

## Technical Appendix

### QBR Demo

We have deployed QBR on CLIC, an online legal information platform, to recommend relevant legal questions and answers to users. Figure 1 illustrates the system with an example input. It provides a user-friendly interface and guides users through three key steps: describing the situation (Figure 1a), choosing relevant topics (Figure 1b), and reading/visiting recommendations (Figure 1c).

For example, a mother finds herself in an abusive relationship and wishes to pursue a divorce but she lacks knowledge about legal process. Instead of formulating precise legal questions, she can simply describe her current situation to the system (Figure 1a). After providing the input, the system prompts her to select one or more topics related to her situation (Figure 1b). Although we did not detail the topic filtering procedure in the paper, this process is designed to help users focus on the areas they wish to explore. Finally, the system presents her with a list of relevant questions that match her input (Figure 1c). Additionally, the system offers options for her to view relevant excerpts or to visit the original CLIC page for the full content. The deployment of QBR eases the users’ burden by allowing them to access legal information without the need to navigate through the comprehensive and intricate CLIC website.

### Contrastive Learning Model Training

For CL training, we use AdamW optimizer with a learning rate of  $1e-5$ . We train for 5 epochs with maximum token length 128. The temperature hyper parameter  $\tau$  is 0.1.

### Detailed experiment results

In this section, we provide full results of the experiments.

**QB quality** Our  $QB$  has 38,571 questions in total. We investigate how its size  $|QB|$  affects QBR’s performance. We perform uniform subsampling to obtain QBs of various sizes. Tables 1 and 2 show document retrieval and scope identification performance, respectively, as we vary  $|QB|$  in steps of 10k questions. From the tables, we see that QBR’s performance progressively improves as we increase the QB’s size. A notable point shown in Table 2 is that even with a small  $QB$  (10k), compared with no QB (0 questions), scope identification is drastically improved ( $acc$ :  $0.5 \rightarrow 0.719$ ;  $MRR_s$ :  $0.7061 \rightarrow 0.8345$ ). This shows that the QB provides critical information for disambiguating scopes and our CL approach is highly effective even with a small question bank.

For the legal dataset, our question bank  $QB$  consists of 15,333 human-composed questions ( $QB_H$ ) and 23,238 machine-generated questions ( $QB_M$ ). Generally, human questions are precise and cover almost perfectly the whole document set, while machine questions are mostly precise and give good coverage. Moreover, machine questions are cheaper to obtain and so they are more numerous. Tables 3 and 4 show QBR’s performance using  $QB_H$ ,  $QB_M$ , and the complete  $QB$ . MPNet’s is also shown as a comparison baseline.

From Table 3, in document retrieval, we see 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. This shows that machine questions are competitive against human ones and both are highly useful. By combining  $QB_H$  and  $QB_M$ , the final question bank  $QB$  is even richer in content, which further hoists QBR’s performance. The performance numbers under  $QB$  in Table 3 are therefore significantly better than others’. This shows that human and machine can work complementarily to construct a rich question bank. Table 4 shows scope identification performance with different QBs. We see that in this case  $QB_M$  gives better performance than  $QB_H$ . This is because there are more machine questions than human questions. Therefore,  $QB_M$  helps generate more CL training examples compared with  $QB_H$ , leading to better scope identification. Other conclusions are similar to those previously mentioned: Human and machine questions complement each other and they are highly useful for QBR to achieve accurate scope identification.

**QBR with different language models** We have conducted experiments using different language models (in addition to MPNet) to derive the embedding function  $T()$  QBR employs to demonstrate that QBR can be incorporated with various embedding methods. We replace the embedding function  $T()$  of QBR with different embedding methods and evaluate its performance on scope identification. Recall that QBR obtains the adjusted embedding ( $T'()$ ) through contrastive learning and LLM-augmentation. The results are reported in Table 5, where “Original” shows the performance of using  $T()$  directly to perform scope identification and “QBR” shows that of using QBR’s adjusted embedding ( $T'()$ ). We observe that QBR shows significant advantage over the original embedding methods. For example, the Recall@1 scores of TinyBERT and QBR (with TinyBERT as the baseline embedding) are 0.4390 and 0.8080, respectively. QBR is therefore a general approach that can work with different representation techniques.

**Medical Domain** In addition to the application in legal contexts, our approach also extends to other professional fields, such as medicine. The National Institutes of Health (NIH)<sup>1</sup> is the primary agency of the United States government responsible for biomedical and public health research. It is composed of 27 different institutes and centers, each focuses on specific areas of health and disease investigation. We collected 100 medical pages from NIMH<sup>2</sup>, NIAMS<sup>3</sup>, NIAAA<sup>4</sup> and NEI<sup>5</sup>. Based on the medical documents, we obtained 10,393 MGQs (and their answer scopes) as the medical question bank. To perform CL training, we construct training examples using the bank. We generate (10,393; 63,805) positive and negative examples for MGQs. We further augment the training set using LLM-augmentation and obtained (1,848; 10,928) positive and negative examples. The final training set has (12,241;

<sup>1</sup><https://www.nih.gov/>

<sup>2</sup><https://www.nimh.nih.gov/>

<sup>3</sup><https://www.niams.nih.gov/>

<sup>4</sup><https://www.niaaa.nih.gov/>

<sup>5</sup><https://www.nei.nih.gov/>

108 74,733) examples in total. Finally, we sample 300 positive  
109  $(\hat{u}, s)$  examples that are not included in the training set as our  
110 test set  $U$ .

111 We deployed QBR with the medical data and conducted  
112 analogous experiments. Table 6 and Table 7 show the doc-  
113 ument retrieval and scope identification performance on the  
114 Medical Dataset. From the two tables, we can derive similar  
115 conclusions regarding performances on document-level and  
116 scope-level retrieval, demonstrating that the wide applicabil-  
117 ity of QBR in different domains.

118 **Document retrieval with QB** In the paper, we show the  
119 performance of baseline model in terms of document-level  
120 retrieval. Here we examine the effectiveness of combining  
121 QB with individual methods and evaluate the performance  
122 for each baseline approach. The results, illustrated in Table 8  
123 shows a significant enhancement across all models. For ex-  
124 ample, the Recall@1 performance of BM25 is improved from  
125 0.2540 to 0.3120. This global improvement demonstrates  
126 the effectiveness of introducing QB in the domain-specific  
127 document retrieval process irrespective of the retrieval model  
128 used. Thus, our proposed method of using QB is an effec-  
129 tive and broadly applicable approach in improving domain-  
130 specific document retrieval.

## 1 Describe your situation

Write about your situation or ask a legal question

CLICK TO START VOICE INPUT or write below

I am in an abusive relationship with my husband. I am considering divorce, but I am not familiar with the procedures and cost involved. Also, I am afraid that seeking a divorce might provoke revenge from my husband. I am unsure about what steps to take and how to ensure my daughter's and my safety.

CONTINUE

### Jumpstart Examples

#### Is this illegal?

Recently, I recorded a TV drama with the intention of viewing it at a later time. I'm wondering if this act of directly videotaping TV dramas that are broadcast at home infringes upon copyright laws.

#### How can I handle this properly?

Is it necessary for me to stamp "copy" on a photocopy of someone else's ID card?

#### I am worried about this...

I'm inquiring on behalf of a friend who is employed on a ship. My friend's company has recently installed a CCTV system, asserting that it provides comprehensive surveillance coverage from various angle...

Show More

### (a) Step 1: Describe the situation using text or voice input

## 2 Choose topics

Click to select the topics that you consider relevant

Family Matrimonial & Cohabitation

Legal Aid

SEARCH

### (b) Step 2: Choose the relevant topics

## Read and visit recommendations

3 Your scenario might be addressed by the following CLIC pages represented by the model questions below. Click "VISIT PAGE" to visit the relevant CLIC pages or paragraphs.

RESTART

GO BACK AND EDIT STEP 1

QUESTION #1

Family Matrimonial & Cohabitation

### How to apply for a divorce?

<https://clic.org.hk/en/topics/familyMatrimonialAndCohabitation/divorce/proceduresAndGroundsForDivorce/q2>

Collapse Preview

HOW DO I APPLY FOR DIVORCE? (WITH A BRIEF SUMMARY OF THE RELEVANT PROCEDURES) You are advised to consult a lawyer before submitting the relevant documents and attending hearings in the Family Court. You need to go through the following stages:- Stage 1 is presenting a petition for divorce. Stage 2 is serving (delivering) my petition to my spouse. Stage 3 is fixing a court hearing date. Stage 4 is obtaining a Decree Nisi - a tentative court order for divorce. Stage 5 is the final order for divorce.

This excerpt is taken from a CLIC page and does not provide a complete answer or detailed information on the question. To view the original content, please click on 'Visit Page' below.

VISIT PAGE

Is this recommendation relevant?

☆☆☆☆

QUESTION #2

Family Matrimonial & Cohabitation

### I plan to file for divorce. What legal documents may help?

[https://clic.org.hk/en/topics/familyMatrimonialAndCohabitation/marraige\\_and\\_cohabitant\\_issues/nuptial\\_agreements/separation\\_agreements/merits\\_of\\_entering\\_a\\_separation\\_agreement](https://clic.org.hk/en/topics/familyMatrimonialAndCohabitation/marraige_and_cohabitant_issues/nuptial_agreements/separation_agreements/merits_of_entering_a_separation_agreement)

Preview Relevant Excerpt

VISIT PAGE

Is this recommendation relevant?

☆☆☆☆

### (c) Step 3: Read recommended questions, excerpts and CLIC pages

Figure 1: Demo of the deployed QBR

	0	10k	20k	30k	All (38.57k)
Recall@1	0.4500	0.4930	0.5180	0.5320	0.5400
Recall@3	0.6670	0.6690	0.7020	0.7130	0.7230
Recall@5	0.7360	0.7610	0.7840	0.7870	0.8050
$MRR_d$	0.5739	0.5994	0.6275	0.6379	0.6482

Table 1: Document retrieval performance vs.  $|QB|$  (Legal Dataset)

	0	10k	20k	30k	All (38.57k)
$acc$	0.5000	0.7190	0.7420	0.8130	0.8370
$MRR_s$	0.7061	0.8345	0.8547	0.8961	0.9100

Table 2: Scope identification performance vs.  $|QB|$  (Legal Dataset)

	MPNet	QBR		
		$QB_H$	$QB_M$	$QB$
Recall@1	0.4500	0.5160	0.5090	0.5400
Recall@3	0.6670	0.7050	0.6990	0.7230
Recall@5	0.7360	0.7760	0.7700	0.8050
$MRR_d$	0.5739	0.6242	0.6192	0.6482

Table 3: Document retrieval performance with different Legal QBs

	MPNet	QBR		
		$QB_H$	$QB_M$	$QB$
$acc$	0.5000	0.7450	0.7810	0.8370
$MRR_s$	0.7061	0.8553	0.8728	0.9100

Table 4: Scope identification performance with different Legal QBs

	TinyBERT		BERT		RoBERTa		ANCE		DPR		TAS-B		SBERT		MPNet	
	Original	QBR	Original	QBR	Original	QBR	Original	QBR	Original	QBR	Original	QBR	Original	QBR	Original	QBR
Recall@1	0.4390	0.8080	0.4650	0.8260	0.4500	0.8180	0.4410	0.8170	0.3580	0.8140	0.4390	0.8110	0.4870	0.8200	0.5000	0.8370
$MRR_s$	0.6638	0.8915	0.6763	0.9022	0.6683	0.8974	0.6570	0.8982	0.5964	0.8986	0.6631	0.8946	0.6963	0.8992	0.7061	0.9100

Table 5: Scope identification of QBR with different embedding methods (Legal Dataset)

	Lexical	Sparse Models		Dense Models								
	BM25	SPARTA	docT5query	TinyBERT	BERT	RoBERTa	ANCE	DPR	TAS-B	SBERT	MPNet	QBR
Recall@1	0.4067	0.2433	0.3867	0.4467	0.5067	0.3900	0.4833	0.3467	0.5033	0.5333	0.6133	<b>0.6467</b>
Recall@3	0.5867	0.4100	0.6067	0.6067	0.6700	0.5467	0.6167	0.5400	0.6700	0.7167	0.6783	<b>0.7733</b>
Recall@5	0.6933	0.5067	0.6700	0.6833	0.7333	0.6400	0.6900	0.6500	0.7333	0.7700	0.6938	<b>0.8333</b>
$MRR_d$	0.5193	0.3564	0.5117	0.5508	0.5995	0.4963	0.5713	0.4661	0.6019	0.6385	0.7003	<b>0.7244</b>

Table 6: Document retrieval performance on Medical Dataset

	TinyBERT	BERT	RoBERTa	ANCE	DPR	TAS-B	SBERT	MPNet	QBR	QBR <sub>-GPT</sub>
$acc$	0.4767	0.5767	0.5233	0.5600	0.4633	0.5467	0.5967	0.6267	<b>0.7133</b>	0.6600
$MRR_s$	0.6635	0.7335	0.6900	0.7218	0.6403	0.7103	0.7455	0.7679	<b>0.8263</b>	0.7982

Table 7: Scope identification performance on Medical Dataset

Method		Lexical	Sparse		Dense						
		BM25	SPARTA	docT5query	TinyBERT	BERT	RoBERTa	ANCE	DPR	TAS-B	SBERT
w/o QB	Recall@1	0.2540	0.1890	0.3300	0.2770	0.3840	0.3300	0.3670	0.1920	0.3900	0.4520
	Recall@3	0.4160	0.3170	0.4950	0.4130	0.5250	0.5100	0.5540	0.3150	0.5700	0.6370
	Recall@5	0.5090	0.3820	0.5800	0.4940	0.6060	0.5750	0.6250	0.3850	0.6520	0.7180
	$MRR_d$	0.3584	0.2697	0.4333	0.3666	0.4763	0.4364	0.4801	0.2758	0.4993	0.5639
w/ QB	Recall@1	0.3120	0.2470	0.3990	0.3480	0.4530	0.4130	0.4150	0.2080	0.4330	0.5340
	Recall@3	0.4340	0.4090	0.5650	0.5160	0.6250	0.5730	0.5620	0.3380	0.6310	0.7090
	Recall@5	0.5010	0.4770	0.6160	0.5830	0.6890	0.6450	0.6260	0.4030	0.6840	0.7870
	$MRR_d$	0.3921	0.3468	0.4915	0.4521	0.5551	0.5116	0.5063	0.2922	0.5413	0.6359

Table 8: Results of document-level retrieval across all baselines with and without QB