLUSION

This study has demonstrated a QnA system for the Online Public Administrative Service Portal in Ha Giang Province, utilizing Large Language Models and contributing a comprehensive dataset that serves as a valuable resource for future research. Despite the potential of LLMs, their application in Vietnam's public administrative services remains underexplored. Our study addresses this gap by offering a systematic approach to integrating LLMs into public administrative frameworks. However, further research is needed to explore the broader implications of LLMs in diverse administrative contexts, including approaches like advanced RAG [14], or autonomous Agents [15]. Additionally, it is crucial to effectively address different question types within the QnA system, as each type requires tailored strategies for optimal handling.

REFERENCES

[1] J. Martinez-Gil, "A survey on legal question–answering systems," *Computer Science Review*, vol. 48, p. 100552, 2023.
[2] G. Caldarini, S. F. Jaf, and K. McGarry, "A literature survey of recent advances in chatbots," *CoRR*, vol. abs/2201.06657, 2022. [Online]. Available: https://arxiv.org/abs/2201.06657
[3] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," 2024.
[4] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2021. [Online]. Available: https://arxiv.org/abs/2003.01200

[5] A. Pham Duy and H. Le Thanh, "A question-answering system for vietnamese public administrative services," in *Proceedings of the 12th SoICT*, ser. SOICT '23, 2023.
[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *CoRR*, vol. abs/2106.09685, 2021.
[7] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *FTIR*, vol. 3, pp. 333–389, 01 2009.
[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805
[9] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019. [Online]. Available: https://arxiv.org/abs/1908.10084
[10] N. Q. Duc, L. H. Son, N. D. Nhan, N. D. N. Minh, L. T. Huong, and D. V. Sang, "Towards comprehensive vietnamese rag and llms," *arXiv preprint arXiv:2403.01616*, 2024.
[11] X. Amatriain, "Prompt design and engineering: Introduction and advanced methods," *arXiv preprint arXiv:2401.14423*, 2024.
[12] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," *arXiv preprint arXiv:2309.15217*, 2023.
[13] C. Nguyen, S. Luu, T. Tran, A. Trieu, A. Dang, D. Nguyen, H. Nguyen, T. Pham, T. Pham, T.-T. Vo, D.-T. Dol, N. Khang, H. Nguyen, N.-C. Le, T.-T. Le, Q. Bui, P. Nguyen, H.-T. Nguyen, V. Tran, and L. Nguyen, "A summary of the alqac 2023 competition," 10 2023, pp. 1–6.
[14] M. Eibich, S. Nagpal, and A. Fred-Ojala, "Aragog: Advanced rag output grading," 2024. [Online]. Available: https://arxiv.org/abs/2404.01037
[15] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on llm based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, Mar. 2024.

Dinh-Dien La is a PhD student majoring in computer science, University of Information and Communications Technology, Thai Nguyen University. He is currently Deputy Director of the Department of Information and Communications of Ha Giang province. His research interests are data science, machine learning, and deep learning.
Email: ladien.it@gmail.com

Tuan-Anh Nguyen is pursuing an Engineering degree in Information Technology at the University of Information and Communication Technology in Thai Nguyen, Vietnam. He is involved with the Institute of Applied Science and Technology in Thai Nguyen. His research interests include machine learning, deep learning, and natural language processing.
Email:dtc20h4802010242@ictu.edu.vn

Duc-Huy Mai is pursuing the B.Eng degree in Information Technology from Thai Nguyen University of Information and Communication Technology (ICTU). His research interests include software development, natural language processing (NLP), and question answering.
Email:dtc1954802010132@ictu.edu.vn

Thi-Thanh Ha received PhD degree in Information System in Ha Noi University of Science and Technology, Viet Nam. She obtained Bachelor Degree of Science in Applied Mathematics and Informatics from University of Natural Science, Vietnam National University in 2004. She currently is a lecturer at Computer Science in Thai Nguyen University of Information and Communication Technology. Her research interests are fields of deep learning in Natural Language Processing, Question Answering, and chatbot.
Email: htthanh@ictu.edu.vn

Assoc. Prof. Trung-Nghia Phung received his Engineering degree in Electronics and Telecommunications from Hanoi University of Science and Technology (HUST) in 2002. He completed his Master of Science degree in Telecommunications from Vietnam National University –Hanoi (VNUH) in 2007 and his PhD degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2013. He has been Rector of Thai Nguyen University of Information and Communication Technology (ICTU). His main research interests are signal processing and machine learning.
Email: ptnghia@ictu.edu.vn

Van-Khanh Tran received Ph.D. in Natural Language Processing from the Japan Advanced Institute of Science and Technology (JAIST), where his research focused on deep learning for natural language generation in spoken dialogue systems. contributed to the development of NLP applications. He is currently an AI Research Scientist on the NLP team at FPT Smart Cloud's Generative AI (GenAI) Center, where he focuses on developing large language models and AI assistant ecosystems tailored for Vietnamese users. He also serves as the Deputy Head of the Institute of Applied Science and Technology. His research interests include natural language processing, large language models, and AI applications in the legal, healthcare, and finance domains.
Email: tvkhanh@ictu.edu.vn