INFORMATICA, 2026, Vol. 37, No. 1, 159–192

DOI: https://doi.org/10.15388/25-INFOR617

Ontology-Based Knowledge Graph Approach for Legal Queries

Vuong T. PHAM[1,3,4], Dang V. DUNG[2,3], Hung Q. NGO[5], Binh NGUYEN[1,3], Nhon V. DO[6], Ali SELAMAT[7,8], Hien D. NGUYEN[2,3,*]

[1] *Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh city, Vietnam*
[2] *University of Information Technology, Ho Chi Minh city, Vietnam*
[3] *Vietnam National University, Ho Chi Minh city, Vietnam*
[4] *Sai Gon University, Ho Chi Minh city, Vietnam*
[5] *University College Dublin, Dublin, Ireland*
[6] *Hong Bang International University, Ho Chi Minh city, Vietnam*
[7] *Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia*
[8] *Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Johor, Malaysia*
e-mail: vuong.pham@sgu.edu.vn, dungdv@uit.edu.vn, hung.ngo@ucd.ie, ngtbinh@hcmus.edu.vn, nhondv@hiu.vn, aselamat@utm.my, hiennd@uit.edu.vn

Received: October 2024; accepted: December 2025

Abstract. In the legal domain, ontologies organize legal concepts and their relationships, while knowledge graphs connect these concepts to specific entities in legal documents. This study proposes a solution for integrating ontology and knowledge graph, called Legal-Onto model, to construct a knowledge base of an intelligent retrieval system in the legal domain. The Legal-Onto model combines ontology as the conceptual layer and knowledge graphs as the implementation layer for representing the content of legal documents. This relational model is integrated with a structure of knowledge graph to identify relations between concepts and entities extracted from ontology in the determined domain. Moreover, this research addresses inherent challenges in semantic-based knowledge-driven search. The specific objective is to accurately extract relevant information from legal documents to respond to entered queries. The experimental results show that this method is more effective than state-of-the-art methods in natural language processing and large language models, which are without specific legal domain knowledge.

Key words: knowledge-based systems, knowledge graph, legal domain, information retrieval, knowledge representation.

1. Introduction

Investigating and obtaining legal information is critical in understanding and applying appropriate law within common contexts (Sartor *et al.*, 2022; Villata *et al.*, 2022). This process facilitates the rapid and precise assimilation of legal knowledge, safeguarding equity,

*Corresponding author.

efficacy, and adherence to legal norms. However, existing systems need help discerning semantic nuances within legal documentation and procuring pertinent legal information, necessitating a higher degree of precision to address user requirements adequately (Sansone and Sperlí, 2022).

Digital communication and infrastructure are propelling society forward at an accelerating pace through ongoing upgrades in the era of digital transformation (Chakraborty and Khan, 2023). The digitization of legal content to facilitate public understanding is a necessary requirement. Ontologies and knowledge graphs constitute efficacious methodologies for representing knowledge within the legal domain (Nguyen *et al.*, 2022, 2021; Moneus and Sahari, 2024). An ontology serves as a foundational framework to encapsulate the substantive content encapsulated within legal documents (Sartor *et al.*, 2022; Pham *et al.*, 2023). Building upon this foundation establishes a knowledge graph characterized by a collection of triples (subject, relation, object), wherein the subject and object entities are denoted as nodes, and their respective relations are demarcated as labelled edges within the graphical structure (Peng *et al.*, 2023). These legal knowledge graphs find applicability across various legal domains, including their utilization in legal case search engines (Junior *et al.*, 2020; Vuong *et al.*, 2023), recommendation systems designed to identify analogous cases (Dhani *et al.*, 2023), and question-answering systems (Sovrano *et al.*, 2020). However, there remains a lack of clarity regarding how ontology and knowledge graphs can be effectively integrated and applied to support accurate legal information retrieval, and such approaches have not yet been widely adopted in practice. Accordingly, this study aims to address the following research question: How can the integration of ontology and knowledge graphs be leveraged to improve the precision and efficiency of legal knowledge query systems?

This paper proposes a solution for integrating ontology and knowledge graphs to design a knowledge retrieval system in the law domain. This study delves into the natural language query processing stage from users. Furthermore, it implements improvements and expansions to enhance both the accuracy and processing speed of the knowledge representation method introduced in the previous publication (Pham *et al.*, 2023). Besides, Nguyen *et al.* (2022) demonstrated and introduced the effectiveness and correctness of this knowledge representation approach from legal texts. Specifically, the research utilizes a knowledge graph to encapsulate knowledge extracted from legal documents and delineate relationships between concepts extracted from those documents. Ngo *et al.* (2024) leverages knowledge representation methods and applies natural language processing techniques to optimize the knowledge graph, reducing its complexity and size. Those results in a rapid and efficient knowledge query method, providing accurate and effective answers. Additionally, this method contrasts with other large language models used in the legal domain, like ChatGPT[1] and Google Gemini[2].

The next section of this paper presents related work in the field of legal knowledge query, with a focus on Natural Language Processing (NLP) techniques and large language models (LLMs). Section 3 introduces the structure of a Legal-Onto model (Nguyen *et al.*,

[1] https://chat.openai.com/

[2] https://gemini.google.com/chat

2024, 2022), which represents the knowledge contained in legal documents. It also describes the structure of the knowledge graph that organizes this ontology. Section 4 identifies the problems associated with retrieval and querying in the legal knowledge domain and proposes solutions to these problems. Section 5 describes the experimental results of the proposed method on a dataset of legal documents related to traffic laws. These results are compared with those obtained from other popular LLMs, such as ChatGPT and Gemini. Finally, Section 6 concludes the paper by summarizing the main findings and outlining directions for future research.

2. Related Work

2.1. *NLP-Based Methods*

Conventional text retrieval systems mainly rely on the matching of terms between the query and documents using Using TF-IDF (Mishra and Vishwakarma, 2015; Manning *et al.*, 2009), BM25 (Robertson and Zaragoza, 2009). Furthermore, additional systems investigate word or document semantics; Word2Vec (Xia *et al.*, 2019) and Doc2Vec (Desai *et al.*, 2021) are two examples of such techniques. The development of deep learning methods for information retrieval (IR) includes neural network topologies such as transfer learning, recurrent neural networks (RNNs), convolutional neural networks (CNNs) (Soni *et al.*, 2023), and pretraining methods. In addition, attention-based methods, such as the Transformer architecture (Vaswani *et al.*, 2017), have been used to improve the ability of IR systems to focus on crucial parts of the query and documents during the matching procedure. Furthermore, it has been demonstrated that using pre-trained language models, such as BERT (Delvin *et al.*, 2019) and RoBERTa (Liu *et al.*, 2019), greatly improves the performance of IR systems. These models provide a deeper comprehension of the context and semantics of natural language texts and queries. Compared to employing the original models, fine-tuning the BERT and RoBERTa models according to certain data domains has shown improved efficiency (Hoang *et al.*, 2023).

Approaches based on deep learning have attracted a lot of interest in solving the challenge of answering questions. In Van *et al.* (2022), the problem is viewed as the extraction of answers using a pre-trained RoBERTa, and the model output consists of the beginning and ending places within an input sequence. Information retrieval (IR) models with varying variants in terms of sentence embeddings and document databases were used by HUKB (Yoshioka *et al.*, 2022). Nguyen *et al.* (2018) developed a recurrent neural network (RNN) to identify and label two essential parts of Japanese legal documents: the "required" part and the "execution" part. That study described the representation layers of the neural architecture and how the model components work. It proposed several RNN variations to identify necessary and complete parts of legal documents. It improved Bi-LSTM-CRF to recognize essential components that do not overlap. The paper then proposed two types of multi-layer models, Multi-layer-Bi-LSTM-CRF and Multi-layer-BiLSTM-MLP-CRF, to identify necessary overlapping components. The techniques outperformed previous methods and achieved state-of-the-art results on the JPL-RRE dataset (Japanese National Pension Law Requisite Efectuation Recognition dataset).

Mohan and Nair (2022) developed a system using deep learning techniques to categorize legal documents. This system has multiple layers that handle both finding important keywords and classifying the documents. First, it gathers legal documents from a dataset. After that, it uses a technique called Sammon Mapping to identify key terms based on how close they are to each other in meaning. Using these keywords, the system classifies the documents through quadratic discriminant analysis, which considers the probability of each category. Finally, the categorized documents come out as the system's output. This process repeats until the system achieves the most accurate classification possible. However, these methods did not represent the semantic relations between entities and concepts in law documents. Thus, the answer has not yet reflected the argument for it.

2.2. *Large Language Models*

Large language models (LLMs) constitute a type of deep learning model designed to handle natural language processing tasks. Their training hinges on utilizing massive datasets encompassing text and code, typically involving hundreds of millions to billions of parameters (Min *et al.*, 2023). These models are subjected to vast troves of text data from diverse sources, enabling them to grasp intricate linguistic patterns and semantic relationships (Naseem *et al.*, 2021). LLMs can generate text, translate languages, write creative content, and answer questions informatively. They can learn to perform many kinds of tasks in querying legal documents, including identifying relevant documents, summarizing documents, and extracting information (Maroudas *et al.*, 2022; Sikos, 2021). Ngo *et al.* (2023) used this information to answer questions about legal documents, such as the meaning of a particular clause or the implications of a court ruling.

Gemini is a LLM chatbot developed by Google AI (GoogleAI, 2024). It is trained on a massive dataset of text and code, and it can generate text, translate languages, and write different kinds of creative content. It can answer inputted questions as a chatbot. Gemini is useful in querying legal documents by searching for specific keywords or phrases in legal documents. Gemini can use its understanding of legal concepts to find relevant information, such as the definition of a term or the scope of a provision. It will return documents that discuss the elements of negligence, duty, and the standard of care. In addition, Gemini can be used to identify relevant sources of law and organize research materials. However, this program has limitations in querying legal documents. It only gives a general answer for an inputted query. It does not query the meaning of an article in legal documents.

ChatGPT is a large language model, which is developed by OpenAI. It is built based on the Generative Pre-trained Transformer (GPT) architecture (Liu *et al.*, 2023; Sejnowski, 2023). It is designed to generate human-like text based on the input it receives as a chatbot. In the domain of querying legal documents, ChatGPT can explain the meaning of legal terminology (Ajevski *et al.*, 2023). It provides explanations of various legal concepts and principles. In addition, ChatGPT also assists with legal research by formulating research queries and suggesting sources. It gives guidance on legal procedures. Nevertheless, the dataset of ChatGPT is only updated until 2021, preventing it from incorporating new legal knowledge for responding to queries in a more up-to-date manner.

LLaMA[3] (Large Language Model Meta AI) stands as a cutting-edge foundational large language model conceived to empower researchers in their endeavours to advance this subfield of AI (Meta, 2024). LLaMA's operational mechanism involves accepting a sequence of words as input and predicting a subsequent word, thus recursively generating text. Accessing to the model will be granted case-by-case to people with organizations and civil society (Mökander *et al.*, 2024). Users can implement it to become a chatbot supporting many domains. As with other LLMs, the function of LLaMA in querying legal documents is only to give definitions of legal concepts and principles from its collected documents. It cannot solve some law cases in practice. The current systems generally need the repository of legal documents as their semantics. They still need support to answer practical law cases.

2.3. *Ontology and Knowledge Graph*

Ontology is an effective method to organize the knowledge of legal documents (Gunkel, 2023; Satti *et al.*, 2023) and question-answering systems (Zouaoui and Rezeg, 2021). This method creates a model that captures the key ideas and how they connect, especially in legal documents (Martinez-Gil, 2023). While knowledge engineers design these ontology-based systems, they also need legal expertise to understand the specific area of law involved. Legal concepts are especially complex and interconnected (Allison, 2023; Zalesinska *et al.*, 2021). In this context, the applications that support residences in achieving a better grasp of the legal concepts expressed within legal ontologies is necessary (Martinez-Gil, 2023; Shao *et al.*, 2023). This will facilitate informed decision-making regarding the optimal ontology selection based on the target applications.

Besides, knowledge graph is a useful method for representing the relational knowledge domain. Ding *et al.* (2023) proposed semi-supervised method for knowledge graph construction, which was designed for decision support in operations and maintenance. However, the structure of that knowledge graph is not suitable for legal domains. Khan *et al.* (2022) demonstrated the effectiveness of knowledge graph embeddings for recommendations. They propose a deep collaborative alert recommendation (DCA) system to recommend region-specific COVID-19 against. Nevertheless, the structure of this graph is not enough for processing legal domain queries, which often require reasoning across multiple legal documents.

The integration of ontology and knowledge graphs is an effective method to solve these issues altogether (Abu-Salih and Alotaibi, 2024). Knowledge graph is a way to capture the intricate interrelationships prevalent in legal documents (Tang *et al.*, 2024). This is achieved by establishing nodes and edges, where nodes denote legal entities and edges signify their connections. This structured representation enables a more intuitive and comprehensive understanding of the legal landscape, facilitating tasks such as legal research, decision-making, and analysis. Moreover, the knowledge graph is constructed based on ontology, which can enhance legal information retrieval and analysis processes.

[3] https://ai.meta.com/blog/large-language-model-llama-meta-ai/

Oliveira and Oliveira (2024) utilized a graph based on the Resource Description Framework (RDF) to represent and search specific sections of legal documents. This RDF graph is built upon a well-defined legal ontology. This ontology allows the system to understand the overall structure of a legal system and the individual documents within it. By leveraging this ontology, the RDF graph can effectively capture the meaning of different parts of a legal document and the relationships between them. However, the proposed method did not solve the problem to optimize the graph for enhancing the searching on the described domain.

Based on that, this research aims to propose a solution to design the knowledge base of the legal domain by combining ontology and knowledge graph approaches. The proposed model includes a knowledge model of relations plays as an ontology to represent the content of legal documents. This model includes the relationships between legal concepts, similar to how an encyclopedia organizes information. Additionally, the proposed model integrates a knowledge graph structure to make connections between these concepts and specific entities extracted from the legal domain.

3. Ontology-Based Knowledge Graphs for Representing Relations in Legal Documents

3.1. *Ontology Representing the Knowledge of Legal Documents*

Ontology is an effective method for representing the knowledge of relations, especially relations between entities in the legal domain (Nguyen *et al.*, 2021; Sikos, 2021). In this study, ontology Legal-Onto is used to represent the knowledge of legal documents (Nguyen *et al.*, 2022). This ontology is built based on the knowledge model of relation integrating the structure of the knowledge graph to retrieve information for inputted queries (Nguyen *et al.*, 2021, 2023a).

Definition 1. Given a legal document d, ontology Legal-Onto for representing the document d is structured as follows:

$$\mathbb{K} = (\mathbf{Conc, Rel, Rules}) \oplus (\mathit{Keyphrases, Rela}),$$

- **(Conc, Rel, Rules)** is a structure of the Rela-model (Do *et al.*, 2018). In which, **Conc** is a set of concepts in the document d, **Rel** is a set of relations between concepts in **Conc**, **Rules** is a set of rules in the corresponding legal document. The structures of each element have been presented in (Nguyen *et al.*, 2022; Do *et al.*, 2018).

  Each concept $c \in$ **Conc** has the structure:

$$c := (\mathit{Name}, \mathit{Content}, \mathit{InnerRul}, \mathit{Attrs}, \mathit{Keyphrases}).$$

In which, the components of the concept are described in Table 1.

Table 1
The structure of a concept.

Elements	Meaning	Type	Condition
Name	The name of the concept in the law	String	
Content	The definition of the corresponding concept.	String	
InnerRul	Set of articles in the document pertaining to the relevant concept.	$\{a_1, a_2, \dots, a_n\}$, where $a_k = [Article_k, Para_k, Point_k]$ is a regulation $(1 \leqslant k \leqslant n)$	$[Article_k, Para_k, Point_k]$ a list of integers to locate a regulation in a legal document as Article – Paragraph – Point, respectively
Attrs	Set of components (or other concepts) building the corresponding concept (if necessary).	$\{c_1, c_2, \dots, c_q\}$ where, $c_j \in$ **Conc** is another concept $(1 \leqslant j \leqslant q)$	
Keyphrases	Set of keyphrases related to the concepts in each article.	$\{key_1, key_2, \dots, key_m\}$ where, key_i is a keyphrase $(1 \leqslant i \leqslant m)$	Each keyphrase may be a name of a concept: $key_i = c'.Name$ where, $c' \in$ **Conc**

Table 2
Kinds of facts.

Kind	Specification	Meaning
1	$[< r >$ is $< character >]$ $r \in$ **Rel** is a relation.	Demonstrate a characteristic of a relation.
2	$[< c_1 >< r >< c_2 >]$, $c_1, c_2 \in$ **Conc**	Connections among concepts.
3	$[< c >< r >< key >]$, $c \in$ **Conc** or $c \in$ *Keyphrases*; and $key \in$ *Keyphrases*	Relations between a concept and a keyphrase, or between two keyphrases.

Each binary relation $r \in$ **Rel** is one of three kinds:

$$\mathbf{Rel} := \mathbf{R}_{\text{hirer}} \cup \mathbf{R}_{\text{conc}} \cup \mathbf{R}_{\text{Phrases}}, \tag{1}$$

where, $\mathbf{R}_{\text{hirer}}$ is a set of relations "has-a" and "is-a" between concepts in **Conc**, $\mathbf{R}_{\text{conc}}$ comprises relations between concepts in **Conc**, connections like "a-part-of" and other binary relations not present in $\mathbf{R}_{\text{hirer}}$. $\mathbf{R}_{\text{Phrases}}$ is a set of relations between keyphrases within the legal document. This set also encompasses connections between keyphrases and a concept to show the meaning of the mentioned concept.

Each deductive rule $rul \in$ **Rules** concerning facts can be described as follows:

$$rul : \{f_1, f_2, \dots, f_m\} \longrightarrow \{g_1, g_2, \dots, g_n\},$$

where, $\{f_1, f_2, \dots, f_m\}$ denotes the hypothesis facts and and $\{g_1, g_2, \dots, g_n\}$ denotes the goal facts of the rule *rul*. There are three kinds of facts as the following Table 2.

- (*Keyphrases*, *Rela*) is a knowledge graph representing the relations between entities or keyphrases extracted from legal documents. In this framework, *Keyphrases* denotes

a collection of entities or keyphrases identified within the legal document, and *Rela* represents a set of arcs that depicts the relationships between entities or keyphrases.
- $Keyphrases := \{k \mid k$ is a keyphrase of the legal document$\}$
- $Rela := \{e = (k_1, k_2) \in Keyphrases \times Keyphrases \mid k_1$ are k_2 are keyphrases appearing in the same article of the law document $d\}$

- The symbol $\oplus$ means the combination of the structure (**Conc**, **Rel**, **Rules**) and the knowledge graph. In which, the Rela-model structure is the basis of the knowledge domain being built and the knowledge graph represents the relationships between entities based on concepts (**Conc**) and the arcs on the graph are the relationships specified in **Rel**.

The next section presents the structure of a knowledge graph building based on ontology in legal domain. **Conc** and **Rel** encapsulate concepts and relations to represent legal documents, while *Keyphrases* and *Rela* embody entities or keyphrases and relationships, respectively. These elements are extracted from legal documents and meticulously represented within the knowledge graph.

3.2. *Knowledge Graphs Based on Ontology in Legal Query Problems*

Knowledge graphs, after being built based on the content from legal documents to build the system's knowledge base or from questions entered by users, act as an intermediate knowledge layer. They support the matching and inference process to accurately retrieve relevant information from the knowledge base to answer questions effectively. They provide a structured, clear, and organized way of representing knowledge of legal documents. One of the main applications of knowledge graphs in the legal field is the ability to solve legal questions and answers. When users pose legal queries, knowledge graphs can effectively identify legal concepts and relationships related to the question, simplifying the information retrieval process. This capability significantly enhances the performance of legal research and improves the accuracy of answers provided by legal Q&A systems.

In this section, the improvement of the knowledge graph is proposed. The nodes of this graph are extracted from concepts of ontology Legal-Onto, combining keyphrases from the content of legal documents. In addition, the edges of this graph are constructed from relations between entities in the ontology and relations between keyphrases, which were labelled by experts in the legal domain. This knowledge graph also added the weight to measure the importance of relations between entities in the document.

Definition 2. Given a document law d. The structure of the knowledge graph representing the relations between key phrases in the document d is a tube:

$$(Keyphrases, Rela, weight),$$

where:

- *Keyphrases*: the set of keyphrases which was described as Definition 1, but it is improved as follows:

$$Keyphrases := K_1 \cup K_2, \tag{2}$$

where, $K_1 = \{c.Names \mid c \in \mathbf{Conc}\}$, which is the set of concepts in a legal document d, and K_2 is a set of key phrases of document d.
- *Rela*: set of relations representing edges of knowledge graph. There are two kinds of relations:

$$Rela := E_1 \cup E_2, \tag{3}$$

where, $E_1 = \{r(c_1, c_2) \mid r \in \mathbf{Rel}, c_1, c_2 \in \mathbf{Conc}\}$, which is the set of relations between concepts in a legal document d, and $E_2 = \{e = (k_1, k_2) \in K_2 \times K_2 \mid k_1$ are k_2 keyphrases appearing in the same article of the law document $d\}$.
- *weight* : *Keyphrases* $\longrightarrow \mathbb{R} \times \mathbb{R}$, which is a map to compute the similarly binary vector for each key phrase in *Keyphrases*. ($\mathbb{R}$ is the set of real numbers.)

  The degree of similarity between keyphrases is quantified by the term frequency-inverse document frequency TF-IDF measure, denoted as $(TF(k, d), IDF(k, d))$. Here $TF(k, d)$ represents the term frequency, indicating the prevalence of keyphrase k within document d. Conversely, $IDF(k, d)$ represents the inverse document frequency, capturing the specificity of keyphrase k in document d. The mathematical formulations of $(TF(k, d), IDF(k, d))$ are outlined as follows Manning *et al.* (2009), Le *et al.* (2019):

$$TF(k, d) := c + (1 - c).\frac{n_{k,d}}{\max\{n_{k',d} \mid k' \in Keyphrases\}}, \tag{4}$$

where, $n_{k,d}$ is the number of occurrences of the keyphrase k in the document d, and $c \in [0, 1]$ is a parameter, which is the minimum value for every keyphrase.

$$IDF(k, d) := \log\left(\frac{card(\bigcup_{a \in Article(d)} Phrase(a))}{1 + card(M)}\right), \tag{5}$$

where, $Article(d)$ is the set of articles of law document d, and $Phrase(a)$ is the set of keyphrases of article a in the document d, and the set $M := \{a \in Article(d) | k \in Phrase(a)\}$.

In this structure, the Rela-model is combined with the graph of keyphrases via keyphrases and their relations. When keyphrases are objects in land law, those relations between them are behaviours that were determined in law. Figure 1 illustrates the actual entity–relationship diagram of the knowledge base for a legal domain.

In the legal context, the use of knowledge graphs based on the Legal-Onto model allows the construction of knowledge structures based on clearly declared conceptual components and relationships. This helps integrate knowledge from various sources of documents such as laws, decrees, circulars and updated documents, while ensuring consistency in structure

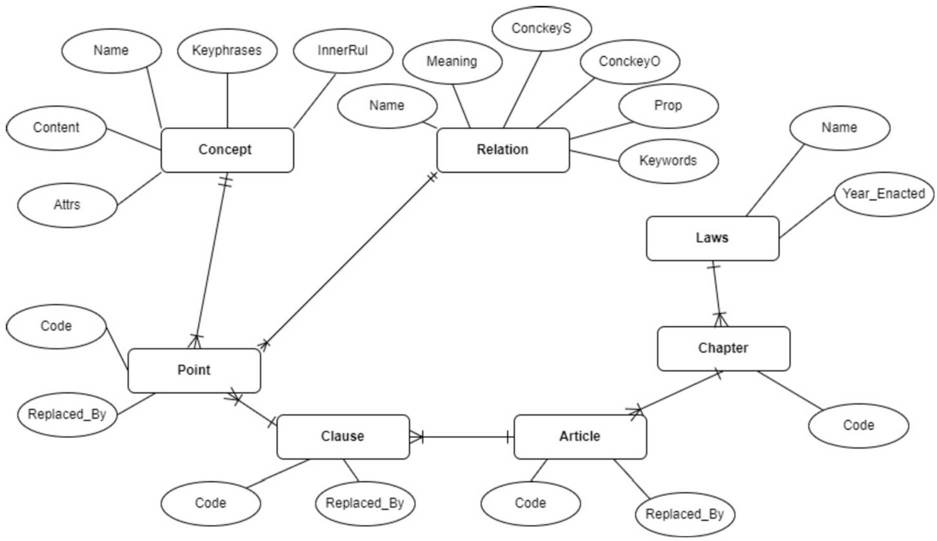

Fig. 1. Entity-relationship diagram of the legal system.

and content. Thereby, the system overcomes the problem of information fragmentation and inconsistency in the process of retrieving legal knowledge. This requires a high level of expertise and commitment from researchers and knowledge managers in the legal knowledge domain. However, the adaptability of knowledge graphs provides a flexible solution to address these challenges. Their easy adaptability allows for the seamless integration of legal knowledge from various sources into a unified graphical representation, reducing issues related to data fragmentation and ensuring data consistency. Therefore, once the legal knowledge base has been represented on an ontology-based platform and the content of knowledge documents is organized into knowledge graphs, updating and modifying knowledge graphs to add or change legal information is a quick and painless process for knowledge managers.

4. Legal Querying on Knowledge Graph

Recent studies on knowledge organization and representation based on the Legal-Onto model in (Dang *et al.*, 2025; Nguyen *et al.*, 2024) have shown the ability to effectively connct knowledge graphs with legal entities and the relationships between them. Legal contents can be represented accurately, flexibly and in accordance with the context of use. This capability makes them a natural choice for representing complex relationships between legal elements. Through their graph structure, knowledge graphs help display the connections between rules, laws, and court decisions. It creates a robust platform for research and inquiry in the legal field. Moreover, optimization in using knowledge graphs is an integral part of the typical legal research process, such as reducing graph size while maintaining content.

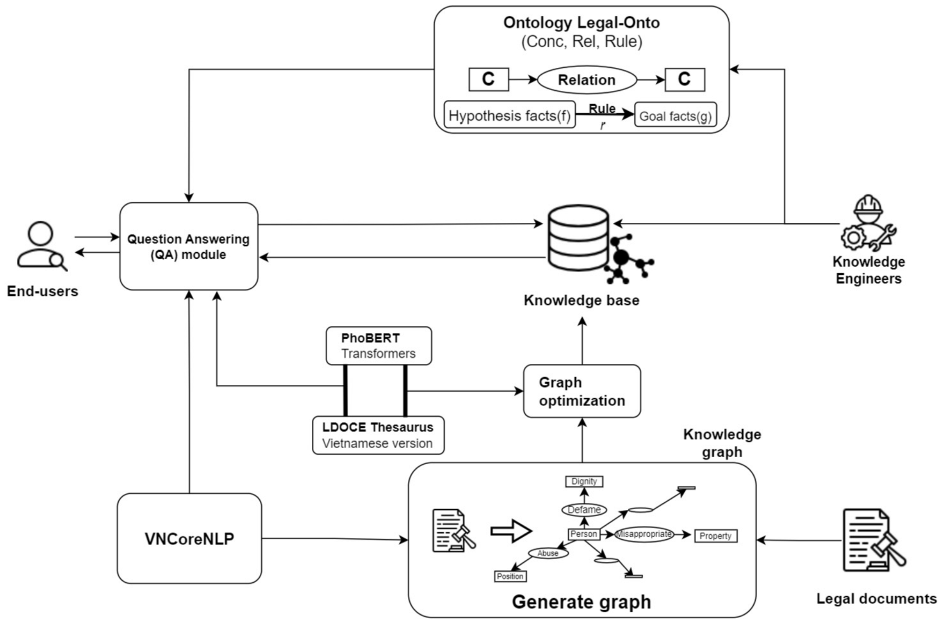

Fig. 2. Architecture of an intelligent query system in the legal domain.

Advanced methods, such as NLP, are employed to improve the performance in resolving legal queries. NLP automates the extraction of legal information from documents, reducing reliance on manual data entry and ensuring that the system remains up-to-date with the constant changes in the legal field. This section presents the problems of optimizing the legal knowledge graph and searching the knowledge domain for law documents.

4.1. *Knowledge Graphs and Legal Query Issues*

The intelligent search engine in the legal domain is designed based on the architecture of a search engine. However, in the legal field, high accuracy is an important factor because inaccurate information can affect the reputation of the system, and can even cause serious consequences. Therefore, the answers given to users need to have clear and specific grounds. To ensure this, the process of extracting, organizing, and representing knowledge from legal documents must be verified by experts in the field. The architecture of this system is shown in Fig. 2. It illustrates the process of an intelligent search engine, extracting essential information from legal documents to respond to a user-input query, there are two types of users for this system:

(a) **Knowledge Engineers:** This type of user takes on the role of checking the organized knowledge, as well as updating the knowledge base of the system. They are lawyers, legal consultants, or people with knowledge in their respective fields. Extraction is performed semi-automatically with the control and verification of this type of user. Finally, those components are organized under the Legal-Onto model.

(b) **End-users:** When users enter a query to the system, the system analyses the semantics of the query and classifies it into suitable content in this domain. The search engine will match the semantics of the query with the organized knowledge base to retrieve the relevant article in law documents that match the query's meaning. Finally, the system outputs results to users.

When an end-user inputs a query, the Question-Answering (QA) module of the system extracts entities and relations within the inputted query based on information from concepts and relations defined in the knowledge base according to the ontology model Legal-Onto. Simultaneously, PhoBERT[4] and VnCoreNLP[5] are integrated to support this process. From the extracted information, the system generates a graph representing the meaning of the query, formed from concepts and relations. Subsequently, the content of the knowledge base is localized and represented through a knowledge graph. These graphs are optimized to reduce the complexity of the query, enhancing the effective search capabilities of the system. Finally, the answer is extracted from the knowledge-matching part between the query's knowledge representation graph and the knowledge base.

For this process, two main problems need to be solved when searching for information on legal documents to answer an inputted query. These problems are:

• **Problem 1**: *Optimize the legal knowledge graph*. This problem is set up to reduce relations on the knowledge graph. It will enhance the performance for extracting information based on relationships between entities on the graph.

• **Problem 2**: *Searching on the legal knowledge domain*. This problem will represent the inputted query by a knowledge graph by extracting important phrases and relations. It may need to retrieve programming expertise from the knowledge base. Next, based on their semantic similarity, the created graph is matched with content in the search system's knowledge base. Comparing the matching is done with the structure of the knowledge model's constituent parts, particularly the relations and ideas and the structure of graphs. Further relationships connected to the query's content can be inferred using the knowledge model's inference rules.

4.2. *Building Legal Knowledge Graph*

The process of building a knowledge graph from legal documents is shown in Fig. 3. Data from legal documents is collected, pre-processed and classified into keywords to build a set of concepts (**Conc**), relations (**Rel**) and inference rules (**Rules**), thereby forming the legal ontology. The main tasks include:

- **Data collection and pre-processing:** removing noise, standardizing formats and checking information integrity.
- **Keyword extraction:** extracting keyword phrases from legal documents.

[4] https://github.com/VinAIResearch/PhoBERT
[5] https://github.com/vncorenlp/VnCoreNLP

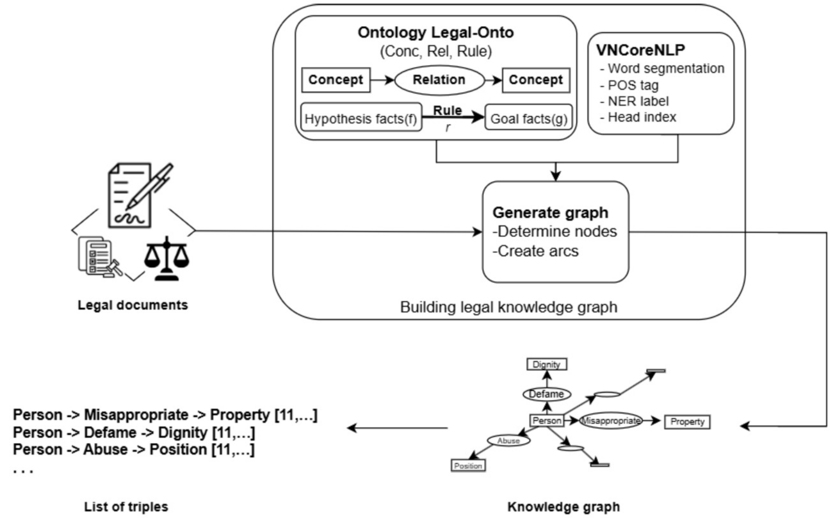

Fig. 3. The process of building a legal knowledge graph domain.

- **Keyword classification:** using the extraction results to build concepts in **Conc**, determine relations in **Rel** and establish inference rules in **Rules**.

Based on this ontology, the study continues to extract keyword phrases from legal documents. Keyword phrases are mapped into nodes and edges in the knowledge graph, with the support of the VNCoreNLP natural language processing model. The implementation process includes the following steps:

- **Keyword phrase extraction based on ontology:** taking advantage of keyphrases associated with concepts and relationships.
- **Natural language processing support:** using VNCoreNLP to increase extraction accuracy.
- **Mapping into the knowledge graph:** keyphrases associated with concepts into nodes; keyphrases representing relationships into edges, reflecting semantic connections.

The results of this process include the Legal-Onto model, which is formed from a set of concepts, relations and inference rules, and a knowledge graph with a flexible structure, organized and stored in the form of knowledge triples (Subject, Relation, Object). Each triple represents a directed relationship in the graph, reflecting the semantic position of knowledge in the text. This graph creates a solid knowledge base based on the ontology, effectively supporting semantic queries in the legal field.

4.3. *Optimization Problems Related to the Legal Knowledge Graph*

The phase to optimize the knowledge graph to improve its performance is based on the method for extracting and representing knowledge in a knowledge graph. By eliminating meaningless relations in the knowledge graph, those relations that appear repeatedly but do

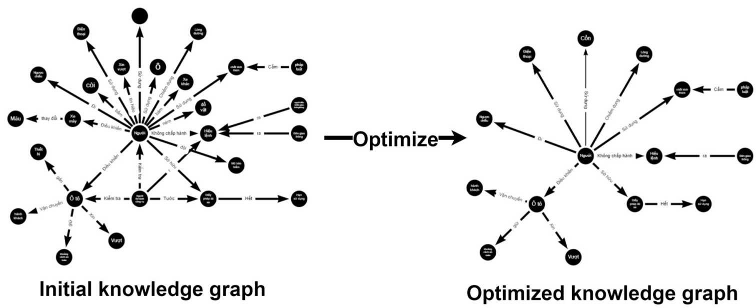

Fig. 4. The process of optimizing the knowledge graph.

not contribute to the unique content of the text, this optimization technique helps reduce the complexity and size of the text. Furthermore, the relationships that have the same meaning continue with the reduction. Knowledge graph optimization is the process of eliminating unnecessary triples unique to a document item to simplify and reduce the size of the graph. The relations are also found and merged to reduce repetition. Figure 4 illustrates the optimization process for a knowledge graph. Algorithm 1 optimizes the knowledge graph.

Equivalent relations are sets of three words, denoting subject, relation, and object, that possess matching or synonymous meanings. In order to determine the equivalence between two relations, each pair of subject, relation, and object values is scrutinized for identity or synonymy. If any element diverges, the comparison concludes, and the two triples are not deemed equivalent. Within the same graph representing the meaning of a Document Item, any equivalent triples are omitted to streamline the graph's size and complexity.

4.4. *Searching Problem on Legal Knowledge Domain*

The system will extract specified main keywords from a text or query before processing it to retrieve the knowledge. These keywords consist of two main parts: the user's intent and the query's entities. The knowledge model's component pieces, particularly the notion and relation components, are compared to determine how the matching mechanism operates. Other ties relating to the content of the query can be deduced with the use of inference rules. The system will then show the retrieval results for the entered query.

The legal retrieval procedure encompasses the conversion of the query into a representation within a knowledge graph, followed by exploring knowledge graphs that correlate with the representation of the input query. This process delineates two primary categories of input queries: conceptual queries and violation queries. The determination of query classification and semantics has been presented in (Sovrano *et al.*, 2020).

The knowledge graph, which encapsulates the semantic nuances of the inquiry, is composed of multiple triples featuring diverse subject types. Typically, numerous questions may need more details to formulate a comprehensive triple. Alternatively, in instances

Algorithm 1 Optimizing the knowledge graph

Stage 1: *Identify meaningless triples.*
Edges with low (TF, IDF) values that do not contribute specific content may appear in many document items throughout the triple-creation process, especially through automatic extraction. System administrators can use those values to determine whether to remove certain (TF, IDF) values from the graph that depict the significance of the passage. The administrator lists and removes edges that are judged unnecessary from the graph.
Step 1.1: Compute TF, IDF of triple k
For each triple k in document d **do**
+ **Compute** $TF(k, d)$ by the formula (4).
+ **Compute** $IDF(k, d)$ by the formula (5).
Step 1.2: Evaluate the meaning of a triple k through (TF, IDF)
If $TF(k, d) * IDF(k, d) \leqslant \alpha$ **then** k is marked as meaningless,
where α is a constant, which will be determined through the particular knowledge domain. Through gathering synonymous word clusters in Vietnamese Road Traffic Law and verified by domain experts, we choose $\alpha = 0.15$.
Step 1.3: The knowledge engineer verifies the meaningless marked relations to decide for deleting this relation or not.
Stage 2: *Optimize equivalent relations.*
This stage verifies equivalent relations and improves them. It begins by finding synonymous keyphrases with the same or nearly the same meaning based on their (TF, IDF) values. Then, it calculates their similarity using cosine similarity (Wang and Dong, 2020). The cosine similarity is determined using PhoBERT to turn the keyphrases into vectors if they are not part of any value fields in the thesaurus.
Step 2.1: Verify whether two essential phrases in the LDOCE Vietnamese thesaurus belong in the same value field. If so, they are regarded as interchangeable.
Step 2.2:
Compute $similar(k_1, k_2)$ by cosine similarity (Wang and Dong, 2020),
If $similar(k_1, k_2) \geqslant \beta$ **then** k_1 and k_2 are semantically similar.
where, $similar(k_1, k_2)$ is the measurement of similarity between two keyphrases k_1 and k_2; β is a constant, which will be determined through the particular knowledge domain. Through gathering synonymous word clusters in Vietnamese Road Traffic Law and verified by domain experts, we choose $\beta = 0.6$. For example, some interchangeable word pairs include *comply* and *obey*, *drive* and *operate*.

where the query omits the specification of the subject or object values, placeholders such as "*" are employed to signify that these elements are not fully defined in the question.

Given the knowledge base K for law documents as ontology Legal-Onto and an input query q, the following algorithm finds relevant knowledge for query q. Algorithm 2 finds relevant knowledge.

After obtaining a list of IDs satisfying each relation in the star graph in Step 2, the algorithm intersects these values. The result is a set of subgraphs within the system's knowledge graph that match the star graph. Figure 5 describes the intersection of star graphs to get results.

- Retrieve a compilation of corresponding IDs for each wildcard element within the knowledge representation graph associated with the question.

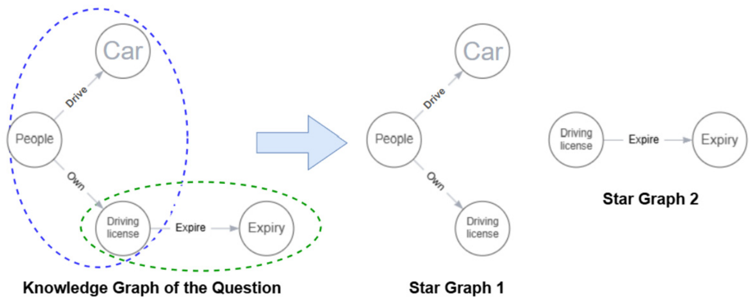

Fig. 5. The process of decomposing a question graph into star graphs.

Algorithm 2 Finding relevant knowledge.

Input: A knowledge graph representing the meaning of query q.
Output: Search the content in Knowledge base K being suitable for the knowledge graph of query q.
Step 1: Decompose the question graph.
The graph representing question semantics consists of multiple triples and has different types of subjects. They are decomposed into subgraphs into star graphs (Casillas *et al.*, 2013), which is similar to breaking down a question into sub-questions with each star being triples with a common subject.

- The graph is decomposed into star graphs.
- Each star represents triples sharing the same subject.
- This decomposition helps subdivide the question into sub-questions.

For example, given the user's question 'What is the penalty for driving a car with an expired driving license?', the content of the question is first mapped into a knowledge graph, where entities and their relations are represented as nodes and edges. From this initial graph, the system decomposes it into subgraphs, as illustrated in Fig. 5. The graph on the left represents the full content of the user's question, while the two star-shaped subgraphs on the right capture: (1) the triples sharing the subject "driving a car with a driving license", and (2) the triples related to the condition "the driving license has expired".
Step 2: Find knowledge that matches the star graph.
For each decomposed star graph, this algorithm systematically searches for lists of subgraphs within the knowledge base that share semantic equivalence. It then records the corresponding ID values.
Details of the matching and subgraph-finding process:
Let S be a set of star graphs that have been created in Step 1.
For each relation k in $Rel(S)$ **do** $Rel(S)$ is a list of relations in S:

- Search all relation k' in K.**Rel** such that $similar(k', k) \geqslant \beta$.
- Let list ID_k be a set of ID values of k'.

End for;
Note: The searching of k' is noted as follows:
If there is an "is-a" relationship between a component in the question's knowledge graph and a component in the knowledge base, the algorithm also accepts the corresponding knowledge component.
If any component of the triple in the query is missing and replaced with the character "*", the matching process relies on the remaining components in the triple.
Step 3: Take the intersection set of the answer sets of the star graphs.

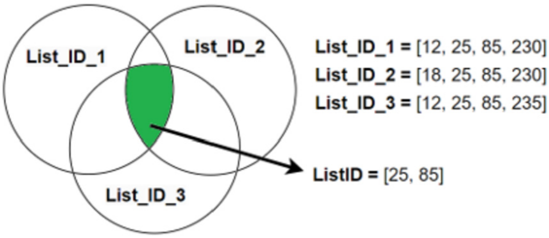

Fig. 6. The intersection of star graphs to get results.

- Determine the common elements among these sets by computing their intersection.

$$Res := \bigcap_{k \in Rel(S)} ID_k \tag{6}$$

- The list of IDs obtained as a result represents the answer to the input query q (see Fig. 6).

While traversing triples to discover semantically equivalent counterparts, the matching process depends on the remaining components if any element within the question's triple is absent. Furthermore, if the knowledge graph encompasses components that share an "is-a" relationship with corresponding elements in the knowledge base, the algorithm considers these associated knowledge elements. For instance, "motorbike" might have an "is-a" relationship with "motorcycle", "moped" and so forth.

4.5. *Advantages of the Integrating Ontology Legal-Onto and Knowledge Graphs*

With the structure of the Legal-Onto model as Nguyen *et al.* (2023b), it was built on three main components that allow the organization and representation of legal knowledge in a semantically clear and logically consistent manner. In which, Concepts formed to represent entities in the knowledge domain, Relations describe in detail and specifically the connection between entities and Rules are logical rules that support the reasoning process. This organization is highly formal, thereby helping both humans and computers to easily store and process. In addition, from the detailed organization of conceptual components and relations combined with Rules, it helps support reasoning based on available information, ensuring that the results are as expected by the user. They easily support the ability to expand and update knowledge when editing or supplementing without having to rebuild from scratch. From the above foundation as well as the presented studies in (Do *et al.*, 2018; Nguyen *et al.*, 2024, 2023b), it can be seen that the ability to organize and represent legal knowledge accurately and flexibly.

The structure of the knowledge graph helps to represent the content of legal documents in the form of nodes and the relationships between them (Hogan *et al.*, 2021). This allows the integration of legal data from various sources into a unified structure that ensures the requirements of reality to represent legal content comprehensively and closely linked in the legal system. Furthermore, KG is capable of automatically answering specific legal

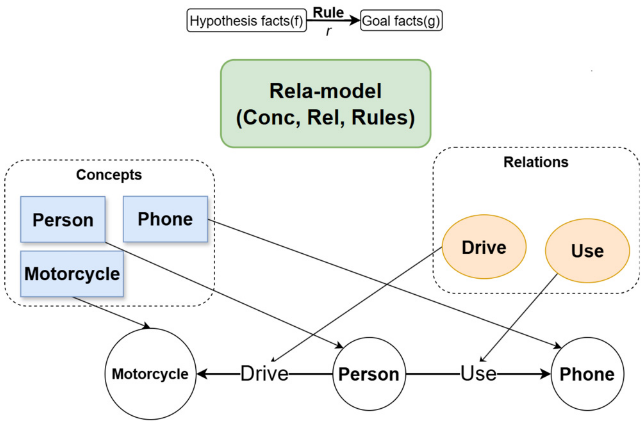

Fig. 7. Legal-Onto and knowledge graph representation for a legal statement.

queries using the extensive network of structured legal knowledge and relationships existing in the system (Dhani *et al.*, 2023). When a user asks a specific legal question, the Knowledge Graph scans their structured legal knowledge to search for and locate relevant semantic results in the knowledge base. In doing so, the Knowledge Graph provides accurate answers to queries and is particularly good at dealing with complex questions that require a lot of information and specific conditions.

With the ability to organize and represent precise information of Legal-Onto and the flexibility and efficiency in querying of the Knowledge Graph, the integration of these two components creates a powerful and comprehensive legal knowledge base in both knowledge organization and representation. With the nodes on the graph being mapped and constructed based on concepts and similarly the arcs on the graph being formed from declared Relations based on the Legal-Onto model. Rules components supporting the ability to reason on graphs have been built, helping to improve the efficiency of reasoning and querying, ensuring that answers are explained clearly and accurately. This is a solid foundation to support the construction of a system that supports querying and consulting legal information based on issued documents.

Example 1. The information segment: "Traffic participants' behaviour of using phones while driving motorbikes, mopeds, motorbike-like vehicles and moped-like vehicles", the ontology model helps to clearly define concepts such as "Person", "Motorcycle", "Phone" and the relationships "drive" and "use". From there, the model builds an accurate knowledge representation, fully representing the content of the text segment as shown in Fig. 7. This not only effectively supports the information retrieval process, avoiding duplication, but also easily expands or updates knowledge without having to rebuild the entire knowledge graph.

5. Experimental Results

The proposed method has been tried and tested on Vietnam's road traffic legal documents. With the current method, it can be seen as a flexible approach applicable to various knowledge domains, and it may even function with languages other than Vietnamese. However, when transitioning to a new knowledge domain, specialized knowledge in that field is required to carry out information extraction. Additionally, when switching to other languages, the system also needs to utilize alternative natural language processing applications instead of PhoBERT and VnCoreNLP, as these libraries are specific to the Vietnamese language. Nonetheless, other processes such as the formation of the knowledge graph or graph matching to yield results can be maintained unchanged.

The current method still faces several challenges. Although the system has utilized PhoBERT and VnCoreNLP to support the process of constructing the knowledge base, transforming information from legal documents into the knowledge base still requires the intervention of individuals with in-depth knowledge of the field to make final decisions, and this process takes a considerable amount of time. Furthermore, as the system expands to serve various legal domains, this process becomes more complex, and leveraging existing knowledge becomes more challenging, as each field employs specialized language and unique knowledge. Moreover, the current graph matching method is limited to processing individual questions and lacks the capability to address complex legal queries, where legal regulations overlap with each other.

In this study, the research team utilized state-of-the-art methods to test with the collected queries, including natural language processing methods. These methods delve deeply into analysing the syntax and semantics of words and sentences in legal documents, in order to provide meaningful answers suitable for the user's question. In addition, the study also tested LLMs such as ChatGPT and Gemini, currently in widespread use. The results of the experimental process are compared, reflecting the advantages of the method presented by the research team compared to well-known methods in practice.

5.1. *Testing on Vietnam Traffic Law*

The proposed method has been tested and experimented on Vietnamese traffic law documents in this section. The study utilized a dataset gathered from various sources, including 36 articles, 306 clauses, and 762 points pertaining to traffic offenses and their corresponding punishments as outlined in traffic laws (Vietnam National Assembly, 2008; Vietnam Government, 2019; Vietnam National Assembly, 2021; Vietnam Ministry of Transport, 2019). Based on this extensive data, the researchers used a knowledge representation framework to establish 90 distinct concepts and 49 relations between concepts in the road traffic legal and regulatory framework. This process resulted in the creation of 616 nodes and 1239 relations between edges on the road traffic legal domain graph, which form the underlying structure of the simulated knowledge graph as shown in Fig. 8. This knowledge graph effectively captures the essence of traffic violations defined in road traffic laws.

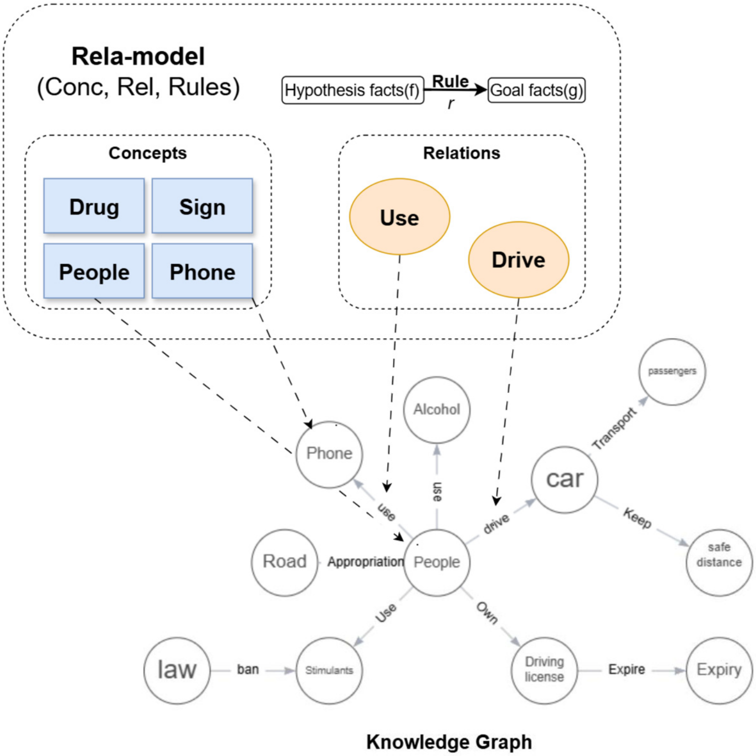

Fig. 8. Simulation of knowledge base in road traffic field based on Lego-Onto model.

The system receives user queries, evaluates them, and converts the relevant knowledge into a knowledge graph. To identify relevant Points – Clause – Laws, this knowledge graph, which is tailored to the particular query, is then compared to the database's knowledge graph. After that, the system shows the customer an extensive list of Points, Clauses, and Regulations that specifically answer their question.

Example 2. Consider the query q_1 = "How is a person riding a motorbike fined for not wearing a helmet?". The designed system provides an accurate answer as Fig. 9:

The meaning of Fig. 9 is as follows:

According to Article 6, Clause 3, Point n (Decree 123/2021/ND-CP):

• ***Article 6:*** *Penalizing individuals riding motorbikes, including electric ones, and similar motorized vehicles violating road traffic regulations.*

• ***Clause 3:*** *Imposing fines ranging from* 400 000 *VND to* 600 000 *VND for specific violations*:

• ***Point n:*** *Not wearing a motorcycle helmet or wearing it without proper strapping when participating in road traffic.*

This response accurately and suitable captures pertinent legal information for query q_1. Moreover, the designed system can support to extract information from multiple documents with suitable articles for an inputted query.

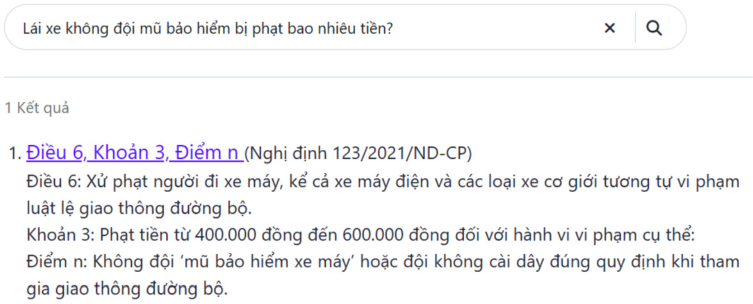

Fig. 9. The answer of our system for query q_1.

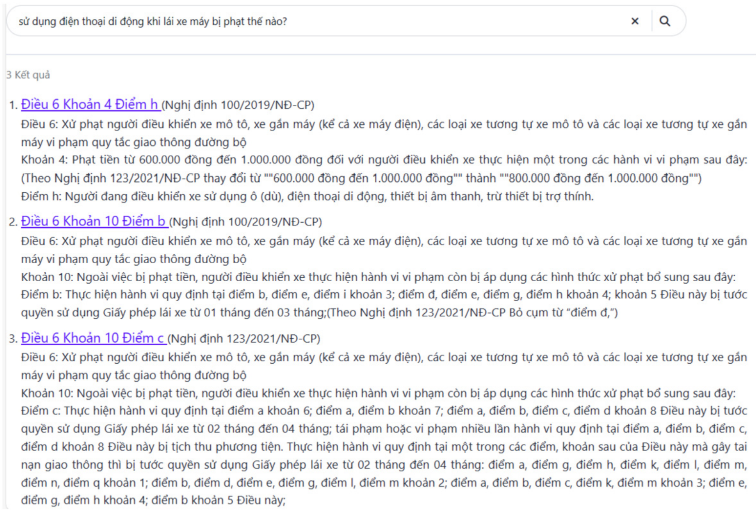

Fig. 10. The answer of our system for query q_2.

Example 3. Consider the query q_2 = "*What is the fine for using a mobile phone while riding a motorcycle*?". The designed system provides an accurate answer, as shown in Fig. 10.

For query q_2, the designed system can combine the required information from many clauses and points in articles of multiple law documents:

- At first, through Article 6, Clause 4, Point h in Vietnam Government (2019) was amended and supplemented in Vietnam National Assembly (2021), it shows a fine ranging from 600 000 VND to 1 000 000 VND is imposed on individuals operating motorcycles, including electric motorcycles, and similar motor vehicles violating regulations on the use of mobile phones, audio devices, excluding hearing aids.
- In addition to the monetary penalty, those violating this regulation may face supplementary penalties, which were retrieved from two legal documents (Vietnam Government, 2019; Vietnam National Assembly, 2021), such as:

Table 3
Testing results on Road Traffic Law with the proposed method.

Kind	Meaning	Quantity	Correct	Accuracy
1	Query about definitions of concepts	30	24	80.0%
2	Violation of traffic signs and signals	47	37	78.7%
3	Personal safety violation	48	40	83.3%
4	Violations related to parking and reversing	22	17	77.3%
5	Vehicle document violations	24	21	87.5%
6	Violation of substances and prohibited substances	14	13	92.8%
7	Obstruction of traffic violations	16	14	87.5%
	Total	**201**	**166**	**82.6%**

- Revocation of the driving license for 1 to 3 months (Article 6, Clause 10, Point b in Vietnam Government, 2019).
- Revocation of the driving license for 2 to 4 months if the violation results in a traffic accident (Article 6, Clause 10, Point c in Vietnam National Assembly, 2021).

This response accurately captures relevant legal information for query q_2 completely.

5.2. *Evaluation of the Accuracy of the Designed System*

The research conducted has yielded several noteworthy results. The central focus of this research was to explore and evaluate the utilization of knowledge graphs and artificial intelligence solutions in addressing legal queries in the field of road traffic law. This study conducts experiments focused on querying the meaning of terminology in Vietnamese road traffic law. Their knowledge contents are divided into 7 kinds:

- Kind 1: Queries about the definitions of concepts.
- Kind 2: Queries about violation of traffic signs and signals.
- Kind 3: Queries about personal safety violation.
- Kind 4: Queries about violations related to parking and reversing.
- Kind 5: Queries about violation of vehicle documents.
- Kind 6: Queries about violation of prohibited substances.
- Kind 7: Queries about obstruction of traffic violations.

The process of this system was checked by experts in Vietnamese road traffic law (traffic police and lawyer in road traffic). Table 3 and Fig. 11 show the testing results for each kind.

In practice, the common queries are kinds 1, 2, and 3. In these kinds, the system gets the highest performance with queries about personal safety violations (Kind 3). That means it can support users to determine actions that risk their safety. Regarding queries in Kind 2, there are many types and forms of traffic signs, and the system has difficulty precisely extracting violations with those signs. For other queries, the proposed method gets better results for queries about the violation of substances and prohibited substances (Kind 4), especially related to alcohol. In Vietnam, it is essential to prohibit the use of alcohol while driving. Thus, the system is designed to suit the requirements of Vietnamese traffic.

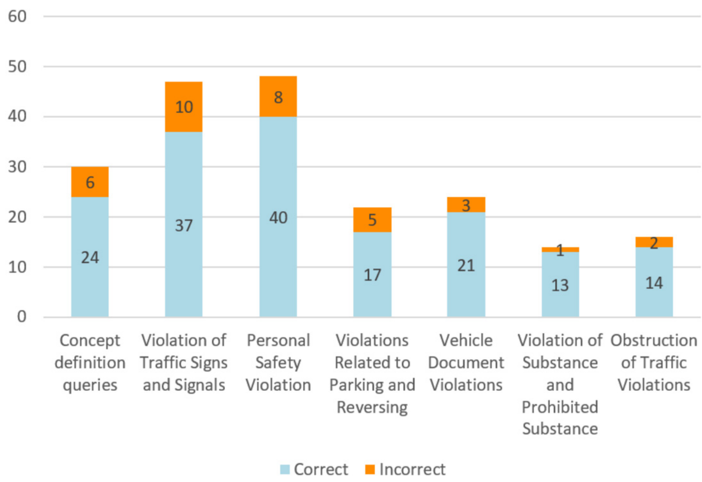

Fig. 11. Accuracy chart of the query system.

5.3. *Comparison with NLP Methods*

In this section, the proposed method is compared with other NLP methods, including TF-IDF, BM25, TIWS and TPS. They are compared based on Vietnamese traffic law documents with the metric TopK@acc.

Metrics: The effectiveness of the methods is evaluated using the TopK@acc metric, where accuracy is defined as the percentage of questions with completely correct labels found in the Top K documents. L_K represents a set of k labels or ID of documents that the system predicts are most relevant to the query, while l_q represents the practical set of labels associated with the query.

$$TopK@acc = \frac{1}{n}\sum_{1}^{n}\begin{cases}1, & l_q \subseteq L_k,\\ 0, & \text{otherwise.}\end{cases} \tag{7}$$

TF-IDF is a method for evaluating the importance of a word in a document or text within a dataset (Mishra and Vishwakarma, 2015). The TF-IDF weight is determined by two factors: the normalized term frequency (TF) factor, which represents the frequency of a word appearing in a document divided by the total number of words in that document. The second factor is related to the Inverse Document Frequency (IDF).

TIWS is a method that uses TF-IDF combined with Word Segmentation (Le *et al.*, 2023). This method extracts the k most relevant articles to answer a given query. Each sentence is encoded using the Word Segmentation extracted from the Undersea library.

In information retrieval, **BM25** acts as a scoring system for documents. It helps search engines to assess how relevant a document is to a search query (Robertson and Zaragoza, 2009). This method considers two factors: how often words appear in a document and how uncommon those words are in the entire database. Moreover, a parameter k balances the importance of these two factors, while another parameter b modifies the impact of document length on the final score.

Table 4
Results of the proposed method and NLP methods.

Methods	Top5@acc	Top10@acc	Top20@acc	Top50@acc
TF-IDF	0.042	0.084	0.153	0.393
BM25	0.037	0.079	0.169	0.321
TIWS	0.058	0.079	0.147	0.367
TPS	0.441	0.5	0.563	**0.688**
Proposed method	**0.561**	**0.603**	**0.645**	0.671

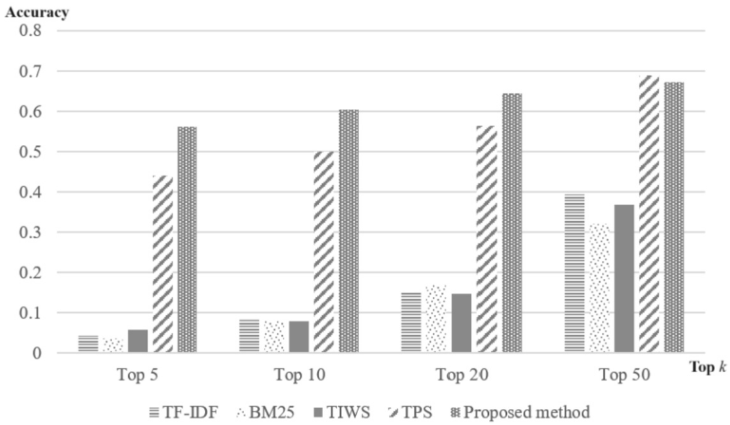

Fig. 12. The search results of TF-IDF, BM25, TIWS, and TPS on Vietnamese Traffic Law.

TPS is a method using TF-IDF with PhoBERT Stage (Le *et al.*, 2023). In this method, negative samples are generated using TF-IDF/BM25 and then embedded with PhoBERT before being fed into the training model. When the training process is finished, it will achieve improved repetition of PhoBERT. This fine-tuned TPS model supports embedding documents and queries from the test dataset. This method utilizes cosine similarity evaluation to predict the most relevant documents corresponding to a given question.

Table 4 compares the results of the TF-IDF, BM25, TIWS, TPS methods and the proposed solution according to the TopK@acc metric with $K = 5, 10, 20, 50$ (Fig. 12).

5.4. *Comparison with Other LLMs*

ChatGPT and Gemini are popular question-answering systems which LLMs constructed. This section compares the designed system with ChatGPT 4o and Gemini 2.5 in the legal domain. For comparison of the results of our system, there are 171 queries related to traffic violations in the Road Traffic Law, including query q_1 in Example 2 and query q_2 in Example 3.

For query q_1, Fig. 13 and Fig. 14 are results of ChatGPT and Gemini for query q_1 in Example 2 (resp.). Those results are:

In Fig. 13, ChatGPT responds: "*In Vietnam, not wearing a helmet while riding a motorbike is considered a traffic violation and may result in a fine. Typically, the fine for this*

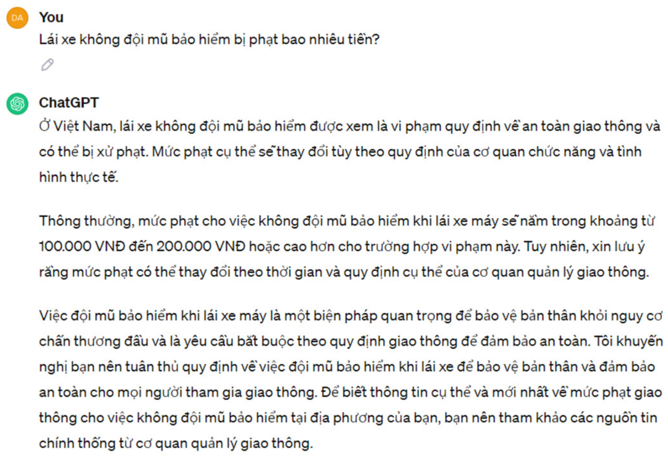

Fig. 13. The answer of ChatGPT for query q_1.

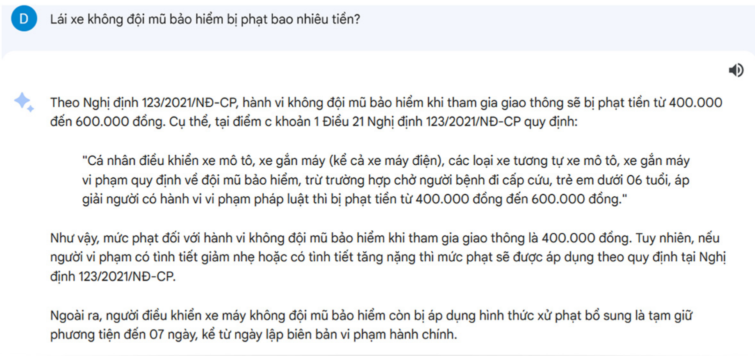

Fig. 14. The answer of Gemini for query q_1.

offense ranges from 100 000 *to* 200 000 *VND. However, these fines can change over time and vary with regulatory updates*". Despite accurately describing the violation, the fine range of "100 000 to 200 000 VND" is entirely inaccurate.

In Fig. 14, Gemini's response is succinct: *According to Decree* 123/2021*/ND-CP, the act of not wearing a helmet while participating in traffic will result in a fine ranging from* 400 000 *to* 600 000 *VND. Specifically, at Point c, Clause* 1*, Article* 21 *of Decree* 123/2021*/ND-CP stipulates*.

Individuals riding motorbikes, including electric ones and similar motorized vehicles, violating regulations on wearing helmets, except in cases of transporting emergency patients, children under 06 *years old, or escorting individuals violating the law, will be fined from* 400 000 *to* 600 000 *VND. Thus, the fine for not wearing a helmet while participating in traffic is* 400 000 *VND. However, if the violator has mitigating or aggravating circumstances, the fine will be applied according to the provisions of Decree* 123/2021*/ND-CP. Additionally, motorbike riders not wearing a helmet may also face the supplementary penalty of vehicle detention for up to 7 days, starting from the date of the administrative*

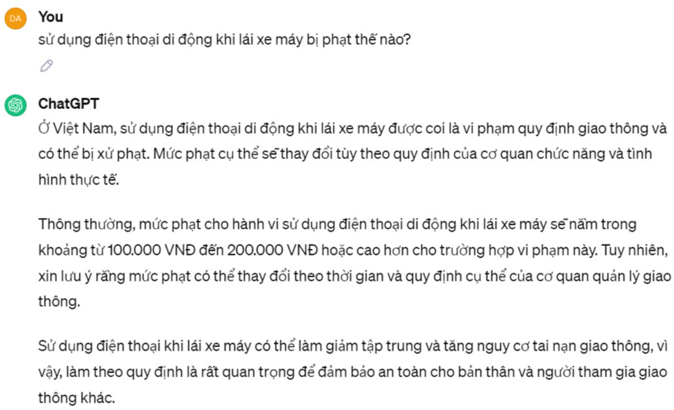

Fig. 15. The answer of ChatGPT for query q_2.

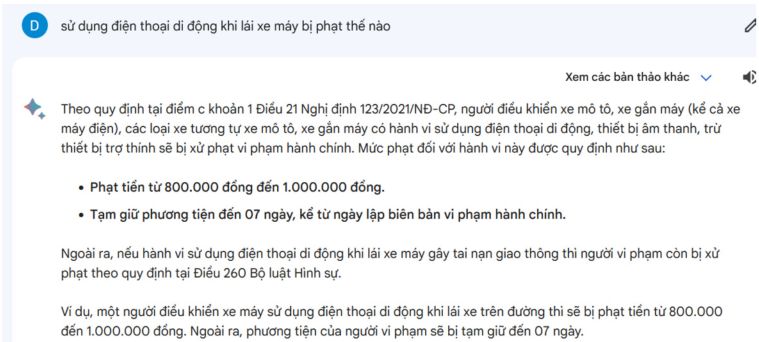

Fig. 16. The answer of Gemini for query q_2.

violation record . The fine "from 400 000 to 600 000 VND" that Gemini mentioned is entirely accurate; however, this fine is not stipulated at Point *c*, Clause 1, Article 21 of Decree 123/2021/ND-CP. Furthermore, this offense incurs only a monetary penalty and is not subject to an "additional penalty of vehicle detention for up to 7 days." Therefore, Gemini's presentation of the supplementary penalty is inaccurate.

For query q_2, Fig. 15 and Fig. 16 are results of ChatGPT and Gemini for for query q_1 in Example 3 (resp.). Those results are:

In Fig. 15, ChatGPT responded: "*In Vietnam, using a mobile phone while riding a motorcycle is considered a traffic violation and may result in a fine. The specific penalty will vary depending on the regulations of the competent authority and the actual situation. Typically, the fine for using a mobile phone while riding a motorcycle ranges from* 100 000 *VND to* 200 000 *VND or higher for this violation. However, please note that fines may change over time and according to the specific regulations of the traffic management authority. Using a phone while riding a motorcycle can reduce concentration and increase the risk of traffic accidents. Therefore, compliance with regulations is essential to ensure safety for oneself and other road users.*"

While partially describing the violation, it inaccurately states the fine as “100 000 to 200 000 VND,” which is significantly different from reality.

In Fig. 16, Gemini’s feedback is concise and understandable: *According to Point c, Clause* 1*, Article* 21 *of Decree* 123/2021/*ND-CP, motorbike riders violating regulations on using mobile phones, sound devices, excluding hearing aids, will be fined administratively. The penalty for this violation is specified as follows*:

A fine ranging from 800 000 *VND to* 1 000 000 *VND*.

Vehicle impoundment for up to 07 *days, starting from the date of issuing the administrative violation report*.

Furthermore, if using a mobile phone while riding a motorbike causes a traffic accident, the violator will be penalized according to the regulations in Article 260 *of the Penal Code*.

Although the response is succinct and accurate in specifying the fine amount, it fails to reference “Point c, Clause 1, Article 21 of Decree 123/2021/ND-CP.” Additionally, the supplementary penalty of “Vehicle impoundment for up to 07 days” is inaccurate and omits the penalty of “revocation of the driving license”.

Along with these queries, testing with Google consistently produced answers by extracting knowledge from articles. However, these results had approximately 25.8% of answers to violation-related questions that needed to be updated in accordance with the latest legal texts, leading to discrepancies in specific penalty levels. Regarding conceptual questions, Google Gemini provided responses with an accuracy rate of up to 93.3%.

Based on the criteria of designing of intelligent systems using the knowledge base (Nguyen *et al.*, 2020; De Cruz, 2024), the comparison between ChatGPT, and Gemini and our system based on the following criteria: accuracy, suitability of content, and usability.

- **Accuracy:** In the context of legal knowledge retrieval, this criterion refers to the correctness, precision, and reliability of the information provided. It evaluates whether the system can deliver factual, up-to-date, and legally sound answers, including specific figures, relevant legal texts, and correct interpretations.
- **Suitability of content:** It assesses how well the generated responses align with the user’s query, are relevant to the legal context, and are presented in a comprehensive and organized manner. It goes beyond mere accuracy to evaluate the usefulness and applicability of the information for the user’s specific needs.
- **Usability:** It focuses on the ease of interaction with the system, the clarity of its outputs, and its practical utility for users, particularly in the legal domain. It evaluates how intuitive, efficient, and user-friendly the system is for accessing and utilizing legal knowledge.

Each criterion is evaluated based on the responding factors as Table 5, and Table 6 compares ChatGPT 4o, Google Gemini 2.5 and our system through these criteria.

The integration of Ontology Legal-Onto and knowledge graphs has proven highly valuable in organizing and presenting legal information in an understandable and consistent format in various legal texts. This significantly supports efficient analysis and understanding of complex legal content, thereby benefiting legal practitioners and researchers. The

Table 5
The factors for evaluating an intelligent search engine in legal domains.

Criteria	Evaluating factors
Accuracy	• Factual Correctness. • Precision of Figures and Details. • Up-to-dateness/Currency of Information. • Domain-Specific Legal Knowledge. • Consistency and Reliability.
Suitability of content	• Relevance to Query. • Specificity of Information. • Comprehensiveness and Completeness. • Contextual Understanding. • Integration of Knowledge Base.
Usability	• Natural Language Understanding. • Clarity of Responses. • Referencing and Sourcing. • User Experience. • Efficiency of Retrieval.

research results indicate that knowledge graphs have substantial potential to contribute to the legal field. They provide more accessible and comprehensible legal information, support legal decision making, and increase the reliability and efficiency of legal knowledge. These findings suggest a promising path for further research and development in integrating artificial intelligence and the legal field. In addition, the structure of the knowledge graph helps the designed system respond to specific legal queries depending on the depth and accuracy of the underlying data.

6. Conclusions and Future Work

This article proposed a method for organizing the knowledge domain about legal documents based on integrating ontology and knowledge graphs. Ontology Legal-Onto is used to represent the content of legal documents. It also conducts a thorough assessment of knowledge graphs concerning legal queries, highlighting instances where they outperform traditional search tools. By using a knowledge graph, the method can extract legal information from multiple documents for inputted queries. Moreover, NLP-based applications, such as PhoBert and VnCoreNLP, have significantly streamlined the knowledge base construction process.

The experiment shows that the proposed method is effective with common query kinds in practice. Compared with other LLMs, the designed system is better than ChatGPT in accuracy and are suitable with the content because its knowledge base is organized for a determined legal domain. In addition, this method is also more effective than Gemini in combining useful legal information from many documents for practical queries. The construction of a knowledge base is acknowledged as a significant challenge, requiring substantial time, effort, and intellectual resources to establish a semantically meaningful

Table 6
Comparison of the designed system with ChatGPT and Gemini in Legal Domain.

Criteria	ChatGPT 4o	Gemini 2.5	Our system
Accuracy	ChatGPT was unable to provide precise figures for the answers and could only offer advice to the inquirer. The knowledge of ChatGPT is not always up-to-date, as it lacks a mechanism for real-time updates and continuous monitoring of changes in legal regulations and interpretations. Moreover, it also lacks domain-specific legal knowledge. The absence of expertise in interpreting and applying legal concepts can result in misunderstandings and inaccuracies.	Gemini has the capability to access and collect huge amounts of information on the Internet, offering an overview of legal answers available online. It can extract information from various sources of information and provides specific values directly from legal texts, aiding in grasping specific information and legal regulations. The system relies on online sources and may not guarantee the accuracy, currency, or completeness of legal answers.	By integrating knowledge fields into the legal domain, it was observed that information retrieval accuracy improved significantly, with an accuracy rate of 82.6%. This enhancement facilitates easier access to legal information for non-experts, enhancing transparency and enabling informed decision-making. Ontology contributes to maintaining the consistency and reliability of legal data through regular verification and updating. It also combines knowledge graphs to identify inconsistencies and errors, ensuring that the legal information remains accurate and dependable.
Suitability of content	The result content is suitable and consistently analyses questions and provides answers relevant to the question's content. It does not provide specific information from legal texts. Users need to cross-check and verify information from reliable sources to ensure the accuracy and reliability of the answers.	Analyse and provide coherent responses to questions in alignment with the question content. However, the information has not yet been organized as a knowledge base. Thus, the accuracy of Gemini for queries that need information from multiple legal documents is not good. It only gives a simple answer.	The knowledge base is integrated knowledge from diverse legal sources organized from the source of law documents. It ensures centralized and easily accessible legal information. This integration reduces data fragmentation and promotes consistency across legal knowledge. Thus, beside the suitable answer for the query's meaning, relations of KG help extract necessary information from many documents.
Usability	ChatGPT accurately understands natural language, and responds to a wide range of legal queries, including complex cases. However, it is unable to provide specific, accurate figures or details about the extracted legal documents. Users are advised to verify information from official legal sources.	Ensure that answers are updated based on the latest legal documents, although the responses are accurate regarding the substantive content of fines. However, Gemini's citations are inaccurately positioned within the legal text, and the supplementary penalties presented are also incorrect.	The system supports reference to the original documents with confidence in the accuracy and reliability of the responses. The system is emerging to release a practical application supporting people to search for legal knowledge and violations in Vietnamese road Traffic law.

query system. The ability of the designed method to improve information accessibility, precision, and consistency makes it a valuable asset in the legal domain.

In the future, the research will endeavour to examine the procedural aspects that facilitate to organize the knowledge base of the system from the large data of legal documents (Wu *et al.*, 2024; Ngo *et al.*, 2024). The effectiveness of this organization will be subject to validation by experts acting in the capacity of knowledge engineers. Additionally, enhancements in NLP techniques will be pursued to refine the analysis of entities and relationships within legal documents and queries, thereby ensuring heightened precision. These advancements are anticipated to augment the effectiveness in catering to a diverse array of legal disciplines, thereby enhancing its responsiveness to user requirements. Furthermore, the devised methodology is poised to contribute to the establishment of a comprehensive framework for knowledge management within the legal domain, achieved through the integration of ontology and knowledge graph principles (Nguyen *et al.*, 2023b). Furthermore, integrating continuous learning mechanisms allows the knowledge graph to adapt and learn from user interactions, enhancing its understanding of user preferences and legal nuances over time. These envisioned enhancements are designed to strengthen the capabilities of the system, making it more adept at handling dynamic legal landscapes, and they can adapt to be utilized in society life with the changing knowledge economy (Li, 2024). This opens up opportunities for the system to become flexible and adaptable in the future, enabling it to proactively respond to new challenges, as well as updates and changes in the legal domain.

Acknowledgements

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under a project within the framework of the Program titled "Strengthening the capacity for education and basic scientific research integrated with strategic technologies at VNU-HCM, aiming to achieve advanced standards comparable to regional and global levels during the 2025–2030 period, with a vision toward 2045."

References

Abu-Salih, B., Alotaibi, S. (2024). A systematic literature review of knowledge graph construction and application in education. *Heliyon*, 10(3), 25383.

Ajevski, M., Barker, K., Gilbert, A., Hardie, L., Ryan, F. (2023). ChatGPT and the future of legal education and practice. *The Law Teacher*, 57(3), 352–364.

Allison, N.G. (2023). From semantic weight to legal ontology via classification of concepts in legal texts. *The Law Teacher*, 57(2), 201–217.

Casillas, J., Cordón, O., Triguero, F.H., Magdalena, L. (2013). *Interpretability Issues in Fuzzy Modeling*. Springer-Verlag, Berlin Heidelberg.

Chakraborty, C., Khan, M. (2023). Big data-driven futuristic fabric system in societal digital transformation. *Big Data*, 11(5), 321–322.

Dang, D., Nguyen, H., Le, T., Do, H., Nguyen, H., Ngo, H.Q., Nguyen, H.D. (2025). Knowledge graph-based legal query system with LLM and retrieval augmented generation. In: *Proceedings of the 17th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2025), Part 1*, pp. 161–172.

De Cruz, P. (2024). *Comparative Law in a Changing Word*. Taylor & Francis, London.
Delvin, J., Chang, M.W., Kenton, L., Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT 2019*, pp. 4171–4186.
Desai, D., Ghadge, A., Wazare, R., Bagade, J. (2021). A comparative study of information retrieval models for short document summaries. In: *Proceedings of Fourth International Conference on Computer Networks and Inventive Communication Technologies (ICCNCT 2021)*, pp. 547–562.
Dhani, J.S., Bhatt, R., Balaji, G., Sirohi, P., Bhatnagar, V. (2023). Similar cases recommendation using legal knowledge graphs. In: *Proceedings of 3rd Symposium on Artificial Intelligence and Law (SAIL 2023)*, pp. 1–12.
Ding, Y., Li, H., Zhu, F., Wang, Z., Peng, W., Xie, M., (2023). A semi-supervised failure knowledge graph construction method for decision support in operations and maintenance. *IEEE Transactions on Industrial Informatics*, 20(3), 3104–3114.
Do, N., Nguyen, H.D., Selamat, A. (2018). Knowledge-based model of expert systems using Rela-model. *International Journal of Software Engineering and Knowledge Engineering*, 28(08), 1047–1090.
GoogleAI (2024). Gemini. https://gemini.google.com/ [Accessed: 08 July 2024].
Gunkel, D. (2023). *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. MIT Press, Cambridge, Massachusetts.
Hoang, L., Bui, T., Nguyen, C., Nguyen, L.-M. (2023). AIEPU at ALQAC 2023: deep learning methods for legal information retrieval and question answering. In: *Proceedings of 15th International Conference on Knowledge and Systems Engineering (KSE 2023)*. bpublisherIEEE, pp. 1–6.
Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., et al. (2021). Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4), 71.
Junior, A.C., Orlandi, F., Graux, D., Hossari, M., O'Sullivan, D., Hartz, C., Dirschl, C. (2020). Knowledge graph-based legal search over German Court cases. In: *Proceedings of 17th European Semantic Web Conference (ESWC 2020)*, pp. 293–297.
Khan, N., Ma, Z., Ullah, A., Polat, K. (2022). DCA-IoMT: knowledge-graph-embedding-enhanced deep collaborative alert recommendation against COVID-19. *IEEE Transactions on Industrial Informatics*, 18(12), 8924–8935.
Le, H.H., Nguyen, C.-T., Ngo, T.P., Vinh, P.V., Nguyen, B.T., Huynh, A.T., Nguyen, H.D., (2023). Intelligent retrieval system on legal information. In: *Proceedings of 15th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2023)*, pp. 97–108. Springer.
Le, T., Luu, S., Nguyen, H., Do, N. (2019). Knowledge representation for designing an Intelligent Tutoring System in learning of courses about algorithms. In: *Proceedings of 25th Asia-Pacific Conference on Communications (APCC 2019)*, pp. 310–315.
Li, M. (2024). Adapting legal education for the changing landscape of regional emerging economies: a dynamic framework for law majors. *Journal of the Knowledge Economy*, 15(3), 10227–10256.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
Liu, Y., Han, T., Ma, S., et al. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2), 100017.
Manning, C., Raghavan, P., Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University, Cambridge.
Maroudas, S., Legkas, S., Malakasiotis, P., Chalkidis, I. (2022). Legal-tech open diaries: lesson learned on how to develop and deploy lightweight models in the era of humongous language models. In: *Proceedings of the Natural Legal Language Processing Workshop 2022*, pp. 88–110.
Martinez-Gil, J. (2023). A survey on legal question–answering systems. *Computer Science Review*, 48, 100552.
Meta (2024). Introducing LLaMa: a foundational, 65-billion-parameter language model. https://ai.meta.com/blog/large-language-model-llama-meta-ai/ [Accessed: 08 July 2024].
Min, B., Ross, H., Sulem, E., et al. (2023). Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Computing Surveys*, 56(2), 1–40.
Mishra, A., Vishwakarma, S. (2015). Analysis of TF-IDF model and its variant for document retrieval. In: *Proceedings of 2015 International Conference on Computational Intelligence and Communication Networks (CICN 2015)*, pp. 772–776.
Mohan, D., Nair, L.R. (2022). Sammon quadratic recurrent multilayer deep classifier for legal document analytics. *Computers, Materials & Continua*, 72(2), 3039–3053.

Mökander, J., Schuett, J., Kirk, H.R., Floridi, L. (2024). Auditing large language models: a three-layered approach. *AI and Ethics*, 4, 1085–1115.

Moneus, A.M., Sahari, Y. (2024). Artificial intelligence and human translation: a contrastive study based on legal texts. *Heliyon*, 10(6).

Naseem, U., Razzak, I., Khan, S.K., Prasad, M. (2021). A comprehensive survey on word representation models: from classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5), 1–35.

Ngo, H., Nguyen, H., Le-Khac, N.A. (2024). Ontology knowledge map approach towards building linked data for Vietnamese legal applications. *Vietnam Journal of Computer Science*, 11(02), 323–342.

Ngo, H.Q., Nguyen, H.D., Le-Khac, N.-A. (2023). Building legal knowledge map repository with nlp toolkits. In: *Proceedings of 12th Conference on Information Technology and its Applications (CITA 2023)*. Springer, pp. 25–36.

Nguyen, H.D., Do, N.V., Tran, N.P., et al. (2020). Some criteria of the Knowledge Representation method for an Intelligent Problem Solver in STEM education. *Applied Computational Intelligence and Soft Computing*, 2020(1), 9834218.

Nguyen, H.D., Tran, T.-V., Pham, X.-T., Huynh, A.T., Do, N.V. (2021). Ontology-based integration of knowledge base for building an intelligent searching chatbot. *Sensors & Materials*, 33(9), 3101–3121.

Nguyen, H.D., Huynh, H., Mai, T., et al. (2023a). Design an ontology-based model for intelligent querying system in mathematics education. *Journal of Interdisciplinary Mathematics*, 26(3), 449–473.

Nguyen, H.D., Truong, D., Vu, S., et al. (2023b). Knowledge management for information querying system in education via the combination of rela-ops model and knowledge graph. *Journal of Cases on Information Technology (JCIT)*, 25(1), 13.

Nguyen, H., Pham, V., Ngo, H.Q., Huynh, A., Nguyen, B., Machado, J. (2024). Intelligent search system for resume and labor law. *PeerJ Computer Science*, 10, 1786.

Nguyen, T.-S., Nguyen, L.-M., Tojo, S., et al. (2018). Recurrent neural network-based models for recognizing requisite and effectuation parts in legal texts. *Artificial Intelligence and Law*, 26, 169–199.

Nguyen, T., Nguyen, H.D., Pham, V.T., et al. (2022). Legal-Onto: an ontology-based model for representing the knowledge of a legal document. In: *Proc. of 17th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2022), Online*, pp. 426–434.

Oliveira, F.d., Oliveira, J.M.P.d. (2024). A RDF-based graph to representing and searching parts of legal documents. *Artificial Intelligence and Law*, 32, 667–695.

Peng, C., Xia, F., Naseriparsa, M., Osborne, F. (2023). Knowledge graphs: opportunities and challenges. *Artificial Intelligence Review*, 56(11), 13071–13102.

Pham, V.T., Nguyen, H.D., Le, T., B., N., Ngo, H.Q. (2023). Ontology-based solution for building an intelligent searching system on traffic law documents. In: *Proceedings of 15th International Conference on Agents and Artificial Intelligence (ICAART 2023), Lisbon, Portugal*, pp. 426–434.

Robertson, S., Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.

Sansone, C., Sperlí, G. (2022). Legal information retrieval systems: state-of-the-art and open issues. *Information Systems*, 106, 101967.

Sartor, G., Araszkiewicz, M., Atkinson, K., et al. (2022). Thirty years of artificial intelligence and law: the second decade. *Artificial Intelligence and Law*, 30(4), 521–557.

Satti, M.I., Ahmed, J., Muslim, H.S.M., et al. (2023). Ontology-based news linking for semantic temporal queries. *Computers, Materials & Continua*, 74(2), 3913–3929.

Sejnowski, T. (2023). Large language models and the reverse turing test. *Neural Computation*, 35(3), 309–342.

Shao, Y., Wu, Y., Liu, Y., et al. (2023). Understanding relevance judgments in legal case retrieval. *ACM Transactions on Information Systems*, 41(3), 1–32.

Sikos, L.F. (2021). AI in digital forensics: ontology engineering for cybercrime investigations. *Wiley Interdisciplinary Reviews: Forensic Science*, 3(3), 1394.

Soni, S., Chouhan, S., Rathore, S. (2023). Textconvonet: a convolutional neural network based architecture for text classification. *Applied Intelligence*, 53(1), 14249–14268.

Sovrano, F., Palmirani, M., Vitali, F. (2020). Legal knowledge extraction for knowledge graph based question-answering. In: *Legal Knowledge and Information Systems*. IOS Press, Amsterdam, pp. 143–153.

Tang, Y., Qiu, R., Liu, Y., et al. (2024). CaseGNN: graph neural networks for legal case retrieval with text-attributed graphs. In: *Proceedings of 46th European Conference on Information Retrieval (ECIR 2024)*. Springer, pp. 80–95.

Van, H.N., Nguyen, D., Nguyen, P.M., Le Nguyen, M. (2022). Miko team: deep learning approach for legal question answering in ALQAC 2022. In: *Proceedings of 14th International Conference on Knowledge and Systems Engineering (KSE 2022)*. IEEE, pp. 1–5.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 6000–6010.

Vietnam Government (2019). Decree on Administrative penalties for road traffic and rail transport offences, No. 100/2019/ND-CP.

Vietnam Ministry of Transport (2019). National Technical Regulation on Traffic Signs and Signals. QCVN 41:2019/BGTVT.

Vietnam National Assembly (2008). Law on Road Traffic, No. 23/2008/QH12.

Vietnam National Assembly (2021). Amending and supplementing some articles of decrees provisions on penalties for administrative violations in the marine region, road, rail traffic, and domestic airline, No. 123/2021/ND-CP.

Villata, S., Araszkiewicz, M., Ashley, K., et al. (2022). Thirty years of artificial intelligence and law: the third decade. *Artificial Intelligence and Law*, 30(4), 561–591.

Vuong, Y., Hoang, Q., Nguyen, M., et al. (2023). Constructing a knowledge graph for Vietnamese legal cases with heterogeneous graphs. In: *Proceedings of 15th International Conference on Knowledge and Systems Engineering (KSE 2023), Hanoi, Vietnam*, pp. 1–6.

Wang, J., Dong, Y. (2020). Measurement of text similarity: a survey. *Information*, 11(9), 421.

Wu, Q., Yu, Y., Zhou, L., Lu, Y., Chen, H., Qian, X. (2024). Storage and query indexing methods on big data. *Arabian Journal for Science and Engineering*, 49(5), 7359–7374.

Xia, C., He, T., Li, W., Qin, Z., Zou, Z. (2019). Similarity analysis of law documents based on Word2vec. In: *Proceeding of 19th IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C 2019)*, pp. 354–357.

Yoshioka, M., Suzuki, Y., Aoki, Y. (2022). HUKB at the COLIEE 2022 statute law task. In: *JSAI International Symposium on Artificial Intelligence*, Springer, pp. 109–124.

Zalesinska, A., Szostek, D., Zalucki, M. (2021). *Legal Tech: Information Technology Tools in the Administration of Justice*. Nomos, Baden-Baden.

Zouaoui, S., Rezeg, K. (2021). A novel quranic search engine using an ontology-based semantic indexing. *Arabian Journal for Science and Engineering*, 46(4), 3653–3674.

V.T. Pham received the BS degree in mathematics and informatics from the University of Science, VNU-HCM, Vietnam, in 2003, and the MS degree in Information Technology from University of Science, VNU-HCM, Vietnam, in 2008. He is currently a PhD student at University of Science, VNU-HCM. He is also a lecturer at Sai Gon University, Vietnam. From 2006–2019, he was a lecturer at the Faculty of Software Engineering, University of Information Technology, VNU-HCM, Vietnam. His research interests include artificial intelligence, computer science, software engineering, mathematics foundation.

D.V. Dang received his MSc degree in computer science in 2023. He is currently a PhD candidate in the Department of Computer Science since 2025 and serves as a lecturer at the Faculty of Software Engineering, University of Information Technology, Vietnam National University Ho Chi Minh City, Vietnam. His research interests include knowledge representation, automated reasoning, and knowledge engineering.

Hung Q. Ngo is a PostDoc Researcher at the School of Computer Science, University College Dublin (UCD), Ireland. He received his PhD degree in computer science from the University College Dublin in 2022 and his MSc degree in computer science from the University of Science, Vietnam National University – Ho Chi Minh City (VNUHCM), Vietnam in 2008. His research interests include knowledge management, natural language processing, and intelligent systems in cross-disciplinary domains, such as bioinformatics, digital agriculture, legal AI, and digital forensics.

B.T. Nguyen received the PhD degree (Hons.) from Ecole Polytechnique, Paris, France, in 2012. He is currently the head of the Department of Computer Science and an associate professor of Computer Science with the Faculty of Mathematics and Computer Science, University of Science, Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam. He has over ten years of experience in AI and data science. Up to now, he has had over 100 publications and 04 patents filed in USA and Canada. He has substantial experience building research and development teams to help companies or startups deliver AI products.

N.V. Do received his MSc and PhD degrees in computer science from University of Science, VNU-HCM, Vietnam in 1995 and 2002, respectively. He was an Associate Professor at University of Information Technology, VNU-HCM, Vietnam, from 2006 to 2018 and at Hoa Sen University, Vietnam, from 2018 to 2019. He has been an associate professor and the head of Information Technology Department at Hong Bang International University, Vietnam, from 2020. His research interests include artificial intelligence, computer science, and their practical applications, especially intelligent systems and knowledge-based systems.

A. Selamat is currently a full professor with Universiti Teknologi Malaysia (UTM), Malaysia. He is a Deputy Vice Chancellor (Student Affairs & Alumni), UTM. He was the Dean of the Malaysia Japan International Institute of Technology (MJIIT), UTM, since 2018. He is also a Professor with the Software Engineering Department, Faculty of Computing, UTM, and was the Chair of the IEEE Computer Society Malaysia Section. He has published more than 120 research articles with IF JCR, with more than 2400 citations received in the Web of Science. His research interests include software engineering, software agents, web engineering, information retrievals, pattern recognition, genetic algorithms, neural networks, soft computing, collective computational intelligence, strategic management, key performance indicator, and knowledge management.

H.D. Nguyen is currently an associate professor at the Faculty of Computer Science, University of Information Technology, VNU-HCM, Vietnam, from 2008. He was a Visiting Assistant Professor at the Computer Science Department, New Mexico State University (NMSU), USA, in 2024–2025. His research interests include knowledge representation, automated reasoning, and knowledge engineering, especially intelligent systems in education, such as intelligent problem solvers. He received the Best Paper Awards at CITA 2023, SOMET 2022, and ICOCO 2022.