A Hybrid Approach to Enhancing Contextual Information for Vietnam Civil Code Question-Answering

Quoc-Dai Luong Tran[1*], and Hai-My Hoang Nguyen[2]
[1] Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology, Ton Duc Thang University
Ho Chi Minh City 700000, Vietnam
[e-mail: tranluongquocdai@tdtu.edu.vn]
[2] Faculty of Law, Ton Duc Thang University
Ho Chi Minh City 700000, Vietnam
[e-mail: nguyenhoanghaimy@tdtu.edu.vn]
[*]Corresponding author: Quoc-Dai Luong Tran

Received September 9, 2024; revised November 19, 2024; accepted January 1, 2025; published January 31, 2025

Abstract

Legal question-answering systems face significant challenges due to the complexity of legal language, the need for accurate statute interpretation, and the highly referential nature of legal documents. In practice, lawyers always refer to legal citations when making their arguments. Current models only rely on question-answer pairs during training without taking advantage of legal citations from official legal documents. This study proposes a hybrid approach to enhancing legal question-answering systems. The approach involves using retrieval models to extract contextual information from legal documents. A query mechanism retrieves relevant legal citations, which are then utilized to enhance the language model. This process enriches legal knowledge and improves the model's ability to generate and interpret legal content. Integrating the retrieved legal context with the fine-tuned language model lays the groundwork for training a question-answering system based on an Encoder-Decoder architecture. This ultimately improves its performance in delivering precise and legally sound answers. We developed a manually labeled question-answering dataset focusing on the Vietnam Civil Code to demonstrate the model's effectiveness in limited resource situations. Our approach minimizes the need for extensive training data while maintaining the model's ability to capture legal topics. Our experiments demonstrate that this hybrid architecture significantly improves performance in legal question-answering tasks compared to traditional models, particularly in scenarios where the language model lacks specific knowledge or when training datasets are limited.

Keywords: Deep neural networks, Information retrieval, Language model, Legal, Question-answering.

1. Introduction

The development of question-answering (QA) systems has received significant attention in natural language processing (NLP) due to their potential to automate information retrieval, knowledge extraction, and interactive problem-solving [1-3]. However, the complexity of legal language and the intricate nature of legal texts present irregular challenges in this domain. Legal documents are often characterized by formal language, specific terminology, and a high degree of intertextuality. This means multiple statutes, regulations, and case law can refer to and build upon one another. One of the main challenges in developing a reliable legal QA system is the need for accurate regulation interpretation and required references to legal documents. For a QA system to work correctly in the legal field, it must be able to interpret terminology in context and retrieve accurate information from a formal source [4].

Current language models, such as BERT [5], RoBERTa [6], GPT-3 [7] and T5 [8], are usually trained on general text and are designed to capture a broad understanding of language and general knowledge. While proficient at common tasks, they need expertise in specific domains such as law, medicine [9, 10]. In the legal domain, regular language models struggle to understand legal information. As a result, when utilized for highly specialized tasks, these models often yield overly simplistic or inaccurate responses due to their lack of necessary background knowledge or deep understanding required to navigate complex technical fields. This limitation emphasizes the importance of fine-tuning these models for specific domains or integrating external information retrieval systems to enhance their performance in specialized applications, such as QA systems about law.

Traditional QA systems typically match questions to answers based on patterns learned from labeled question-answer pairs during training. This approach works well for general-purpose tasks such as conversation or QA systems that provide essential information. However, it is ineffective when applied to specialized areas [11, 13]. In the legal field, the information needed to answer a question is often derived from formal legal sources. Furthermore, these models often need help accessing external information sources flexibly, forcing them to rely solely on knowledge encoded during training. In a field like law, where statutes and case law constantly evolve, the inability to retrieve and incorporate up-to-date information is a shortcoming. Therefore, traditional question-answering systems, which rely heavily on predefined question-answer pairs during training, need help to operate effectively in a legal context.

Recently, large language models like GPT-3 (Generative Pre-trained Transformer 3) have significantly excelled in understanding context, leading to remarkable performance in various NLP tasks. In 2020, OpenAI released GPT-3, which was further enhanced into GPT-3.5. This builds the basis for ChatGPT, a widely popular AI chatbot. These systems leverage vast datasets and advanced techniques such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) [14]. However, their success is largely due to their massive scale, requiring substantial computational resources accessible mainly to tech giants.

Unlike general-domain texts, legal documents often feature dense, domain-specific terminology and long dependencies, making accurate question-answering difficult. Our model incorporates enhanced contextual embeddings, enabling it to capture these nuances better. Our research focuses on creating models that can function effectively in low-resource environments with limited data and minimal computational power. The proposed model uses existing inputs and legal citations extracted from official legal documents to achieve this, which are required reference sources in legal activities. This approach allows the model to understand the user query better and generate more relevant answers, even with restricted resources and data availability.

While most current models focus on improving architecture and performance by relying on rich corpora, they often struggle in scenarios requiring high accuracy with limited data, such as legal documents. To overcome the above limitations, we suggest using a hybrid approach, which leverages the strengths of the retrieval-based model with fine-tuned language modeling for legal question-answering tasks. Retrieval-based models excel at finding and extracting relevant contextual information from extensive document collections, such as legal databases. Integrating a retrieval mechanism can enhance the QA system's ability to access relevant legal documents and provide accurate and contextually rich answers. Specifically, retrieval models can search extensive collections of legal documents to locate the most pertinent sources, such as statutes, regulations, or court cases, that may be pertinent to a specific question.

In this hybrid system, the retrieval mechanism first identifies and retrieves relevant legal texts based on the query. These texts are subsequently processed by a fine-tuned language model, which has been trained on a specific legal corpus. This fine-tuning process allows the model to better understand the nuances of legal language, interpret legal concepts accurately, and generate responses that are aligned with legal principles. By combining the retrieved context with the language model's capabilities, the system can generate more accurate and legally sound responses. The benefits of this hybrid approach extend beyond just improving accuracy. One significant advantage of this approach is that it decreases the dependency on large amounts of labeled question-answer pairs, which are often complex and expensive to obtain in specialized domains like law. By integrating retrieval methods, the model can leverage existing legal documents to provide the necessary context for answering questions, reducing the need for extensive manually annotated datasets. Hybrid models are more scalable and adaptable to different law subfields, where annotated question-answer pairs may be sparse or unavailable.

Additionally, the QA system provides exceptional flexibility and adaptability in handling new or previously unseen legal cases. Given that the legal system and statutes are subject to frequent revision and new case law is regularly established, the capability to retrieve the latest legal documents ensures that the model remains appropriate and precise. This feature is a crucial feature for legal QA systems; relying solely on static datasets trained on historical information can lead to outdated or inaccurate responses, particularly in cases involving recent legislative changes or court judgments.

Our approach differs from current research, primarily concentrating on designing model architectures for response generation. Instead, we prioritize retrieving and learning legal contextual information to significantly enhance the accuracy of the model's response generation. This research introduces a new combined approach to building legal QA systems that address the shortcomings of existing models. By integrating a retrieval-based approach with a fine-tuned language model and leveraging the Encoder-Decoder architecture, our system can retrieve relevant legal documents, interpret them in context, and produce accurate and legitimate responses. The proposed method improves the performance of legal QA tasks and reduces dependence on large datasets, making it an efficient and scalable solution. In this study, our main contributions are as follows:

- We propose a hybrid method that integrates retrieval-based techniques to extract contextual information from external legal document sources.
- This contextual information is leveraged to fine-tune and enhance the performance of language models in the legal domain.
- Combining retrieved-context with fine-tuned language models improves the performance of Encoder-Decoder-based question-answer systems, particularly when

dealing with small datasets.

- We develop a hand-labeled question-answer dataset specifically focused on Vietnam Civil code, providing a valuable resource for future research in legal NLP.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of previous research on conversational agents. Section 3 discusses the problem formulation, introduces the backbone framework, and details the model development. Section 4 presents the experimental results and provides an analysis of these findings. Finally, Section 5 concludes the paper.

2. Related Works

Contextual information is essential in training chatbots, enabling the model to generate more coherent, appropriate, and more like human responses. Without context, chatbots may produce generic or incoherent answers that lack depth. The recently created datasets Persona-Chat [15], Wizard-of-Wikipedia [16], and MultiWOZ [17] have been developed to include conversation history, user personas, or external knowledge. In these datasets, contextual information is essential in directing the flow of the dialogue, allowing the model to generate contextually appropriate responses. These datasets also help the model maintain conversational consistency, better understand user intent, and respond accurately by leveraging past interactions or external facts, accordingly improving the overall user experience.

Recent studies have aimed to improve conversation generation by integrating external knowledge from unstructured text sources and structured data [18, 19]. This combination enriches dialogue systems by providing relevant context and detailed information, leading to more informative and context-aware responses. In [18], the authors introduced a method that uses knowledge graphs to enhance conversation generation. The model retrieves relevant knowledge graphs based on a user's query and encodes them using a static graph attention mechanism to enrich the query's semantics. Additionally, during the decoder, a dynamic graph attention mechanism empowers the model to read the graphs and their knowledge, leading to a significant enhancement in response generation by integrating knowledge from the graphs.

Traditional retrieval methods like BM25, while effective for extracting relevant documents or text based on keyword matching, often fail to capture the nuanced and complex semantics inherent in legal language. This limitation results make these models may prioritize surface-level lexical matches over deeper contextual relevance. A study [20] introduces a data-driven model that utilizes a knowledge-based chatbot model which designed to generate more informative responses. It extends the traditional Sequence-to-Sequence (SEQ2SEQ) [21] framework by incorporating the dialogue's history and external information, enhancing the model's versatility and applicability in open-domain settings. The proposed approach shows notable improvements over a competitive SEQ2SEQ baseline, demonstrating its effectiveness in producing more content-rich and coherent responses.

Fine-tuning language models for specific domains is essential because pre-trained models on general data often need more specialized knowledge for domain-specific tasks [22, 23]. Research has shown that language models that are fine-tuned for specific fields - such as law, medicine, or engineering - achieve significant improvements in performance. This fine-tuning process allows the model to learn the unique semantics of each domain effectively. By adapting the model to the particular context of a domain, fine-tuning enhances its accuracy and ability to handle complex tasks, making it more effective in delivering relevant and contextually appropriate responses.

Recently, Lee et al. demonstrated that fine-tuning a pre-trained language model as BERT [5] on biomedical texts (BioBERT) resulted in substantial gains in accuracy for biomedical NLP tasks compared to models trained solely on general corpora [24]. This method improves the model's ability to provide contextually appropriate responses in specialized applications. In the legal domain, fine-tuning language models is essential due to the complexity and specificity of legal texts. Legal documents contain formal language, intricate terminologies, and case-specific precedents that general models struggle to interpret accurately. Legal-BERT [25], a language model based on BERT, has demonstrated that fine-tuning legal corpora improves performance on legal NLP tasks, such as contract analysis or legal question-answering. This fine-tuning allows the model to understand legal reasoning, interpret statutes, and reference previous cases, making it significantly more effective for specialized legal tasks.

Traditional methods for open-domain question answering often rely on knowledge graphs [26, 27]. The question is first processed by a semantic analysis model that predicts queries, guiding the interaction with the knowledge graph to find an answer. While effective, these approaches face challenges due to the independent training of each model. Potential errors accumulate and complicate further development and improvement, making optimizing and refining the system efficiently challenging. Sugathadasa et al. developed a legal text retrieval system using neural networks and experimented on a dataset of over 2,500 legal cases gathered from different online sources [28]. The proposed system employs deep learning techniques alongside a TF-IDF page ranking mechanism to create document embeddings, enhancing the efficiency of retrieving relevant legal texts.

Encoder-decoder models [25, 28], while powerful in generating fluent and coherent responses, face challenges in effectively utilizing retrieved context. Legal question-answering demands a high degree of precision, and the encoder-decoder architecture can struggle with distinguishing the most suitable details from extensive legal documents. Additionally, encoder-decoder models typically require large-scale annotated datasets for training, which are lacking in the legal domain.

A combined approach utilizing the BM25 algorithm and BERT has been proposed to improve the extraction of legal information and case law [29, 30]. This method effectively supports legal decision-making by accurately identifying and retrieving relevant case laws. The paper reports experiments conducted as part of the authors' participation in the COLIEE-2019 shared task organized at ICAIL 2019. The authors obtained encouraging results in all these four sub-tasks, demonstrating the effectiveness of their approach. A recent study introduced BB25HLegalSum [31], which combines BERT clustering with the BM25 algorithm to create summaries of legal documents. This method effectively selects, clusters, and highlights key sentences. It outperformed baseline models on the BillSum dataset and received positive feedback from legal professionals for its clear and highlighted presentation.

Unlike earlier studies that relied solely on retrieval-based methods for information extraction tasks, our approach utilizes these methods' capabilities to enrich the knowledge of deep learning models. Drawing inspiration from recent advancements, we integrate a retrieval-based model with an encoder-decoder deep learning architecture. The retrieval model effectively extracts contextual information, which is then used as input to enhance the performance and understanding of the deep learning model. Our model combines BM25's retrieval robustness with the capability of an encoder-decoder. By using BM25 to extract context and feed it into an encoder-decoder model, our approach balances computational efficiency and exploits contextual information. This hybrid strategy leverages BM25's ability to quickly identify potentially relevant legal citations and the encoder-decoder's capacity to process and generate precise, contextually rich answers. This integration addresses the limitations of individual components and offers a practical solution for legal question-answering tasks, particularly in low-resource settings.

3. Proposed Model

First, we present a formal definition of the question-answering problem. In this research, we approach answering questions by treating it as a problem where the dataset consists of pairs of questions and answers, denoted as $X = \{x_1, x_2, \dots, x_{|X|}\}$ and $Y = \{y_1, y_2, \dots, y_{|Y|}\}$. Developing a chatbot model can be formulated as a challenge of mapping inputs to their corresponding target responses. The model learns to map input questions (source) to corresponding answers (target) by analyzing patterns and rules within extensive training data. This process helps the model generate responses effectively. From a probability point of view, the model is trained to estimate the conditional probability of the target sequence Y based on the corresponding source sequence X.

$$p(Y|X) = p\left(y_1, y_2, \dots, y_{|Y|} \middle| x_1, x_2, \dots, x_{|X|}\right) \tag{1}$$

To address the limitations of existing question-answer models that typically rely only on question-answer pairs for training, we incorporate additional legal contextual information derived from authoritative legal documents. Integrating external context allows the model to understand legal information better. This approach enhances the model's ability to provide accurate and contextually relevant answers, particularly in specialized domains like law, where understanding precedents, statutes, and legal references is essential for generating reliable responses.

3.1 Building datasets

The civil code serves as the fundamental reference in judicial decisions. We create a comprehensive set of legal citations by analyzing legal documents and extracting essential content for each chapter. We organize the content by labeling every article and clause, ensuring that relevant legal principles are systematically categorized. This process allows quick access to pertinent legal information and enables more efficient referencing in legal judgments, enhancing the model's ability to align with legal standards in civil cases.

We developed two specialized datasets, hand-built by experienced legal professionals, to enhance the model's legal expertise. The first dataset $C = \{c_1, c_2, \dots, c_{|C|}\}$ is a comprehensive collection of legal citations extracted from the Vietnam civil code 2005. We have compiled over 1,000 legal citations in accordance with the structure of the civil code. Each citation is carefully organized by article and clause, making the most relevant legal references easily accessible. This dataset offers a structured framework that reflects how legal professionals reference sources of law, making it an essential resource for training the model to accurately navigate complex legal questions.

The second dataset is a manually triplet collection $(x_t, y_t, c_t)_{t=1}^{n}$ covering various legal issues, where (x_t, y_t) is the t^{th} pair of question and answer, while c_t is its legal citation. Built by lawyers with deep expertise, this dataset includes real-world legal inquiries and carefully crafted responses. We have constructed and labeled a dataset of 5,000 triple sets focused on content related to the civil code. This dataset is essential for training and fine-tuning models, as it offers a comprehensive legal citation that supports more accurate interpretation and response generation in legal question-answer systems. These two datasets form a robust foundation for training the model, integrating the law's theoretical framework and the practical application of legal knowledge to real-world situations.

3.2 Fine-tuning the Language Model with Legal Knowledge

Our study utilizes PhoBERT [32] as a pre-trained language model (Pre-trained LM) specifically designed for the Vietnamese language. It is based on RoBERTa [6], which optimizes the BERT pre-training procedure for more robust performance. This model has shown significant improvements over previous Vietnamese language models in tasks like part-of-speech tagging, named entity recognition, and machine translation, making it highly effective for various natural language processing (NLP) applications in Vietnamese.

General large language models (LLMs) are usually pre-trained on vast amounts of general-purpose data, which often lack the specialized knowledge required for specific domains like law. This gap can lead to limited performance in understanding and reasoning about legal matters. As shown in **Fig. 1**, we propose using the legal dataset to fine-tune the pre-trained language model, aiming to overcome this limitation. This process enhances the model's legal context knowledge, improving its ability to handle legal-specific tasks and deliver more accurate, informed responses in legal question-answering systems.

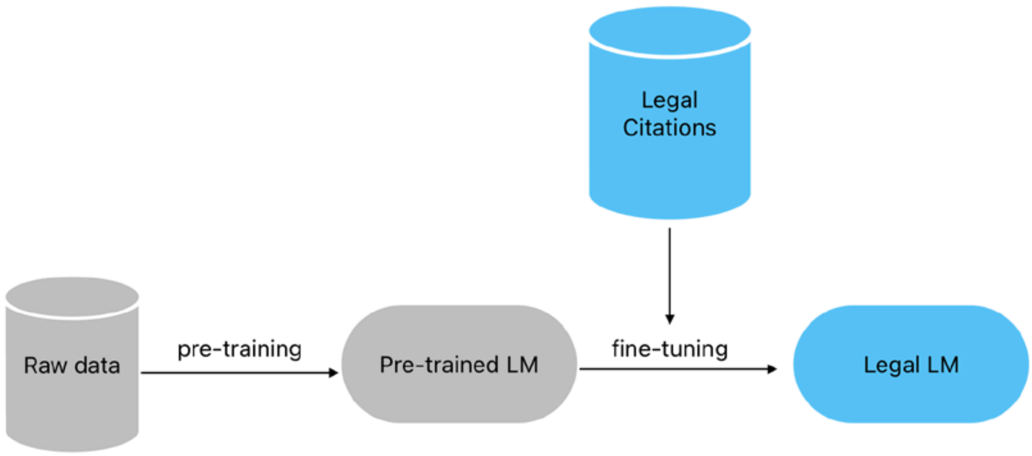

Fig. 1. Fine-tuning the pre-trained LM with Legal citations.

For each legal document, the tokenizer processes the legal contextual information into a sequence of tokens $w = \{w_1, w_2, \dots\}$. The objective function $f_\theta(\cdot)$ for this pre-training is optimized to enrich the model's knowledge of the legal language, thereby improving legal-specific text generation.

$$L_{LegalLM}(\theta) = E_{w \in C}\left[-\sum_{i} \log f_\theta\,(w_i | w_0, w_1, \dots, w_{i-1})\right] \qquad (2)$$

where $w_0, w_1, \dots, w_{i-1}$ represent the previous tokens, with x_i as the predict token and θ as the parameters of the pre-trained model $f_\theta(\cdot)$. To fine-tune the model, we optimize θ using the legal citation C. This fine-tuning process updates the parameters to produce a legal-oriented language model, improving the model's ability to understand and predict legal-specific content based on prior context.

3.3 Enhancing the Question-Answering Models with Legal-LM and legal citations

Our legal language model is built upon the BERT architecture, designed to handle data sequences, such as natural language text. This architecture has been successfully applied in various NLP tasks, including machine translation [33, 34], recommendation systems [35] and language modeling [36]. BERT's training process includes a next-sentence prediction task, enabling it to capture the relationships between sentences. This pre-training step is particularly valuable for tasks like question answering, as it allows the model to learn the context and sequence dependencies by predicting whether one sentence logically follows another in the text.

Our proposed response generation model is based on a transformer encoder-decoder architecture, which has demonstrated significant improvements across various tasks [8, 37]. However, these models typically require extensive pre-training on large datasets before being fine-tuned for specific tasks. Recent studies showed that skipping the costly pre-training phase by utilizing a pre-trained encoder can still yield competitive results in text generation tasks [38, 39]. Building on these experiences, we employ an Encoder-Decoder based model [40], initializing both the encoder block and the decoder block with our pre-trained Legal-LM.

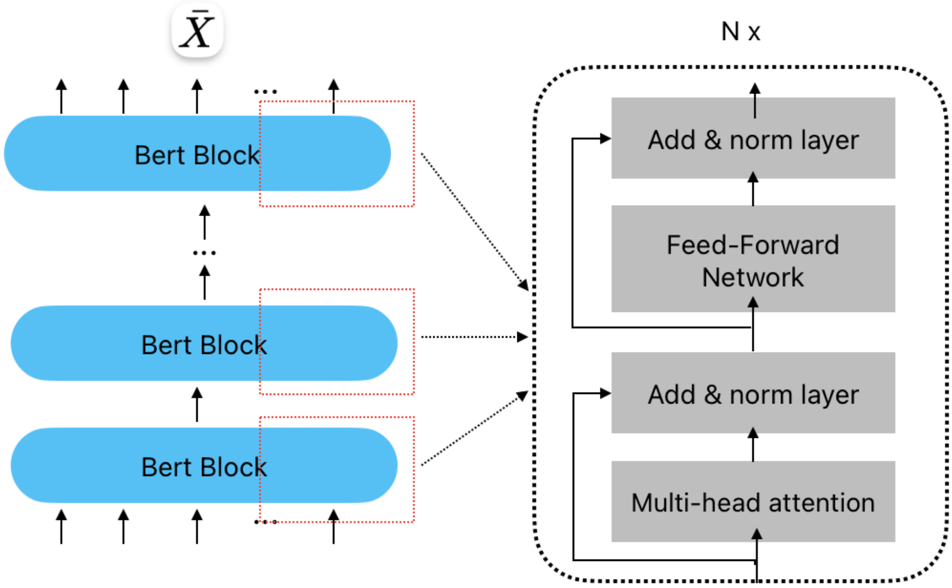

Fig. 2. The encoder architecture consists of multiple BERT blocks, with each block featuring a multi-head attention mechanism followed by a feed-forward network.

Humans often use the context in everyday conversations, just as lawyers refer to legal citations in their arguments. Our model overcomes the limitations of the traditional legal QA model by integrating legal citations, which provide relevant legal information for the user's query. The model takes the input message x_t along with its corresponding legal citation c_t and transforms them into a contextualized encoded vector $\bar{X}$:

$$\bar{X} = f_{\theta_{en}}([x_t, c_t]) \tag{3}$$

This combined sequence is passed through an encoder, which consists of multiple encoder blocks based on BERT. As you can see in **Fig. 2**, the figure illustrates that each encoder block includes a self-attention mechanism and two feed-forward layers. This architecture enables the model to effectively capture the immediate input and its context, which in turn facilitates better understanding and response generation.

The decoder's architecture mirrors the encoder's (as illustrated in **Fig. 3**) but with some minor variations. This block depends on the contextualized vector $\bar{X}$ and incorporates a cross-attention layer. Each decoder block is more complicated, comprising a unidirectional self-attention layer, a cross-attention layer, and two feed-forward layers. During training, the decoder transforms the vector $\bar{X}$ and the target utterance x_t into logit vectors L.

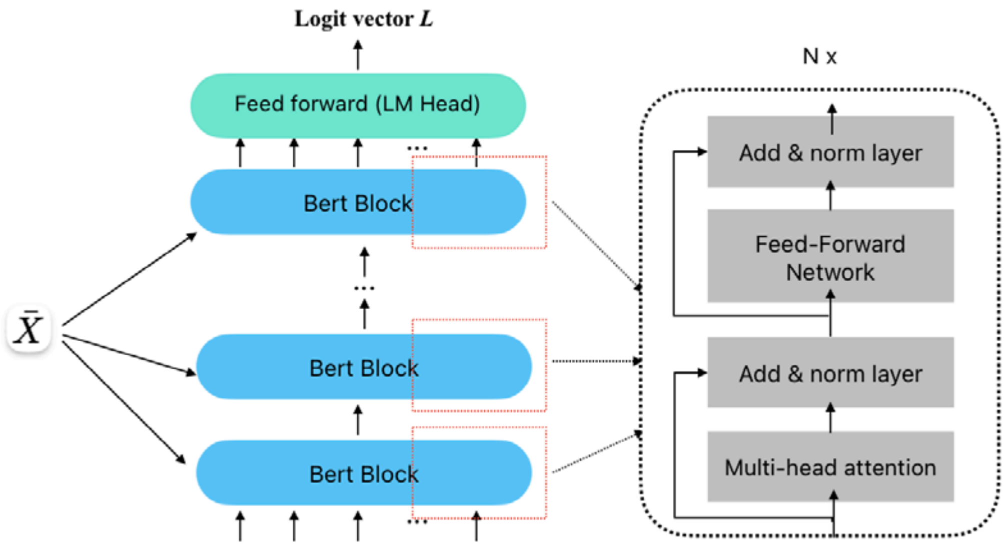

Fig. 3. The decoder architecture incorporates BERT blocks, enhanced with cross-attention layers positioned between the multi-head attention mechanism and the feed-forward network.

Given the target $y_t = \{w_1^t, w_2^t, \dots, w_{|y_t|}^t\}$, the probability distribution is estimated by calculating the conditional probabilities of predicting each subsequent word:

$$p(y_t|[x_t, c_t]) = p_{\theta_{de}}(y_t|\bar{X}) = \prod_{i=1}^{|y_t|} p_{\theta_{de}}\left(w_i^t \middle| w_{0:i-1}^t, \bar{X}\right) \tag{4}$$

The logits represent the probability distribution of the response y_t given the user input x_t. Consequently, a softmax function is applied to the logit vector to determine each generated word, as shown below:

$$p_{\theta_{dec}}\left(w_i^t \middle| w_{0:i-1}^t, \bar{X}\right) = Softmax(l_i) \tag{5}$$

where l_i represents the i^{th} token in the logit vector L. The training objective is to minimize the cross-entropy (CE) loss below:

$$L = -\sum_{i=1}^{|y_t|} \log p_{\theta_{de}} \left(w_i^t \middle| w_{0:i-1}^t, \bar{X} \right) \tag{6}$$

3.4 Extract Legal Citations for Response Generation

Recently, BM25 is widely used for information retrieval due to its robust handling of term frequency and document length normalization, making it ideal for extracting relevant information from large corpora [41]. In our work, BM25 is the foundational method for extracting legal context from legal citations. To address this, we utilize the algorithm BM25 as a probabilistic-based ranking function to retrieve relevant legal citations. BM25 scores documents based on the presence and importance of query terms using term frequency (TF) and inverse document frequency (IDF). This approach allows us to retrieve the most relevant legal texts contextually aligned with user queries. By leveraging BM25 in our legal extraction process, we can effectively narrow down a vast corpus of legal documents and keep only helpful information for further processing with more advanced models. There are three main components in BM25:

1. Inverse Document Frequency (IDF) is a measure that indicates the significance of a term within the context of an entire corpus. Its importance is determined by how rare the term is across the documents. Terms that appear less frequently are considered more informative. The calculation for the IDF component is as follows:

$$IDF(t) = log\frac{N - n(t) + 0.5}{n(t) + 0.5} + 1 \tag{7}$$

   where N is the total number of documents in the corpus, and $n(t)$ is the number of documents containing term t.
2. Term Frequency – $TF(t,D)$ is the number of times the term t occurs in the document D.
3. Document length normalization involves using a parameter called b. When the length of a document is significantly shorter or longer than the average document length, the usual formula may yield inaccurate results. If $b=1$, BM25 fully normalizes based on document length. Conversely, when $b=0$, document length is ignored, treating each term as equally informative regardless of the document's size.

The BM25 scoring function is defined as follows:

$$BM25(q, D) = \sum_{t \epsilon q} IDF(t) \cdot \frac{TF(t, D) \cdot (k + 1)}{TF(t, D) + k\left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \tag{8}$$

where q represents the query, with each term t in the query q; D is a document being evaluated for relevance with the query. $|D|$ is document D's length (in words), and *avgdl* is the average document length across the corpus. The parameter k controls term frequency saturation, regulating how much a single query term can influence a document's overall score.

An essential step in our preprocessing is removing stop words. Stop words refer to words that occur frequently but offer little semantic value. By excluding these terms, we reduce noise and help the retrieval model focus on terms that contribute meaningfully to the context of legal citations. Once preprocessed, user queries are processed through a component designed to extract legal citations. BM25, a probability-based retrieval method, ranks citations by their relevance to the query. Its effectiveness is particularly notable in legal texts, where the standardized structure of documents aligns well with its ranking approach.

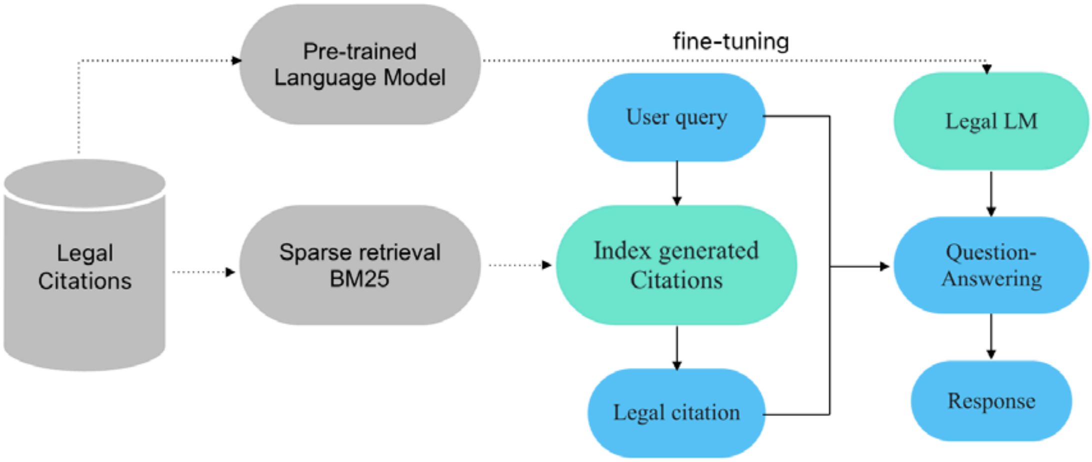

Fig. 4. Workflow of the proposed model: Integrating BM25 retrieval with a fine-tuned Legal LM for accurate, context-aware responses.

In our task, BM25 serves as an initial ranker, identifying candidate documents that likely contain relevant legal citations for the given query. **Fig. 4** provides an overview of our proposed model. It builds upon the previously fine-tuned Legal LM and Question-Answering model. The model leverages the results of legal citation retrieval as a legal foundation. These citation result is integrated to provide rich contextual information, helping the model understand the legal framework relevant to each user query. The algorithm integrates sparse retrieval BM25 and fine-tuned language modeling to effectively handle legal question-answering tasks. It begins with a database of legal citations, where BM25, a probabilistic sparse retrieval method, ranks and retrieves legal citations relevant to the user query. The databases are also passed into a pre-trained language model, fine-tuned specifically for legal texts to ensure an accurate understanding of legal terminology and context. The question-answering module, trained on the Legal LM, combines the information from the user query and the retrieved legal citations to generate a clear, accurate response that is contextually relevant to the legal domain. This approach combines the strengths of BM25 for efficient legal citation retrieval and neural language modeling for contextual understanding. This combination of legal citations and contextual understanding reinforces the model's ability to deliver more accurate, legally grounded responses.

4. Experiments

In this section, we introduce the experiment models, compare their outcomes, and offer a qualitative analysis of the results. We evaluate the proposed method through experiments, comparing its performance against baseline models. The evaluation uses the same dataset and metrics to ensure a fair comparison. In addition, based on the experimental results, we also analyze and highlight the strengths and limitations of the approach, providing insight into its effectiveness within the legal question-answering domain.

4.1 Quantitative Evaluation

Evaluating effectiveness is essential in the development of question-answering systems. We assess question-answering systems using two measures. The first measures focuses on how closely the system's performance matches human judgment in generating responses. The second evaluates the coherence and consistency of the generated responses within the question-answering model.

We utilized the BLEU (Bilingual Evaluation Understudy) score [42] as the first metric, which is a widely used method for evaluating dialogue systems. It compares consecutive n-grams in the generated response to those in a reference sentence, assigning scores based on the number of matching n-grams. The BLEU score reflects the word overlap between the generated and reference responses, making it helpful in measuring dialogue quality [43, 44]. Rising BLEU scores suggest that the generated response is becoming more aligned with the reference, making it increasingly similar to human responses.

We use a second measure, the ROUGE score (Recall-Oriented Understudy for Gisting Evaluation) [45]. Recently, this metric is commonly used in evaluating the quality of question-answer systems. ROUGE offers a collection of measures that evaluate the quality of a generated response by comparing it to reference responses crafted by humans. This metric examines the similarity between the chatbot's output and the ground truth response by analyzing n-gram overlaps, word sequences, and word pairs.

4.2 Experimental Results

We initially trained an Encoder-Decoder-based baseline model, a widely used approach for text generation tasks [40]. We also developed QA model utilizing the BM25 algorithm. This algorithm ranks relevant documents or citations based on their relevance to the query. Finally, the proposed model was developed to demonstrate the usefulness of incorporating legal citations to improve response generation in the legal QA task. Furthermore, to ensure consistency in the comparison between models, we also configure the same training parameters for the experiments. We set the batch size to 64 and fixed the learning rate at 2e-4. The models were trained using the Adam optimizer [46] to minimize cross-entropy loss.

In this experiment, we evaluated the model's performance using BLEU and ROUGE scores to assess the similarity between the generated responses and the ground truth utterances from the test dataset. Higher BLEU and ROUGE scores indicate better alignment with the reference responses, signifying improved model performance. **Table 1** and **Table 2** present our experimental results. We also provide an illustration of the scores achieved by the experimental results we trained, as shown in **Fig. 5** and **Fig. 6**.

Table 1. Experimental results with BLEU scores.

Models	BLEU 1	BLEU 2	BLEU 3	BLEU 4
Baseline model	0.450	0.301	0.264	0.207
BM25	0.510	0.341	0.305	0.249
Our proposed model	0.591	0.480	0.407	0.355

Table 2. Experimental results with ROUGE scores.

Models	ROUGE Precision	ROUGE Recall	ROUGE Fmeasure
Baseline model	0.330	0.297	0.301
BM25	0.351	0.301	0.324
Our proposed model	0.395	0.335	0.357

According to the results shown in **Table 1**, the improvements across all BLEU scores demonstrate that our model generates responses that are more closely aligned with the ground truth. When comparing the baseline model to the proposed model, we observed that incorporating legal citations significantly improved BLEU scores, ranging from approximately 31% to 71%. Specifically, BLEU-1 increased from 0.45 to 0.591, BLEU-2 from 0.301 to 0.48, BLEU-3 from 0.264 to 0.407, and BLEU-4 from 0.207 to 0.355. Specifically, BLEU-1 shows an increase of over 31%, reflecting better unigram precision, which indicates that our model better captures individual words in context. The larger gains in higher n-gram scores, particularly in BLEU-4, suggest that the proposed method generates contextually coherent sequences more effectively.

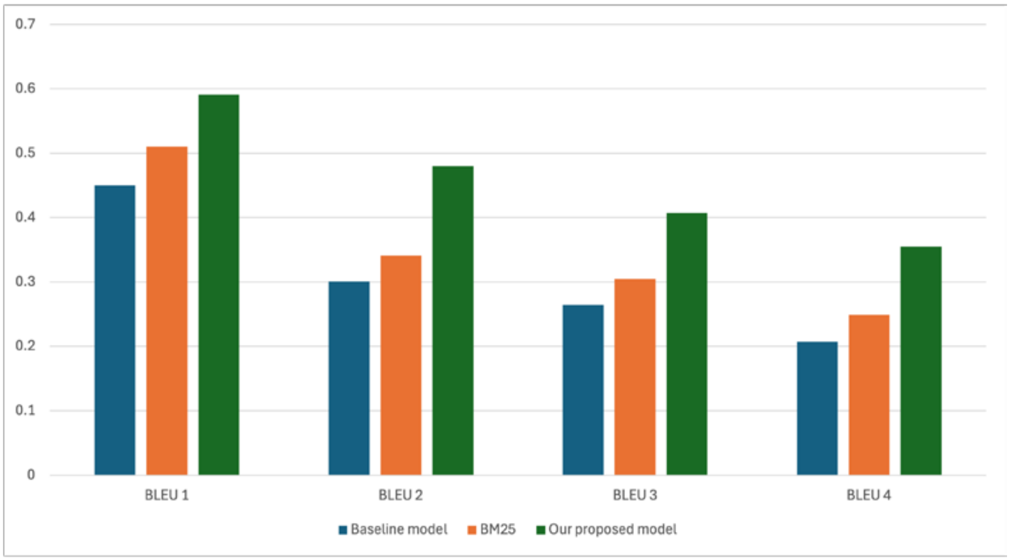

Fig. 5. BLEU score evaluation of experimental model performance.

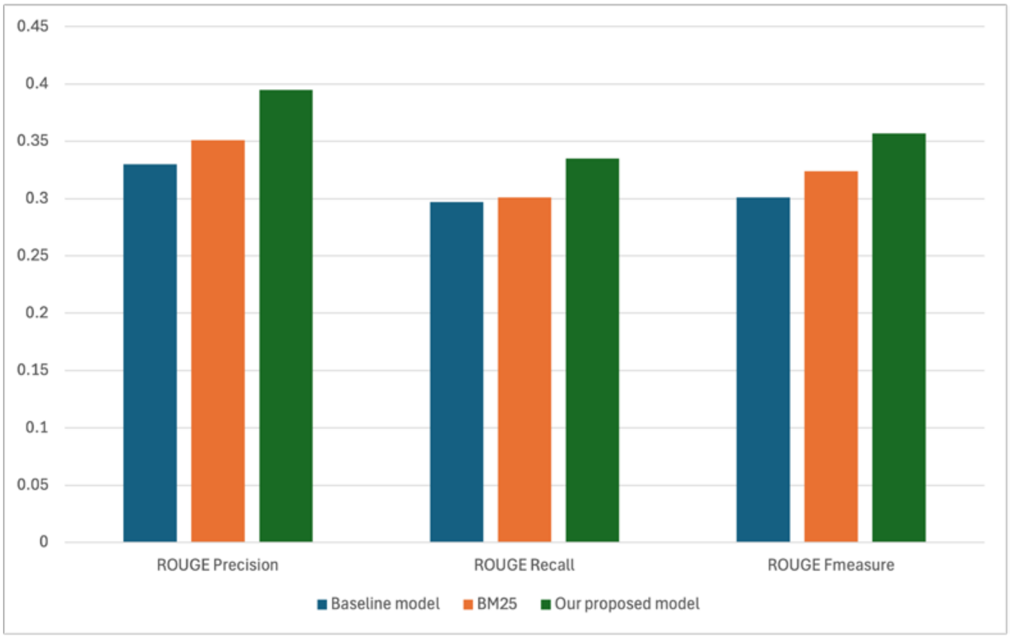

Fig. 6. ROUGE score evaluation of experimental model performance.

The proposed model shows a clear improvement over the baseline across all ROUGE metrics (as shown in **Table 2**). Precision has increased from 0.330 in the baseline to 0.395 in the proposed model, indicating greater accuracy in generating responses. Recall has also improved from 0.297 to 0.335, demonstrating the proposed model's enhanced ability to retrieve more relevant instances. Consequently, the F-measure rises from 0.301 to 0.357, reflecting a better balance between precision and recall. These improvements suggest that the proposed model more effectively generates accurate and comprehensive responses for legal question-answering tasks.

The results also show that the BM25-based model outperforms the baseline system. This is because our question-and-answer corpus is based on legal documents, with answers often partially derived from legal citations. However, it still performs worse than the proposed model in direct comparisons. BM25, a traditional query algorithm, has several limitations when applied to question-answering (QA) tasks. A significant issue in the algorithm that requires improvement is their shallow understanding of context. BM25 relies on term and inverse document frequency (TF-IDF) to match keywords between the query and documents. It does not fully capture the deeper semantic meanings or the relationships between elements of the text. It can retrieve relevant documents without fully understanding the nuance or context needed to answer the question accurately. A significant limitation of retrieval-based models is that they only extract information from a predefined dataset without generating new responses. This restricts their ability to handle questions that require interpretation or the synthesis of information not explicitly available in the dataset.

From these findings, we can conclude that our proposed model significantly outperformed the baseline across all evaluated metrics. The use of legal citations led to a significant improvement in handling legal questions. Based on the experimental results, we have observed that including legal citations as additional contextual information improves the accuracy and legal relevance of the question-and-answer model's responses. This aligns with how lawyers make references to legal citations in their practice. Additionally, integrating contextual information is advantageous for models trained on small datasets, addressing the challenge of limited resources commonly encountered in training question-answering systems.

5. Conclusion

This research presents a hybrid retrieval approach combined with deep learning techniques to improve question-answering tasks, specifically focusing on legal citation retrieval for legal QA. Our method incorporates retrieval mechanisms to obtain contextual legal information, which significantly enhances the quality of the responses generated. Experimental results showed that our proposed approach achieves notable improvements in both BLEU and ROUGE scores, confirming its effectiveness.

Our research aims to develop models that perform efficiently in low-resource settings with limited data and computational power. The proposed model leverages legal citations extracted from official legal documents, which serve as essential reference sources in legal tasks. By incorporating these citations, the model enhances its understanding of user queries and produces more relevant responses, even in environments with constrained resources and data availability.

We have also created a QA dataset for the Vietnam Civil Code through careful manual curation, which serves as a valuable resource for legal question-answering systems. Despite these advancements, our proposed model's performance remains dependent on the retrieval component. Future work will address this limitation by leveraging large language models to refine the retrieval process and expand the dataset further. These steps will enhance our model's overall robustness and accuracy, paving the way for more sophisticated and comprehensive legal QA systems. We intend to investigate more advanced query techniques, such as the Retrieval-Augmented Generation (RAG)-based model, which could further improve the model's performance in retrieving and generating precise legal information. Additionally, we aim to incorporate additional legal codes and domain-specific resources to make the model more comprehensive and adaptable to a broader range of legal contexts. We will explore deep learning models and reinforcement learning techniques to enhance integration and refine the dialogue model. These methods could allow for more efficient adaptation to new legal systems and facilitate the automation of model updates. Additionally, this could improve the model's performance for open-domain applications, where dynamic interaction and continuous learning from user inputs are essential.

References

[1] Sarkar, S., Singh, P., Kumari, N., Kashtriya, P., "The Task of Question Answering in NLP: A Comprehensive Review," in *Proc. of the International Conference on Recent Innovations in Computing, ICRIC 2022*, LNEE, vol.1011, pp.603-611, Springer, Singapore, 2023. Article (CrossRef Link)

[2] Zhang, Q., Chen, S., Xu, D., Cao, Q., Chen, X., Cohn, T., Fang, M., "A Survey for Efficient Open Domain Question Answering," in *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics*, vol.1: Long Papers, pp.14447-14465, Association for Computational Linguistics, Toronto, Canada, 2023. Article (CrossRef Link)

[3] Upadhyay, P., Agarwal, R., Dhiman, S., Sarkar, A., Chaturvedi, S., "A comprehensive survey on answer generation methods using NLP," *Natural Language Processing Journal*, vol.8, 2024. Article (CrossRef Link)

[4] Bench-Capon, T., Prakken, H., "Introducing the Logic and Law Corner," *Journal of Logic and Computation*, vol.18, no.1, pp.1-12, 2008. Article (CrossRef Link)

[5] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol.1 (Long and Short Papers), pp.4171-4186, Association for Computational Linguistics, Minneapolis, Minnesota, 2019. Article (CrossRef Link)

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *Proc. of the International Conference on Learning Representations (ICLR 2020)*, 2020. Article (CrossRef Link)

[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., "Language models are few-shot learners," in *Proc. of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, pp.1877-1901, Curran Associates Inc., Red Hook, NY, USA, 2020. Article (CrossRef Link)

[8] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol.21, no.1, pp.5485-5551, 2020. Article (CrossRef Link)

[9] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., Xie, X., "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology*, vol.15, no.3, pp.1-45, 2024. Article (CrossRef Link)

[10] Cuskley, C., Woods, R., Flaherty, M., "The Limitations of Large Language Models for Understanding Human Language and Cognition," *Open Mind*, vol.8, pp.1058-1083, 2024. Article (CrossRef Link)

[11] Shum, H.-y., He, X.-d., Li, D., "From Eliza to XiaoIce: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol.19, pp.10-26, 2018. Article (CrossRef Link)

[12] Lian, R., Xie, M., Wang, F., Peng, J., Wu, H., "Learning to Select Knowledge for Response Generation in Dialog Systems," in *Proc. of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp.5081-5087, Macao, 2019. Article (CrossRef Link)

[13] Priya, P., Varshney, D., Firdaus, M., Ekbal, A., "Knowledge-enhanced Response Generation in Dialogue Systems: Current Advancements and Emerging Horizons," in *Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pp.80-87. ELRA and ICCL, Torino, Italia, 2024. Article (CrossRef Link)

[14] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R., "Training language models to follow instructions with human feedback," in *Proc. of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, pp.27730-27744, Curran Associates Inc., Red Hook, NY, USA, 2024. Article (CrossRef Link)

[15] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J., "Personalizing Dialogue Agents: I have a dog, do you have pets too?," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, vol.1: Long Papers, pp.2204-2213, Association for Computational Linguistics, Melbourne, Australia, 2018. Article (CrossRef Link)

[16] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J., "Wizard of Wikipedia: Knowledge-Powered Conversational agents," in *Proc. of the ICLR 2019, International Conference on Learning Representations*, 2019. Article (CrossRef Link)

[17] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., Gašić, M., "MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling," in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp.5016-5026, Association for Computational Linguistics, Brussels, Belgium, 2018. Article (CrossRef Link)

[18] Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X., "Commonsense knowledge aware conversation generation with graph attention," in *Proc. of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18*, pp.4623-4629, AAAI Press, 2018. Article (CrossRef Link)

[19] Liu, S., Chen, H., Ren, Z., Feng, Y., Liu, Q., Yin, D., "Knowledge Diffusion for Neural Dialogue Generation," in *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, vol.1: Long Papers, pp.1489-1498, Association for Computational Linguistics, Melbourne, Australia, 2018. Article (CrossRef Link)

[20] Ghazvininejad, M., Brockett, C., Chang, M.-W., Dolan, B., Gao, J., Yih, W.-t., Galley, M., "A Knowledge-Grounded Neural Conversation Model," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 32, no.1, 2018. Article (CrossRef Link)

[21] Sutskever, I., Vinyals, O., Le, Q. V., "Sequence to sequence learning with neural networks," in *Proc. of the 28th International Conference on Neural Information Processing Systems, NIPS'14*, vol.2, pp.3104-3112, MIT Press, Cambridge, MA, USA, 2014. Article (CrossRef Link)

[22] Guo, K., Diefenbach, D., Gourru, A., Gravier, C., "Fine-tuning Strategies for Domain Specific Question Answering under Low Annotation Budget Constraints," in *Proc. of the 2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp.166-171, 2023. Article (CrossRef Link)

[23] Jeawak, S., Espinosa-Anke, L., Schockaert, S., "Cardiff University at SemEval-2020 Task 6: Fine-tuning BERT for Domain-Specific Definition Classification," in *Proc. of the Fourteenth Workshop on Semantic Evaluation*, pp.361-366, International Committee for Computational Linguistics, Barcelona (online), 2020. Article (CrossRef Link)

[24] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol.36, no.4, pp.1234-1240, 2020. Article (CrossRef Link)

[25] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I., "LEGAL-BERT: The Muppets straight out of Law School," in *Proc. of the Findings of the Association for Computational Linguistics: EMNLP 2020*, pp.2898-2904, Association for Computational Linguistics, Online, 2020. Article (CrossRef Link)

[26] Berant, J., Chou, A., Frostig, R., Liang, P., "Semantic Parsing on Freebase from Question-Answer Pairs," in *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp.1533-1544, Association for Computational Linguistics, Seattle, Washington, USA, 2013. Article (CrossRef Link)

[27] Skandan, S., Kanungo, S., Devaraj, S., Gupta, S., Narayan, S., "Question Answering System using Knowledge Graphs," in *Proc. of the 2023 International Conference on Inventive Computation Technologies (ICICT)*, pp.656-661, 2023. Article (CrossRef Link)

[28] Sugathadasa, K., Ayesha, B., de Silva, N., Perera, A. S., Jayawardana, V., Lakmal, D., Perera, M., "Legal Document Retrieval Using Document Vector Embeddings and Deep Learning," in *Proc. of the 2018 Computing Conference on Intelligent Computing*, AISC, vol.857, pp.160-175, Springer, Cham, 2019. Article (CrossRef Link)

[29] Gain, B., Bandyopadhyay, D., Saikh, T., Ekbal, A., "IITP in COLIEE@ICAIL 2019: Legal Information Retrieval using BM25 and BERT," in *Proc. of the 6th Competition on Legal Information Extraction/Entailment, COLIEE*, 2019. Article (CrossRef Link)

[30] Kim, M.-Y., Rabelo, J., Okeke, K., Goebel, R., "Legal Information Retrieval and Entailment Based on BM25, Transformer and Semantic Thesaurus Methods," *The Review of Socionetwork Strategies*, vol.16, pp.157-174, 2022. Article (CrossRef Link)

[31] de Andrade, L., Becker, K., "BB25HLegalSum: Leveraging BM25 and BERT-Based Clustering for the Summarization of Legal Documents," in *Proc. of the 14th International Conference on Recent Advances in Natural Language Processing*, pp.255-263, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023. Article (CrossRef Link)

[32] Nguyen, D. Q., Nguyen, A. T., "PhoBERT: Pre-trained language models for Vietnamese," in *Proc. of the Findings of the Association for Computational Linguistics: EMNLP 2020*, pp.1037-1042, 2020. Article (CrossRef Link)

[33] Yan, R., Li, J., Su, X., Wang, X., Gao, G., "Boosting the Transformer with the BERT Supervision in Low-Resource Machine Translation," *Applied Sciences*, vol.12, no.14, 2022. Article (CrossRef Link)

[34] De Coster, M., Dambre, J., "Leveraging Frozen Pretrained Written Language Models for Neural Sign Language Translation," *Information*, vol.13, no.5, 2022. Article (CrossRef Link)

[35] Shen, T., Li, J., Bouadjenek, M. R., Mai, Z., Sanner, S., "Towards understanding and mitigating unintended biases in language model-driven conversational recommendation," *Information Processing & Management*, vol.60, no.1, 2023. Article (CrossRef Link)

[36] Kurtic, E., Campos, D., Nguyen, T., Frantar, E., Kurtz, M., Fineran, B., Goin, M., Alistarh, D., "The Optimal BERT Surgeon: Scalable and Accurate Second-Order Pruning for Large Language Models," in *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp.4163-4181, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022. Article (CrossRef Link)

[37] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.7871-7880, Association for Computational Linguistics, Online, 2020. Article (CrossRef Link)

[38] Chen, C., Yin, Y., Shang, L., Jiang, X., Qin, Y., Wang, F., Wang, Z., Chen, X., Liu, Z., Liu, Q., "bert2BERT: Towards Reusable Pretrained Language Models," in *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*, vol.1: Long Papers, pp.2134-2148, Association for Computational Linguistics, Dublin, Ireland, 2022. Article (CrossRef Link)

[39] Naous, T., Bassyouni, Z., Mousi, B., Hajj, H., Hajj, W. E., Shaban, K., "Open-Domain Response Generation in Low-Resource Settings using Self-Supervised Pre-Training of Warm-Started Transformers," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol.22, no.4, pp.1-12, 2023. Article (CrossRef Link)

[40] Rothe, S., Narayan, S., Severyn, A., "Leveraging Pre-trained Checkpoints for Sequence Generation Tasks," *Transactions of the Association for Computational Linguistics*, vol.8, pp.264-280, 2020. Article (CrossRef Link)

[41] Robertson, S., Zaragoza, H., "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends® in Information Retrieval*, vol.3, no.4, pp.333-389, 2009. Article (CrossRef Link)

[42] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., "BLEU: a method for automatic evaluation of machine translation," in *Proc. of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pp.311-318, Association for Computational Linguistics, USA, 2002. Article (CrossRef Link)

[43] Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., "A Diversity-Promoting Objective Function for Neural Conversation Models," in *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologie*s, pp.110-119, Association for Computational Linguistics, San Diego, California, 2016. Article (CrossRef Link)

[44] Vinyals, O., Le, Q. V., "A Neural Conversational Model," in *Proc. of the ICML: 31st International Conference on Machine Learning*, 2015. Article (CrossRef Link)

[45] Kapočiūtė-Dzikienė, J., "A Domain-Specific Generative Chatbot Trained from Little Data," *Applied Sciences*, vol.10, no.7, 2020. Article (CrossRef Link)

[46] Kingma, D. P., Ba, J., "Adam: A Method for Stochastic Optimization," in *Proc. of the 3rd International Conference on Learning Representations, ICLR 2015*, 2015. Article (CrossRef Link)

Quoc-Dai Luong Tran achieved a bachelor's degree in Information Technology in 2009 and a master's degree in Computer Science in 2013, both from the National University of Ho Chi Minh City, Vietnam. In 2024, he completed his Ph.D. in Computer Science at Ton Duc Thang University (TDTU), Ho Chi Minh City, Vietnam. He is currently a lecturer in the Department of Computer Science and a member of the Natural Language Processing and Knowledge Discovery Laboratory at TDTU. His research interests include machine learning, reinforcement learning, natural language processing, and the application of artificial intelligence in the legal domain.

Hai-My Hoang Nguyen obtained her Bachelor of Laws and Master of Laws degrees from Ho Chi Minh City University of Law, Vietnam, in 2012 and 2017, respectively. She is currently a lecturer in the Faculty of Law at Ton Duc Thang University, Ho Chi Minh City, Vietnam, where she teaches undergraduate courses in Assets, Ownership and Inheritance, as well as Labor Law within the Department of Civil Law. Her research interests encompass collective labor agreements, workers' rights, legal studies, and the intersection of law and artificial intelligence.