QBR – A Question-Bank-Based Approach to Fine-Grained Legal Knowledge Retrieval for the General Public

Mingruo Yuan , Ben Kao*, Tien-Hsuan Wu
The University of Hong Kong
{mryuan, kao, thwu}@cs.hku.hk

Abstract

Retrieval of legal knowledge by the general public is a challenging problem due to the technicality of the professional knowledge and the lack of fundamental understanding by laypersons on the subject. Traditional information retrieval techniques assume that users are capable of formulating succinct and precise queries for effective document retrieval. In practice, however, the wide gap between the highly technical contents and untrained users makes legal knowledge retrieval very difficult. We propose a methodology, called QBR, which employs a Questions Bank (QB) as an effective medium for bridging the knowledge gap. We show how the QB is used to derive training samples to enhance the embedding of knowledge units within documents, which leads to effective fine-grained knowledge retrieval. We discuss and evaluate through experiments various advantages of QBR over traditional methods. These include more accurate, efficient, and explainable document retrieval, better comprehension of retrieval results, and highly effective fine-grained knowledge retrieval. We also present some case studies and show that QBR achieves social impact by assisting citizens to resolve everyday legal concerns.

1 Introduction

Law is inextricably linked to daily life. Day-to-day activities are subject to legal regulations. Whether we are shopping, at work, driving, or posting on social media, legal considerations are always at play. It is thus important for individuals to understand the law to protect their rights and benefits. Yet, legal knowledge is extensive and technical. It is impractical for one to master all aspects of law. As such, we often rely on legal professionals for help when facing legal issues.

There are online platforms that provide support for individuals on legal issues, such as ABA Free Legal Answers[1] in the US, LawWorks[2] in the UK, and Justice Connect[3] in Australia. Questions posted on those platforms are answered by pro bono lawyers for free. There is a significant rise in the demand for these legal question-answering services. Patiño and others [2019] report that legal problems are ubiquitous. Their survey shows that approximately half of the people interviewed had experienced legal problems within two years prior to the interviews. In addition, it is reported in [ABA Free Legal Answers, 2023] that the number of legal questions responded to in each year had increased from 4,193 in 2016 to 71,640 in 2023. This increase suggests that more people are turning to online help. However, professional, licensed pro bono lawyers are a scarce resource, which cannot meet the ever-increasing demand for online legal help. In this work, we present QBR[4], an AI-assisted approach that helps users retrieve relevant legal knowledge that addresses a user's legal situation by leveraging a collection of legal educational articles and a question bank. We show that our platform can serve a large community in providing expert legal answers, thus effectively addressing the high demand for online legal question answering.

Nowadays, legal information such as court judgments and legislation is available online in many countries. However, their online availability does not translate directly into effective public access to legal knowledge. It remains challenging for ordinary people without a legal background to learn legal knowledge for two reasons. Firstly, the information available online consists mostly of primary legal sources, such as cases and statutes. These documents are written in formal legal language that is generally hard to comprehend by the public. The incomprehensibility of professional documents has been reported in various studies [Hutchinson *et al.*, 2016; Basch *et al.*, 2020; Ferguson *et al.*, 2021]. For example, [Ruohonen, 2021] analyzes the readability of 201 legislations and related policy documents in the European Union (EU). It is found that a PhD-level education is required to comprehend certain laws and policy documents. Secondly, the public may not know the legal principles that are applicable to the legal situation they face. With large numbers of documents, it is difficult for a user to locate the correct legal sources in search of a solution to his/her legal problem. Information retrieval (IR) systems that can bridge such a knowledge gap are therefore essential.

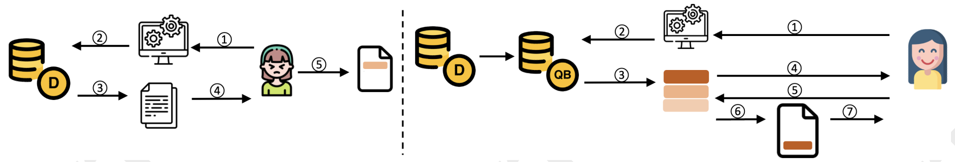
[*] Corresponding author

[1] https://abafreelegalanswers.org/

[2] https://www.lawworks.org.uk/

[3] https://justiceconnect.org.au/

[4] https://github.com/mingruo-yuan/QBR



(a) Traditional Information Retrieval Approach

(b) QBR Approach

Figure 1: Traditional IR vs QBR. **(a)**: ① User input ② Match input with documents ③ Retrieve top results ④ Return all relevant documents ⑤ Read all documents and identify answer. **(b)**: ① User input ② Match input with (q, s_q) ③ Retrieve top results ④ Display short questions ⑤ Select relevant question ⑥ Identify document ⑦ Return answer s_q with the source document

In this paper we propose a question-bank-based retrieval (QBR) methodology for bridging the knowledge gap. We start with a corpus of legal educational documents. These documents are written by lawyers to explain the various legal concepts in layperson's terms. Based on the document contents, we generate model questions and answers, which are collected in a *Question Bank* (QB). QBR answers a user's legal question through retrieving model questions and answers from the QB. In traditional IR, a typical search engine returns a shortlist of documents given a user input query. The user is expected to read the documents in the search result, understand each of them, pick the best matching one, and then extract the most relevant knowledge units (sentences or paragraphs within the documents) that provide an answer to the enquiry. This workflow is illustrated in Figure 1a. Due to the knowledge gap between the professional documents and novice users, it is generally a daunting task for a user to perform the post-search knowledge filtering and extraction step (Step ⑤ in Figure 1a).

Our QBR approach (illustrated in Figure 1b) aims at providing a more comprehensible search results to the user to greatly simplify this step. The idea is to help a user (who may not have any legal training) express a legal question by selecting model questions from the QB. Instead of returning a list of documents, QBR returns a list of *question-answer pairs* (q, s_q)'s. Specifically, for each (q, s_q) pair returned, q is a question from the QB and s_q is an *answer scope*, which is a logical knowledge unit (typically consists of one or few paragraphs) extracted from a document d_q such that s_q answers the question q. To filter the search results, a user needs only read the very short questions q's returned to locate relevant ones. This is much simpler than reading the full contents of shortlisted documents under traditional search engines (Figure 1a). Moreover, the question q effectively *explains* why the answer s_q (and its originating document d_q) is relevant to the user's enquiry; that s_q (and thus d_q) answers the question q, which the system infers as a more proper expression of the user's enquiry. Table 1 shows an example user input and one returned (q, s_q) pair. In this example the answer scope s_q, which is highly relevant to the user's legal situation, is a paragraph of a document in our legal corpus. Instead of returning the full document, which is very long and consists of other less relevant information, QBR returns the question q in Table 1 to the user. If the user finds that q best paraphrases his concerns, the user will select q and QBR will bring to the user the answer s_q as well as the document source, d_q.

User Input: I was out having a drink and I took a picture of a nightclub signboard. Out of nowhere, a man rushed up to me and started accusing me loudly, saying, "Hey! What are you doing taking a picture of me!?" He even demanded that I delete the photos. What is the relevant law in this situation?
q: What are the consequences if persons take photos in a public place without the photo subject's consent and cause any person to be reasonably concerned for his or her safety?
$\boldsymbol{s_q}$: Persons who publish photos with captions that contain personal data of the photo subjects (without their consent) may have violated the Personal Data Privacy Ordinance. If such captions contain unjustified adverse comments on the photo subjects, the publishers may also have incurred civil liability for defamation. Furthermore, persons who take photos on such occasions (i.e. in a public place without the photo subject's consent) and cause any person to be reasonably concerned for his/her safety may have committed the offence of loitering or may be charged with behaving in a disorderly manner in a public place.

Table 1: An example of user input and (q, s_q) pair

Our key contributions are as follows:

- We propose QBR, a new question-bank-based approach to effectively answer legal questions by assisting users with fine-grained legal knowledge retrieval from a corpus of legal educational documents. We show that QBR effectively simplifies the retrieval process, which is critical in community legal question-answering systems where users are typically non-legally-trained.
- We explain the design of QBR, particularly on how a question bank is effectively derived and employed to most accurately identify the matching knowledge units that best answer a given user's legal question.
- We deploy QBR on an online public platform [5]. We provide a case study to illustrate how such a platform helps the public solve their legal problems. This shows the social impacts of our platform.

2 Question Bank and QBR

In this section we give an overview of QBR highlighting four benefits of using a question bank.

Question Bank Let D be a document collection that provides professional knowledge on a certain subject matter, where $d \in D$ denotes a document that is comprised of paragraphs. Conceptually, each document provides information on a particular topic under the subject matter covering various aspects and details. We consider each document d to be composed of multiple *knowledge units*, each could be the

[5] https://ai.hklii.hk/recommender/

Algorithm 1 QBR

```
Input: Question Bank $QB$, User Input $u$, Embedding Function $T$, CL Adjusted Embedding $T'$
Output: Document $d^*$, Scope $s^*$
// (Step 1) Document selection
$(\tilde{q}, d^*, \tilde{s}) \leftarrow \arg\max_{(q,d,s)\in QB} \frac{(T(u)\cdot T(q;s))}{(\|T(u)\|\cdot\|T(q;s)\|)}$;
// (Step 2) Scope disambiguation
$S_{d^*} \leftarrow$ scope set of $d^*$
$s^* \leftarrow \arg\max_{s\in S_{d^*}} \frac{(T'(u)\cdot T'(s))}{(\|T'(u)\|\cdot\|T'(s)\|)}$;
return $d^*, s^*$
```

subject of a user's enquiry. For example, a document on a legal advice website that explains traffic offenses would consist of sections (knowledge units), each defines a specific type of offence and its penalty. A Question Bank (QB) is a collection of question-document-answer tuples (q, d_q, s_q)'s, where q is a question whose answer can be found in document (with id) d_q, and s_q is an *answer scope* (or simply *scope*) that specifies the paragraphs within document d_q that explicitly answer q. Each scope s_q thus represents a single logical knowledge unit that answers a specific question (q). We will elaborate on how QB is constructed from a document collection D later in this section. For the moment, we assume that the knowledge presented in the documents of D is *well covered* by the QB. That is, corresponding to each knowledge unit presented in D, there is one or more (q, d_q, s_q) tuples in QB such that the answer scope s_q comprehensively conveys the knowledge.

For each document $d \in D$, we construct a *question set* Q_d that includes all the questions in the question bank whose answer scopes are found in d. Also, the corresponding answer scopes are collected into a *scope set* S_d. Formally,

$$Q_d = \{q|(q, d_q, s_q) \in QB, d_q = d\}; \quad S_d = \{s_q|(q, d_q, s_q) \in QB, d_q = d\}.$$

Since each answer scope s_q is considered a single logical unit of knowledge (that answers a specific question), the scope set S_d specifies all the knowledge units found in d. Therefore, S_d provides a structured semantic representation of d based on the knowledge it contains. Moreover, the question set Q_d gives a set of questions that are answered by the content of document d and so they are highly relevant to d. This brings us to the first advantage of having a question bank:

Adv. 1 (Document Augmentation): *A QB provides textual information Q_d that describes each document d.*

Essentially, the representation of each document d can be augmented to $d + Q_d$. As we will see later, this augmentation helps disambiguate documents, significantly improving retrieval accuracy.

QBR The QBR method consists of two steps, namely, document retrieval followed by answer scope retrieval. Given a user input u, Step 1 locates the best matching documents $d^* \in D$. Then, given d^*, QBR identifies the best matching scope s^* within d^* as the answer to the user query u. Algorithm 1 outlines the procedure. We elaborate on the two steps in the following descriptions.

Step 1: Document Retrieval Traditional approaches to document retrieval measure a similarity $Sim(u, d)$ between a user's input u against each document $d \in D$ and then return the documents that give the highest similarity scores. There are many methods ranging from simple word matching (e.g., BM25) to lexical analysis and embedding techniques (e.g., BERT). These methods differ in the similarity functions they employ. With the availability of a question bank, QBR matches user input (u) against *questions* (q's) and *scopes* (s's) that are found in the QB instead of directly against the documents (d's) themselves. The rationale is twofold: First, user input is expressed in the form of a question and therefore matching u against questions q's are generally more effective compared with matching u against documents d's. Secondly, a document d could contain multiple knowledge units, some of which could be irrelevant to the user input. This extra (irrelevant) information in d acts as noise, which lowers the effectiveness of the similarity measure. Scopes (s's), on the other hand, are fine-grained knowledge units. They provide more focused retrieval, which reduces the noise effect.

QBR employs a transformer neural network, denoted $T(x)$, that transforms a text sequence x into a continuous vector representation in a semantically rich latent space. Let P be a collection of question-scope pairs (q-s pairs for short) derived from QB, i.e., $P = \{(q, s)|(q, d, s) \in QB\}$. QBR identifies the q-s pairs that best "match" a user input u by measuring the cosine similarity between each q-s pair and u. That is,

$$Sim(u, (q, s)) = (T(u)\cdot T(q; s))/(\|T(u)\|\cdot\|T(q; s)\|). \quad (1)$$

Let $(q^*, s^*) = \arg\max_{(q,s)\in P} Sim(u, (q, s))$ be the pair with the highest similarity score, and let (q^*, d^*, s^*) be the corresponding entry in QB. QBR identifies d^* as the best matching document.

Step 2: Scope Disambiguation with Contrastive Learning Given the best matching document d^*, QBR performs fine-grained knowledge retrieval by identifying the scope in document d^* that best answers a given user input u. We remark that this scope-based retrieval is much more challenging than traditional document retrieval. The reason is that while different documents generally cover different topics, different scopes of the same document present various facts and details *of the same topic*. It is therefore much harder to discern the subtle differences among the knowledge units (scopes) within the same document to pinpoint the one that best answers a user's enquiry. Moreover, due to the wide knowledge gap between professional documents and novice users, a user input u may not provide sufficient textual clues to disambiguate scopes. For example, a user concerned with "defamation" charges might actually mean "libel" or "slander". The two cases could be individually explained in two different scopes, both referencing "defamation". QBR improves scope-based retrieval by utilizing the QB and applying contrastive learning (CL) to modify the embedding function. The idea is that each scope s is associated with some questions in the QB, given by

$$R_s = \{q|(q, d_q, s_q) \in QB, s_q = s\}. \quad (2)$$

These questions provide additional information to improve the separation of scopes' embedding based on the questions the scopes (knowledge units) answer.

Given a document d and its scope set S_d, our objective is to apply contrastive learning (CL) to adjust the embedding function $T()$ such that the embedding vectors of the scopes in d are sufficiently separated. The idea is to utilize the question set Q_d and the scope set S_d from QB to construct training examples. Specifically, for each $s \in S_d$, we first obtain the set of questions R_s that s answers. We then construct a positive example set ($E^+(s)$) and a negative example set ($E^-(s)$):

$$E^+(s) = \{(q,s)|q \in R_s\};\;\; E^-(s) = \{(q,s')|q \in R_s, s' \in S_d \setminus \{s\}\}. \tag{3}$$

Note that for each $(q,s) \in E^+(s)$, s is an answer of q, while for each $(q,s') \in E^-(s)$, s' is a scope in document d that does not answer q. Same as InfoNCE [van den Oord *et al.*, 2019] with cosine similarity on normalized embedding [Chen *et al.*, 2020], we use softmax loss to differentiate positive examples from negative examples. We fine-tune the embedding function $T()$ with the objective of pulling the embedding vectors $T(q), T(s)$ of positive examples $(q,s) \in E^+(s)$ closer and pushing those of negative examples in $E^-(s)$ farther. We use the following equation for training loss:

$$L_{CL}(s) = -\sum_{(q,s)\in E^+(s)} \log \frac{e^{sim(q,s)/\tau}}{e^{sim(q,s)/\tau} + \sum_{(q,s')\in E^-(s)} e^{sim(q,s')/\tau}},$$

where τ is the temperature parameter and $sim(a,b)$ represents the cosine similarity of $T(a)$ and $T(b)$. We use $T'()$ to represent the adjusted embedding function obtained by CL. We remark that the purpose of $T'()$ is to provide an embedding that can better disambiguate the different scopes *within the same document*. On the other hand, the original function $T()$ is retained for the purpose of document retrieval. With CL, we measure the similarity between the user input u and the scopes in S_d using the embedding function $T'()$ to identify the best-matching scope s^* (see Algorithm 1). We also find an entry $(q, s^*) \in QB$ as a search result q-s pair to be displayed to the user (see Figure 1b). From this discussion, we see that the QB provides a means for scope-based knowledge retrieval.

Adv. 2 (Fine-grained Retrieval): *A QB enables fine-grained scope-based retrieval. In particular, it provides training examples for contrastive learning for effective scope disambiguation.*

GPT-Augmentation We augment the CL training data by employing ChatGPT [OpenAI, 2022] to generate user input that mimics novice users. Specifically, for each document $d \in D$, we select two scopes in S_d whose embedding vectors are the most similar. These scopes are the hardest to disambiguate. For each such scope s, we consider any $(q,s) \in E^+(s)$ and generate a few user inputs $\widehat{u_q}$ using ChatGPT with the prompt: "*Given the following context, provide a realistic real-life scenario that a person who knows nothing about legal knowledge might encounter. Context: q.*" We then add $(\widehat{u_q}, s)$ to $E^+(s)$ and $(\widehat{u_q}, s')$, where $s' \in S_d \setminus \{s\}$, to $E^-(s)$.

The QBR procedure we have mentioned so far selects one matching document d^* (Step 1) from which a qs-pair (Step 2) is obtained. We extend the above procedure by retrieving the top-k documents (and hence top-k q-s pairs) if multiple (k) search results are desired. If so, the top-k questions, say q_1, .., q_k (and their respective answers s_1, .., s_k), are displayed to the user, who will then choose among the returned questions the ones that best match his/her enquiry. Note that under QBR, this step of *result filtering* involves the user reading only the returned questions q's, which are very short text. This is in sharp contrast to traditional document retrieval systems in which a user needs to read through the much longer content of the returned documents to determine their relevancy. Also, the relevant knowledge units are directly given by the answer scopes s_i's. The user need not go through a complete document to locate the knowledge. Furthermore, each returned question q_i explains how its corresponding answer s_i addresses the user's enquiry: if q_i is an accurate rephrase of the user's input, then s_i is the desired answer (see example in Table 1). This leads to the third advantage of a QB:

Adv. 3 (Explainability, Comprehensibility, and Efficiency): *Questions in a QB explain the relevancy of retrieval results, helping users to efficiently comprehend the extracted knowledge.*

So far, we have assumed the availability of a QB. We end this section with a brief discussion on how a QB is obtained. With advances in NLP techniques such as large language models (LLMs), there have been quite a few works on generating questions from text documents, especially in the area of Education, where the interests lie in generating assessments (questions) to evaluate students' mastering of knowledge conveyed in course materials (documents). For professional documents, [Yuan *et al.*, 2023] studies the problem of constructing a question bank from a corpus of web pages, each of which is a document that explains a specific legal topic in layperson's terms. In their study, *human-composed questions* (HCQs) and *machine-generated questions* (MGQs) are collected. The HCQs are manually written by legal experts who were instructed to ask questions for every aspect covered by the documents and to identify the answer (scope) for the questions. MGQs are machine-generated questions using the GPT-3 175B model. Readers are referred to [Yuan *et al.*, 2023] for technical details. In that study, the authors compare MGQs and HCQs w.r.t. several quantitative measures. Some advantages of MGQs over HCQs include lower cost and a higher number of questions. Moreover, it is reported that about 93% of the documents' contents are covered by the MGQs. Using machines to construct a QB is therefore a practical approach.

3 Experiments

In this section, we evaluate the performance of QBR. We describe the experiments conducted and present the results.

3.1 Experiment Settings

Data The Community Legal Information Centre (CLIC)[6] is a website that provides legal information to the general public in Hong Kong. Legal knowledge under different topics is presented in thousands of web pages. In [Yuan *et al.*, 2023], human-composed questions (HCQs) and machine-generated questions (MGQs) are derived from these CLIC web pages. For our experiment, we obtained a set of 1,359 CLIC pages as our document collection D. Also, we obtained 15,333 HCQs

[6] https://clic.org.hk/en

and 23,238 MGQs from [Yuan *et al.*, 2023] for a total of 38,571 questions (and their answer scopes) as our question bank QB. We use QB_H and QB_M to represent the sets of HCQs and MGQs, respectively. Note that our question bank $QB = QB_H \cup QB_M$.

To perform CL training, we construct training examples using QB. We follow the procedure described in the previous section to generate positive and negative examples of q-s pairs (Equations 2 and 3). The number of (+ve; -ve) examples obtained are (15,333; 125,977) and (23,238; 236,315) using QB_H and QB_M, respectively. We further augment the training set using GPT-augmentation. For this step, we select two least-distinguishable scopes from each document. Based on each selected scope s and their associated questions q's given in QB, we synthesize user input $\hat{u}$'s to form positive and negative $(\hat{u}, s)$ examples. We then sample (12,168; 72,747) examples from this pool and include those in the training set. The final training set has (50,739; 435,039) examples in total.

Finally, we sample 1,000 positive $(\hat{u}, s)$ examples that are not included in the training set as our test set U. For each $(\hat{u}, s) \in U$, the synthesized user input $\hat{u}$ has its ground-truth answer given by the scope s, and the ground-truth document is the one that contains s.

Performance metrics Recall that QBR carries out a two-step retrieval process, namely, *document retrieval* followed by *scope identification* within the retrieved documents. We therefore evaluate QBR's performance based on its accuracy in these two steps. Consider a test case user input $u \in U$ whose answer scope s^* is contained in document d^*. For document retrieval accuracy, we consider the top-k documents retrieved by QBR from dataset D. The *recall@k* of the retrieval result is 1 if d^* is among the top-k documents; 0 otherwise. Also, the *reciprocal rank* (rr) of the result is $1/i$, where i is the rank of d^* in the result. (Rank = 1 if d^* is ranked first in the result list; rr = 0 if d^* is out of top-k.) We report the average recall@k and mean reciprocal rank (denoted by MRR_d) computed over the whole test set U of 1,000 cases.

For scope identification accuracy, we assume the target document d^* is successfully retrieved, and so we consider the order in which QBR ranks the scopes within the document (i.e., the scopes given in the scope set S_{d^*}). We report the average reciprocal rank of target scopes s^* of all test cases in U. We denote this measure MRR_s to distinguish it from the ranking measure MRR_d, which was defined for the document retrieval step. We also report the average accuracy (acc), defined as the fraction of cases in which QBR pinpoints the correct scope s^* as the answer to the user input (i.e., s^* identified as top-ranked among all scopes in d^*).

Baseline Methods We compare QBR against a wide range of retrieval methods. These methods differ in the way they measure similarity between user input (queries) against documents in a corpus. In particular, we consider three categories of approaches. **Lexical** models (BM25 [Robertson and Zaragoza, 2009]) measure similarity based on word occurrences. **Sparse** retrieval models, (SPARTA [Zhao *et al.*, 2020], docT5query [Cheriton, 2019]), encode terms' occurrences in documents and queries using high-dimensional, sparse vectors. **Dense** retrieval models (BERT [Reimers and Gurevych, 2019], TinyBERT [Reimers and Gurevych, 2019], TAS-b [Hofstätter *et al.*, 2021], RoBERTa [Liu *et al.*, 2019], DPR [Karpukhin *et al.*, 2020], MPNet [Song *et al.*, 2020]) utilize low-dimensional, dense vector representations for documents and queries.

QBR employs an embedding function $T()$. In our experiment, QBR uses MPNet's embedding (as function $T()$) as the default. We remark that although QBR uses MPNet's embedding function, it differs from the original MPNet method because QBR uses the QB to enhance retrieval performance. First, in document retrieval, QBR compares the embedding vectors of user input u against the q-s pairs in QB (Equation 1), instead of against the document embedding vectors. Secondly, in scope identification, QBR uses CL to revise the embedding function (to $T'()$) to better disambiguate scopes within the same document. The details of CL model training are reported in the technical appendix[4].

3.2 Results

Document Retrieval Table 2 shows document retrieval performance comparing QBR against 11 baseline methods. Among the baselines, the lexical model BM25, which could be viewed as a robust benchmark for generalization [Thakur *et al.*, 2021], performs better than SPARTA (sparse model) and DPR (dense model) in terms of both recalls and MRR_d, while having comparable performance to TinyBERT. Among the sparse models, docT5query employs document expansion to capture an out-of-domain keyword vocabulary, which enhances performance over SPARTA. Dense models, particularly SBERT and MPNet, exhibit superior performance, showcasing a proficiency in understanding contextual information. With the exception of DPR and the small TinyBERT model, dense models, particularly MPNet, are better performers. QBR uses MPNet for its base embedding function, but it applies the QB in document retrieval by matching user input against the QB entries. Table 2 shows that this approach significantly improves performance. For example, for Recall@1, QBR (0.5400) is more than twice better than the traditional BM25 method (0.2540) and is much better than the best of the baselines, MPNet (0.4500). Previously we mentioned 3 advantages of using a QB. The results in Table 2 show strong evidence of Adv. 1 (*Document Augmentation*); Questions provided in the QB serve as important information that facilitates document representation, leading to much more effective document retrieval.

We remark that traditional retrieval methods, such as the baselines shown, return the highly ranked *documents* to users in response to their enquiries. To filter the returned results, a user needs to read through the returned documents to find the answer. In contrast, QBR returns (q, s) pairs (see Figure 1b) and the user needs to read only the questions q's returned and (if needed) their answer scopes s's to understand if any of the returned entries answer the enquiry (see Table 1 for an example). The amount of content read is thus much smaller. For example, an average document in our collection D contains 5 scopes (i.e., 5 knowledge units). If we use MPNet as the method for document retrieval, then the probability of finding the answer from the top-ranked document is 45% (MPNet's Recall@1 = 0.4500). For the same amount of time

	Lexical	Sparse Models		Dense Models									
	BM25	SPARTA	docT5query	TinyBERT	BERT	RoBERTa	ANCE	DPR	TAS-B	SBERT	MPNet	QBR
Recall@1	0.2540	0.1890	0.3300	0.2770	0.3840	0.3300	0.3670	0.1920	0.3900	0.4520	0.4500	**0.5400**
Recall@3	0.4160	0.3170	0.4950	0.4130	0.5250	0.5100	0.5540	0.3150	0.5700	0.6370	0.6670	**0.7230**
Recall@5	0.5090	0.3820	0.5800	0.4940	0.6060	0.5750	0.6250	0.3850	0.6520	0.7180	0.7360	**0.8050**
MRR_d	0.3584	0.2697	0.4333	0.3666	0.4763	0.4364	0.4801	0.2758	0.4993	0.5639	0.5739	**0.6482**

Table 2: Document retrieval performance

	TinyBERT	BERT	RoBERTa	ANCE	DPR	TAS-B	SBERT	MPNet	QBR	$\text{QBR}_{\neg\text{GPT}}$
acc	0.4390	0.4650	0.4500	0.4410	0.3580	0.4390	0.4870	0.5000	**0.8370**	0.6460
MRR_s	0.6638	0.6763	0.6683	0.6570	0.5964	0.6631	0.6963	0.7061	**0.9100**	0.7860

Table 3: Scope identification performance

(to read one document with 5 scopes), with QBR, the user could have browsed 5 (q, s) entries in the returned results. The probability of the user finding the answer among them is 80.5% (QBR's Recall@5 = 0.8050), which is almost twice as likely as for the case of MPNet. From the perspective of user search efforts, QBR is much more efficient than all other approaches. The experimental results thus show evidence of Adv. 3 (*Explanability, Comprehensibility, and Efficiency*).

Scope Identification Step 2 of QBR involves identifying the correct scope s^* within the target document d^*. This step can also be done using the baseline methods by applying them to rank the scopes of d^*. However, QBR utilizes the QB to perform contrastive learning (CL) to adjust its embedding function ($T'()$) so as to better disambiguate scopes within the same document. Furthermore, QBR uses GPT to augment the CL training set. Given the correct document d^*, Table 3 shows scope identification performance comparing QBR against the dense model baselines. The column labeled "$\text{QBR}_{\neg\text{GPT}}$" refers to QBR without GPT-augmentation in CL training. From the table, we see that none of the baselines has its acc exceed 0.5. That means they select the wrong answer scope more often than they pick the right one. Scope identification is therefore a very difficult task due to the very similar semantics of the scopes within the same document. By applying CL to obtain a more discerning embedding function $T'()$, QBR gives a much higher acc at 0.837. Moreover, QBR has an MRR_s of 0.91, which is very close to 1.0. This indicates that even for the cases where QBR does not rank s^* first, s^* is ranked very highly by QBR. By comparing the performance of QBR with $\text{QBR}_{\neg\text{GPT}}$, we see a significant drop in acc (0.8370 → 0.6460) and MRR_s (0.9100 → 0.7860) if we take away GPT-augmentation. This shows that the training examples obtained via GPT to mimic user input is highly effective. Nevertheless, the scores of $\text{QBR}_{\neg\text{GPT}}$ are still way higher than those of the baselines. This again shows the effectiveness of CL. These results support Adv. 2 (Fine-grained Retrieval).

3.3 Demo and Additional Experiments

We have conducted additional experiments to further assess QBR's effectiveness. Additionally, we have deployed QBR on a real platform to help the public comprehend the law based on their specific legal circumstances. Due to space constraints, we provide a concise summary of our experimental findings. For more detailed information about the experimental results and a demonstration of the deployed platform, please refer to the technical appendix[4].

QB Quality We investigate how the QB quality affects QBR's performance. Intuitively, a good QB should (1) have rich contents (i.e., enough questions) that cover all knowledge units presented in the documents and (2) be well-phrased and relevant to the knowledge units [Yuan *et al.*, 2023]. First, we investigated the performance of QBR w.r.t. the size of QB. We observe that QBR's performance progressively improves as we increase the QB's size. However, we observe that even a small QB drastically improves scope identification accuracy (acc). Specifically, a (small) QB with 10K questions achieves an acc of 0.719, which is much higher than MPNet (acc = 0.500). This shows that the QB provides critical information for disambiguating scopes and our CL approach is highly effective even with a small QB. We further evaluate QBR using three versions of the question bank: QB_H, QB_M, and $QB = QB_H \cup QB_M$. We observe that QB_H and QB_M give very similar performance with QB_H having a slight edge over QB_M. Also, both of them outperform MPNet by significant margins. The complete QB gives the best performance, showing the complementary nature of machine and human questions.

Language Models We conducted experiments using different language models (in addition to MPNet) to derive the embedding function $T()$ QBR employs. We observe similar performance advantages of QBR over other methods. QBR is therefore a general approach that can work with different representation techniques.

Scalability Our system is efficient and scalable. Query execution time is dominated by the step that identifies the most relevant (q, s) to a user query. This step can be efficiently processed using a vector database. For example, with a QB of 38,571 questions, the average search time is 0.018s. We have conducted an experiment changing the QB size and found that the search time stays fairly stable. The scalability comes from effective vector indexing.

Generalization to Other Domains In addition to legal knowledge retrieval, our QBR approach can be applied to various professional domains, including medical and financial. To illustrate QBR's versatility, we extend our study to medical knowledge retrieval. We utilized QBR with medical data and conducted comparable experiments. Similar conclusions regarding the performance of QBR on document-level and scope-level retrieval can be drawn, which underscores the wide applicability of this method across different domains.

4 Social Impact and Case Studies

We have deployed QBR on the online legal information platform CLIC[6] and conducted user studies. Consent to participate in the studies have been duly obtained from all participants. In the studies, participants were presented with hypothetical scenarios and were instructed to conduct searches for legal information using both their customary approach and QBR. Subsequently, they were asked to compare their search results. Some of the participants were NGO workers who provide services to families and victims of domestic violence cases. All participants provided overwhelmingly positive feedback regarding QBR's efficacy. For instance, in one scenario, participants were tasked with providing guidance to a domestic violence victim on divorce procedures and available protective measures for them and their children. Although participants were aware of the existence of court orders, they could not recall the precise legal action appropriate for the case. Many of them searched for "protection order" when the term they needed to find was "injunction order". They were able to find pertinent and helpful information on the platform with QBR but their searches failed when they follow their traditional keyword search methods. They all perceived QBR to be highly advantageous to their work since even experienced social workers may be perplexed by complex legal terminology. We also study how the information on CLIC and QBR can be utilized by a social work NGO for consultations on youth matters. The social workers believe that legal information written in layperson's terms can help them "thoroughly understand the law in order to effectively assist clients and their parents" when they come into contact with individuals who have been arrested. The QBR tool has garnered positive feedback and has proven to be effective in aiding members of the public in resolving legal problems.

Using QBR on a legal information platform like CLIC is impactful in upholding social justice, as the tool empowers the underprivileged community to be better legally informed. As an example, we interviewed a CLIC user, who works in the field of accounting and financial planning. The user faced two litigation cases: one involves himself and the other involves his small company. As the plaintiff had engaged a team of legal practitioners, the interviewee faced a knowledge imbalance. He tried to seek help from various channels such as government legal advice schemes and court guidance notes to no avail. He felt overwhelmed as he did not understand the legal procedures due to the high technicality and brevity of information. Eventually, he tried CLIC and was able to comprehend the law and the relevant procedures with the information the CLIC platform provided. Without hiring a legal representative, he was able to defend his cases and deter the opposing parties from pursuing the case further. As he said in the interview, "CLIC has been extremely helpful to us in preparing for the litigation".

5 Related Works

Comprehensibility of Legal Information [Mommers, 2011] categorizes the accessibility of legal information into three levels: primary availability, where documents are available and searchable online; secondary availability, where links are established between relevant documents; tertiary availability, where contents are clarified and translated into languages understandable by the target audience. [Mommers, 2011] further studies two legal database websites in Europe and finds that tertiary availability is largely ignored. [Dyson and Schellenberg, 2017] conducts a readability study on 407 passages extracted from Legal Services Corporation-sponsored websites and find that most of them are beyond comprehension by normal citizens, which contradicts the aim of the legal aid to serve those with low income and low literacy. [Curtotti and McCreath, 2013] analyzes the language features and study the readability of Australian legal documents. Aligned with previous empirical studies [Pi and Schmolka, 2000; Tanner, 2002; Abrahams, 2003], they conclude that the linguistic characteristics of legal documents are quite different from normal English and that legal articles are generally more difficult to read.

Domain-Specific IR [Bonifacio *et al.*, 2022] highlights the superiority of domain-specific training data for IR tasks, proposing LLM-generated synthetic data creation. [Yao *et al.*, 2023] addresses Legal Evidence Retrieval (LER) to enhance judicial efficiency in evidence discovery. Current limitations include inadequate user-centric design: non-expert users often struggle with domain-specific terminology, hindering effective query formulation.

Contrastive Learning (CL) Recent CL advancements enhance retrieval systems across scenarios. [Ma *et al.*, 2021] improves neural rankers' out-of-domain resilience through contrastive fine-tuning. [Abdollah Pour *et al.*, 2023] develops self-supervised CL for reviewed-item retrieval, while [Sidiropoulos and Kanoulas, 2024] employs multi-positive CL to handle typo-ridden queries. Inspired by these, we integrate QB with CL to refine granularity in legal IR for novice users.

6 Conclusion

In this paper we proposed QBR to perform fine-grained domain-specific knowledge retrieval for non-expert users. The challenge involves managing noisy and imprecise user input, as well as distinguishing and ranking semantically closely related scopes within the same document. QBR tackles the problems through the use of a question bank (QB), which provides three advantages. By matching user input against QB's entries, and performing contrastive learning based on QB's data, we show that QBR significantly outperforms existing methods in terms of both document retrieval and scope identification. Moreover, by returning rephrased questions and answer scopes instead of documents, QBR makes retrieval results more comprehensible and explainable, leading to a more efficient user experience. In case studies, we showed that QBR makes a social impact by helping users resolve day-to-day legal issues.

Acknowledgements

This research is supported by the WYNG Foundation (AR25AG100407).

References

[ABA Free Legal Answers, 2023] ABA Free Legal Answers. Aba free legal answers 2023 summer report. https://www.americanbar.org/content/dam/aba/administrative/probono_public_service/abafree/rpts/2023/2023-summary-report.pdf, 2023.

[Abdollah Pour *et al.*, 2023] Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Armin Toroghi, Anton Korikov, Ali Pesaranghader, Touqir Sajed, Manasa Bharadwaj, Borislav Mavrin, and Scott Sanner. Self-supervised contrastive bert fine-tuning for fusion-based reviewed-item retrieval. In *ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, page 3–17, Berlin, Heidelberg, 2023. Springer-Verlag.

[Abrahams, 2003] Eloise Abrahams. *Efficacy of plain language drafting in labour legislation*. PhD thesis, Peninsula Technikon, 2003.

[Basch *et al.*, 2020] Corey H. Basch, Jan Mohlman, Grace C. Hillyer, and Philip Garcia. Public health communication in time of crisis: Readability of on-line covid-19 information. *Disaster Medicine and Public Health Preparedness*, 14(5):635–637, 2020.

[Bonifacio *et al.*, 2022] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2387–2392, New York, NY, USA, 2022. Association for Computing Machinery.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[Cheriton, 2019] David R. Cheriton. From doc2query to docTTTTTquery. 2019.

[Curtotti and McCreath, 2013] Michael Curtotti and Eric McCreath. A right to access implies a right to know: An open online platform for research on the readability of law. *J. Open Access L.*, 1:1, 2013.

[Dyson and Schellenberg, 2017] Dana D Dyson and Kathryn Schellenberg. Access to justice: The readability of legal services corporation legal aid internet services. *Journal of poverty*, 21(2):142–165, 2017.

[Ferguson *et al.*, 2021] Catherine Ferguson, Margaret Merga, and Stephen Winn. Communications in the time of a pandemic: the readability of documents for public consumption. *Australian and New Zealand Journal of Public Health*, 45(2):116–121, 2021.

[Hofstätter *et al.*, 2021] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proc. of SIGIR*, 2021.

[Hutchinson *et al.*, 2016] Nora Hutchinson, Grayson L Baird, and Megha Garg. Examining the reading level of internet medical information for common internal medicine diagnoses. *The American journal of medicine*, 129(6):637–639, 2016.

[Karpukhin *et al.*, 2020] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[Ma *et al.*, 2021] Xiaofei Ma, Cicero Nogueira dos Santos, and Andrew O Arnold. Contrastive fine-tuning improves robustness for neural rankers. *arXiv preprint arXiv:2105.12932*, 2021.

[Mommers, 2011] Laurens Mommers. Access to law in europe. In *Innovating Government: Normative, Policy and Technological Dimensions of Modern Government*, pages 383–398. Springer, 2011.

[OpenAI, 2022] OpenAI. Openai: Introducing chatgpt. *URL https://openai.com/blog/chatgpt.*, 2022.

[Patiño and others, 2019] Camilo Gutiérrez Patiño et al. Global insights on access to justice, 2019.

[Pi and Schmolka, 2000] G Pi and V Schmolka. A report on results of usability testing research on plain language draft sections of the employment insurance act: A report to department of justice canada and human resources development canada, 2000.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[Robertson and Zaragoza, 2009] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr 2009.

[Ruohonen, 2021] Jukka Ruohonen. Assessing the readability of policy documents on the digital single market of the european union. In *2021 Eighth International Conference on eDemocracy and eGovernment (ICEDEG)*, pages 205–209, 2021.

[Sidiropoulos and Kanoulas, 2024] Georgios Sidiropoulos and Evangelos Kanoulas. Improving the robustness of dense retrievers against typos via multi-positive contrastive learning. page 297–305, Berlin, Heidelberg, 2024. Springer-Verlag.

[Song *et al.*, 2020] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pretraining for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[Tanner, 2002] Edwin Tanner. Seventeen years on: is victorian legislation less grammatically complicated? *Monash University Law Review*, 28(2):403–423, 2002.

[Thakur *et al.*, 2021] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Thirty-seventh Conference on Neural Information Processing Systems*, 2021.

[van den Oord *et al.*, 2019] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[Yao *et al.*, 2023] Feng Yao, Jingyuan Zhang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Yun Liu, and Weixing Shen. Unsupervised legal evidence retrieval via contrastive learning with approximate aggregated positive. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[Yuan *et al.*, 2023] Mingruo Yuan, Ben Kao, Tien-Hsuan Wu, Michael MK Cheung, Henry WH Chan, Anne SY Cheung, Felix WH Chan, and Yongxi Chen. Bringing legal knowledge to the public by constructing a legal question bank using large-scale pre-trained language model. *Artificial Intelligence and Law*, pages 1–37, 2023.

[Zhao *et al.*, 2020] Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. Sparta: Efficient open-domain question answering via sparse transformer matching retrieval, 2020.