H-LegalKI: A Hierarchical Legal Knowledge Integration Framework for Legal Community Question Answering

Yue Jiang, Ziyu Guan[*]**, Jie Zhao, Wei Zhao** and **Jiaqi Yang**
School of Computer Science and Technology, Xidian University, Xi'an, 710126, China
`{22031212489@stu., zyguan@, jzhao1992@stu., ywzhao@mail., jiaqiyang@stu.}xidian.edu.cn`

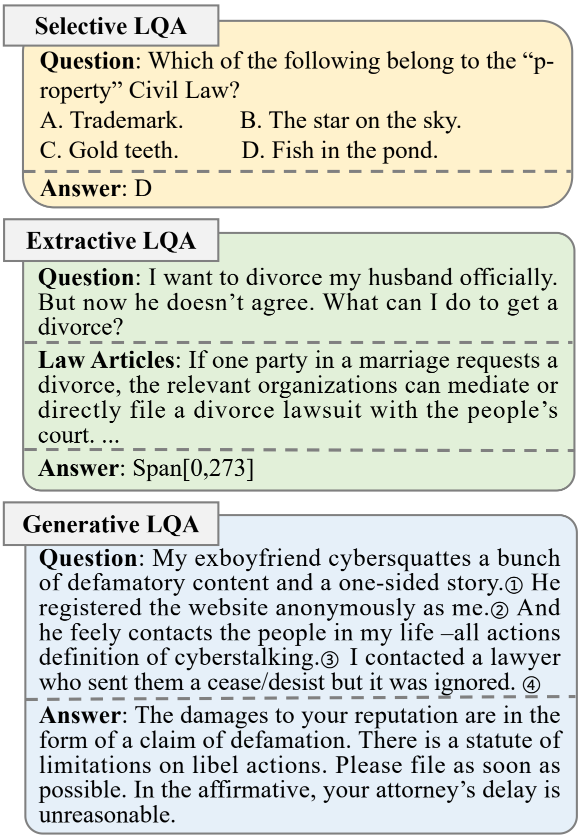
Abstract

Legal question answering (LQA) aims to bridge the gap between the limited availability of legal professionals and the high demand for legal assistance. Traditional LQA approaches typically either select the optimal answers from an answer set or extract answers from law texts. However, they often struggle to provide relevant answers to complex, real-world questions due to the rigidity of predetermined answers. Although recent advancements in legal large language models have shown some potential in enhancing answer relevance, they fail to address the multiple user-specific circumstances, i.e., factual details in questions.

To address these issues, we (1) construct the first publicly available legal community question-answering (LegalCQA) dataset; and (2) propose a Hierarchical Legal Knowledge Integration (H-LegalKI) framework. LegalCQA is collected from two widely used legal forums for developing user-centered LQA models. For H-LegalKI, we design a legal knowledge retriever that gathers comprehensive legal knowledge based on both entire questions and individual sentences. And an answer generation model is designed to understand question- and sentence-level factual details and integrate corresponding legal knowledge in a hierarchical way. Additionally, we design a de-redundancy module to remove redundant legal knowledge. Experiments on LegalCQA demonstrate the superiority of our framework over competitive baselines.

1 Introduction

Legal question answering (LQA) is expected to address the disparity between the limited number of legal professionals (Zhong et al., 2020b) and the extensive volume of legal issues (Louis et al., 2024). As a significant application of legal artificial intelligence, LQA can provide professional legal advice to assist ordinary individuals in protecting their rights. Additionally, for legal professionals, it offers a handy reference to relieve the burdensome work and increase their working efficiency.



Selective LQA

Question: Which of the following belong to the "p-roperty" Civil Law?
A. Trademark. B. The star on the sky.
C. Gold teeth. D. Fish in the pond.

Answer: D

Extractive LQA

Question: I want to divorce my husband officially. But now he doesn't agree. What can I do to get a divorce?

Law Articles: If one party in a marriage requests a divorce, the relevant organizations can mediate or directly file a divorce lawsuit with the people's court. ...

Answer: Span[0,273]

Generative LQA

Question: My exboyfriend cybersquattes a bunch of defamatory content and a one-sided story.① He registered the website anonymously as me.② And he feely contacts the people in my life –all actions definition of cyberstalking.③ I contacted a lawyer who sent them a cease/desist but it was ignored. ④

Answer: The damages to your reputation are in the form of a claim of defamation. There is a statute of limitations on libel actions. Please file as soon as possible. In the affirmative, your attorney's delay is unreasonable.

Figure 1: Comparisons of different kinds of LQA.

Generally, there are two main research focuses in answering legal questions. The first one involves selecting the correct or optimal answer from a list of potential answers or an answer set. For example, Kano et al. (2019); Zhong et al. (2020c); Zheng et al. (2021) collect single/multiple-choice questions from judicial examinations, construct QA pairs, and study LQA as a classification problem. Mansouri and Campos (2023) construct an answer database, including responses sourced from legal

[*] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14614–14625
November 12-16, 2024 ©2024 Association for Computational Linguistics

forums, and retrieve the optimal answer for a given question. The other research line focuses on extracting answers from law articles or regulations (Zhong et al., 2020a; Chen et al., 2023), which predict the start and end indices over the relevant legal text. We present examples of QA pairs for both answer selection and extraction in Figure 1.

While these works have made progress in LQA, they struggle to effectively and relevantly answer complex legal questions since the answers are actually predetermined. Recently, large language models (LLMs) like GPT-4 (Achiam et al., 2023) and LLaMaA (Touvron et al., 2023) have demonstrated remarkable capabilities in various natural language processing tasks, including QA. Several studies train LLMs with legal corpus and develop a series of specialized legal LLMs, such as DISC-LawLLM (Yue et al., 2023), Chatlaw (Cui et al., 2023), and WisdomInterrogatory[1]. These models can improve the relevance of generated answers to the questions. However, they are primarily trained on professional legal knowledge, which is not aligned with the users' multiple practical circumstances, i.e., factual details described in questions. As a result, they struggle to meet users' specific needs, as shown in our experimental results (4.5 & 4.6).

To address these issues, this work concentrates on developing an answer-generation model based on real-world QA data from online legal forums. Users seeking legal assistance typically engage with these forums by asking comprehensive questions. And legal professionals contribute highly question-relevant and user-centered answers. By carefully designing a specialized LQA framework to utilize these QA pairs, we aim to bridge the gap between the practical needs of users and professional legal knowledge.

Specifically, we first collect questions and answers from legal forums and construct a legal community QA dataset, named LegalCQA, to address the lack of real-world QA data. LegalCQA consists of Chinese and English sub-datasets, and covers a wide range of questions with various law categories, e.g., civil law, criminal law, and contract law (more details can be found in 4.1). We show a QA pair in LegalCQA at the bottom of Figure 1. This example illustrates that the question includes more factual details (e.g., slander, personal information tort, and civil action) compared to the other two types of questions. The complexity of user-raised questions, often including multiple circumstances and involving various legal knowledge, presents a challenge for traditional open-domain question-answering methods. Effectively capturing these circumstances and integrating relevant legal knowledge to generate high-quality responses is far from trivial.

We then propose a novel Hierarchical Legal Knowledge Integration (H-LegalKI) framework for LQA. There are two key components of the proposed framework: (1) A designed legal knowledge retriever is responsible for retrieving hierarchical legal knowledge. Specifically, we split the original question q into individual sentences s, and retrieve the involved legal items (e.g., law articles and regulations) for q and all s, respectively. This allows us to acquire more comprehensive legal knowledge to support the answer generation; (2) A novel answer generation model is designed to make use of the hierarchical information (question- and sentence-level) from both questions and the retrieved legal items. We fuse the representations of question q to its corresponding legal items to obtain question-specific legal knowledge representation. And the same process is applied to each sentence. Additionally, we design a simple de-redundancy module based on averaging operation to refine legal knowledge representations, as the retrieved legal items may contain duplicate entries.

Our main contributions are summarized as:

- To the best of our knowledge, we provide the first solution for user-centered LQA which comprehensively studies the multiple circumstances in users' questions.
- A two-stage hierarchical framework is designed to integrate legal knowledge to support the answer generation.
- We contribute a public legal community QA dataset LegalCQA. We conduct plentiful experiments on LegalCQA to verify the effectiveness of H-LegalKI and show its significant improvements over competitive baselines.

2 Related Work

2.1 Legal Question Answering

Recently, rapidly growing attention has emerged on legal artificial intelligence (Zhong et al., 2020b), such as legal case retrieval (Shao et al., 2020; Li et al., 2023; Zhao et al., 2024), legal judgment

[1] https://github.com/zhihaiLLM

Dataset	QA pairs	Answer type	Language	Legal system	Data source	Downstream work
COLIEE-2018-Q (Kano et al., 2019)	720	Binary-choice	Japanese	Statute law	Law exam	Answer Selection
JEC-QA (Zhong et al., 2020c)	26365	Multi-choice	Chinese	Statute law	Law exam	Answer Selection
QAS4CQAR (Zhong et al., 2020a)	3500	Long-form	Chinese	Statute law	Legal institution	Answer Extraction
CaseHOLD (Zheng et al., 2021)	52800	Multi-choice	English	Case law	Law exam	Answer Selection
EQUALS (Chen et al., 2023)	6914	Long-form	Chinese	Statute law	legal forum	Answer Extraction
FALQU (Mansouri and Campos, 2023)	9880	Long-form	English	Case law	legal forum	Answer Selection
LLeQA (Louis et al., 2024)	1868	Long-form	French	Statute law	legal forum	Answer Generation
LegalCQA (ours)	21780 8899	Long-form	Chinese English	Statute law Case law	legal forum	Answer Generation

Table 1: Comparisons of public LQA datasets.

prediction (Xu et al., 2020; Zhao et al., 2022, 2023), as well as LQA (Zhong et al., 2020c; Cui et al., 2023) studied in this work. Next, we elaborate on the work related to LQA.

Datasets: Several datasets have been constructed and released for LQA. Table 1 shows comparisons of these datasets. Wyner et al. (2016) first presented a corpus in the form of textual entailment from the question to an answer, which was derived from a USA national bar exam. COLIEE (Kano et al., 2019) and CJRC (Duan et al., 2019) focused on answering questions with yes/no or short answers. Zhong et al. (2020c) innovatively proposed a reading dataset for the law exam, which was effective for multi-choice. These datasets are derived from judicial examinations, which are far from the real-world LQA scenario. Recently, Mansouri and Campos (2023) paid attention to questions that have multiple answers, selecting answer from all candidate answers. Chen et al. (2023) collected answers from the legal forum and constructed an answer database to retrieve the optimal answer. Louis et al. (2024) collected 1.8k QA pairs from a legal forum, which only covered civil legal questions.

Methods: Based on the above datasets, researchers also designed methods to predict/generate answers. Early methods mainly relied on human-defined rules (Buscaldi et al., 2010; Kim and Goebel, 2017). Later, researchers extracted key concepts and events from law texts to improve models' effectiveness (Wyner et al., 2016). Recently, inspired by the QA pipeline of retriever and reader, Kien et al. (2020) proposed a search-based approach to find the most relevant legal articles on legal questions to support the answer generation; Zhong et al. (2020c) proposed a reading comprehension inference method for different types of questions in the bar examination. These works did not take into account generative LQA scenarios. Recently, a number of approaches have provided new ideas for LQA by fine-tuning LLMs. Cui et al. (2023) fine-tuned LLaMA(Touvron et al., 2023) with legal knowledge and first proposed an LLM for legal advice. LaWGPT[2], HanFei[3] were also published for LQA with similar fine-tuning processes.

Existing works, both in terms of data construction and methodology, often overlook users' practical needs and fail to generate relevant and effective answers based on multiple circumstances. In this work, we focus on user-centered LQA, collecting relevant data and proposing a corresponding solution to address these issues.

2.2 Community QA Datasets

Community-based QA datasets have been extensively explored, which play an indispensable role in promoting QA methods in specific fields. Le et al. (2016) focused on the field of education and collected questions from educational websites in the United States and Poland to better help students' learning. Basaldella et al. (2020) collected datasets from the medical field on social media. Maia and Endres (2024) constructed community QA datasets from the Home Improvement, Personal Finance, and Money sections of StackExchange[4]. For the first time, we focus on community QA in the legal domain.

2.3 Open-Domain QA

Green Jr et al. (1961) exploded open QA for the first time, which answered questions with structured knowledge bases. Nowadays, a typical QA follows two steps: The retriever first finds out the relevant paragraphs as context, and the reader gets the answer according to the question and the context (Zhong et al., 2020c). We divide current

[2] https://github.com/pengxiao-song/LaWGPT

[3] https://github.com/siat-nlp/HanFei

[4] https://stackexchange.com

approaches into two categories according to the reader. The extraction reader answers questions by selecting a span in context (Fader et al., 2014; Seonwoo et al., 2020). While the generation reader is more similar to the reading comprehension (Zhong et al., 2020b), where the model is guided to learn associations between the question and the context, and then generate reasonable answers (Iida et al., 2019; Gao et al., 2021). In this work, we follow the core idea of the latter, focusing on LQA and aiming to capture the multiple circumstances in questions and generate answers accordingly.

3 Approach

3.1 Problem Definition

Let $q = (s_1, s_2, ...s_n)$ be a question with n sentences and y be its corresponding answer. Additionally, we have a legal knowledge database denoted as $\mathcal{D} = \{l_1, l_2, \cdots, l_{|\mathcal{D}|}\}$, where the item l represents a specific law article or regulation. Our goal is to develop a model that can effectively utilize the knowledge in $\mathcal{D}$ and generate a high-quality answer for a given question q. In this paper, we use bold face lower/upper case letters to denote vectors/matrices respectively.

3.2 Overview of H-LegalKI Framework

As shown in Figure 2, the proposed H-LegalKI framework primarily consists of a legal knowledge retriever and an answer generation model that integrates the retrieved knowledge. At the initial stage, we split the question into individual sentences and evaluate the relevance between a question/sentence and each item in $\mathcal{D}$. Items with high relevance are retained as legal knowledge to support the answer generation. Next, we reformulate the texts of question and legal knowledge by inserting special tokens to facilitate the learning of question- and sentence-level information. After encoding, we employ multiple Transformer (Vaswani et al., 2017) layers as the fusion module to fuse the information of question/sentence to that of legal knowledge. We further adopt an averaging strategy to remove the redundant information from the legal knowledge representation. Finally, an autoregressive decoder is employed to generate the answer.

3.3 Legal Knowledge Retriever

In legal communities (forums), users tend to express their problems clearly at once to avoid delays. We also empirically find that these questions often present detailed facts, which are associated with multiple legal articles/regulations (Liu et al., 2023). To obtain comprehensive legal knowledge support, we propose to retrieve legal knowledge (items) at both question and sentence levels.

Specifically, we first retrieve relevant legal items from the knowledge database $\mathcal{D}$ based on the entire question. Additionally, we split the question into sentences and retrieve legal items based on each sentence. We employ Bertscore (Zhang et al.) to evaluate the relevance between the question/sentence and the legal item. Compared with other methods that encode text as a single vector representation such as TF-IDF (Salton et al., 1975) and SBert (Reimers and Gurevych, 2019), Bertscore can retain the maximum amount of information of text and make a more comprehensive similarity evaluation, thus improving the accuracy of knowledge retriever. We have also made some improvements to solve the issue of high computation costs of Bertscore (described in Appendix A.1).

Formally, the process can be expressed as a function $Sim : (query, \mathcal{D}) \rightarrow \mathcal{B}$, where the $query$ can be a question or a sentence, and the $\mathcal{B}$ is the similarity scores of all legal items, ordered by decreasing. For a question, we select top k relevant legal items $L^q = \{l_1^q, l_2^q, \cdots, l_k^q\}$, which serves as the question-level legal knowledge. Each individual sentence in the question provides a specific circumstance and we select top k legal items for each sentence, denoted as $L^s = \{l_i^{s_j}\}_{i=1,j=1}^{i=k,j=n}$, where $l_i^{s_j}$ is the i-th legal item for the j-th sentence. Finally we obtain $k * (n+1)$ legal items $L = \{L^q, L^s\}$. Considering that there may be duplicate legal items in L, we define a set I to record the ids of duplicate items.

3.4 Answer Generator

3.4.1 Encoder

To learn multi-level information of a question, we respectively insert special tokens [QUE] and [SEN] at the start of the question and each sentence inspired by (Lee et al., 2020; Zhu et al., 2023). We feed the following data into the question encoder:

$$\hat{q} = \text{[QUE] [SEN] } s_1 \text{ [SEN] } s_2 \; \cdots \; \text{[SEN] } s_n.$$

After encoding, we extract the embedding of [QUE] token as the representation of the whole question, and the embeddings of [SEN] tokens as the representations of sentences.

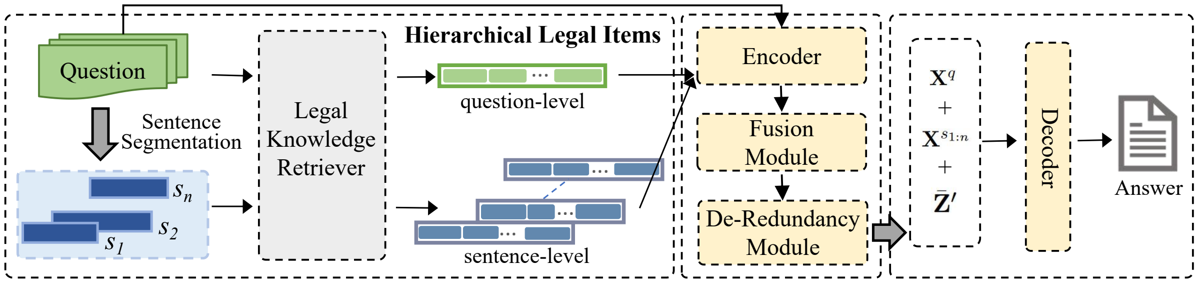


Figure 2: The framework of H-LegalKI.

For the legal items, we concatenate the top k items (corresponding to the question q or sentence s_j) sequentially and insert a special token [LAW] at the beginning of each item, which is defined as (take the legal items corresponding to q as an example):

$$l^q = \text{[LAW]}\ l_1^q\ \text{[LAW]}\ l_2^q\ \cdots\ \text{[LAW]}\ l_k^q.$$

Similarly, we define the legal items for j-th sentence as $l^{s_j} = \text{[LAW]}\ l_1^{s_j}\ \text{[LAW]}\ l_2^{s_j}\ \cdots\ \text{[LAW]}\ l_k^{s_j}$. We use the same structure as the question encoder but with different parameters to encode legal items.

3.4.2 Fusion Module

After extracting the question and sentence representations from the outputs of the question encoder, we fuse them to the representations of corresponding legal items to learn question- and sentence-specific legal representations. This process will yield both question- and sentence-level representations of legal knowledge.

Take the fusion of question and legal items as an example. We reformulate the embedding of [QUE] as $\mathbf{X}^q \in \mathbb{R}^{1\times1\times d}$, where d is the number of dimensionality. The output of the legal encoder for the corresponding legal items is denoted as $\mathbf{Z}^q \in \mathbb{R}^{1\times m_q\times d}$, where m_q is the length of l^q. We then concatenate $\mathbf{X}^q$ and $\mathbf{Z}^q$ along with the second dimension, and feed the result to multiple Transformer layers (MTL) to learn question-aware representations of legal items. Formally, the process is defined as:

$$\hat{\mathbf{Z}}^q = \text{MTL}(Q = K = V = [\mathbf{X}^q; \mathbf{Z}^q]), \quad (1)$$

where Q, K and V respectively represent the query, key, and value within the Transformer layer.

For the fusion of sentences and legal items, we construct the sentence representations as $\mathbf{X}^{s_{1:n}} \in \mathbb{R}^{n\times1\times d}$. On the other hand, we denote the representations of corresponding legal items as $\mathbf{Z}^s \in \mathbb{R}^{n\times m_s\times d}$, where $m_s = \max_{j=1\sim n}(\text{length}(l^{s_j}))$. We also concatenate $\mathbf{X}^{s_{1:n}}$ and $\mathbf{Z}^s$, and feed the result to MTL to get $\hat{\mathbf{Z}}^s$, the sentence-aware representations of legal items.

We extract the embeddings of all [LAW] tokens from $\hat{\mathbf{Z}}^q$ and $\hat{\mathbf{Z}}^s$, and construct the question- and sentence-level legal knowledge as $\tilde{\mathbf{Z}}^q \in \mathbb{R}^{1\times k\times d}$ and $\tilde{\mathbf{Z}}^s \in \mathbb{R}^{n\times k\times d}$, respectively. We further concatenate and flatten these two types of knowledge and construct the whole legal knowledge matrix as $\bar{\mathbf{Z}} \in \mathbb{R}^{(n*k+k)\times d}$.

3.4.3 De-Redundancy Module

As previously mentioned, during the knowledge retrieval phase, we may obtain duplicate legal items. Different factual details within the question might focus on slightly different parts of the same legal item. Therefore, we do not remove these duplicate items during the retrieving phase. After the above learning process, although the representations of these items may differ, they may still contain significant redundant information. This can potentially harm the model's performance (Dieckmann and Rieskamp, 2007).

Suppose these duplicate legal items can be divided into P groups. For a specific group p, there are C legal item representations $\mathbf{r}_c^p$. We average these representations as:

$$\mathbf{r}^p = \sum\nolimits_1^C \frac{1}{C}\mathbf{r}_\mathbf{c}^\mathbf{p}. \quad (2)$$

Then, we remove the representations corresponding to the $2 \sim C$ legal items in the current group from knowledge matrix $\bar{\mathbf{Z}}$, and replace the representation of the first legal item with $\mathbf{r}^p$. Finally, the de-redundant knowledge is denoted as $\bar{\mathbf{Z}}'$, and we define the encoder, fusion module, and de-redundancy module as a network F, i.e., $\bar{\mathbf{Z}}' = F(q, L)$.

Dataset		QA Pairs	QLength	ALength
English	Train.	7007	882	1495
	Dev.	890	858.6	1438.2
	Test	1002	833.7	1479.7
Chinese	Train.	17150	101.9	51.4
	Dev.	2452	101.5	99.9
	Test	2178	103	59.4

Table 2: The statistics of LegalCQA. QLength and ALength are the average length of questions and answers in LegalCQA.

3.4.4 Decoder

We employ an autoregressive decoder to generate the answer:

$$P_{AR}(y \mid \mathbf{E}) = \prod P_D(y_t \mid y_{<t}, \mathbf{E}), \qquad (3)$$

where $\mathbf{E} = [\mathbf{X}^q; \mathbf{X}^{s_{1:n}}, \bar{\mathbf{Z}}']$.

3.5 Training Objective

We employ the following loss to train the entire answer generation model:

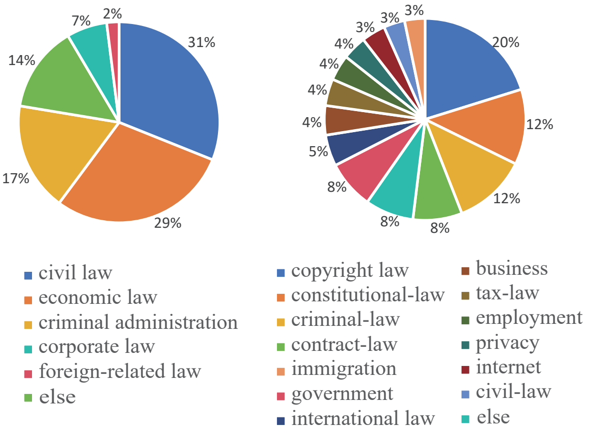
$$\mathcal{L}_{rec} = -\mathbb{E}_{y \sim P_y}[\log P_D(y \mid F(q, L))]. \qquad (4)$$

4 Experiments

4.1 Dataset Construction

Existing LQA datasets focus on answer selection or extraction, lacking generative LQA data. We construct LegalCQA based on QA pairs from legal community, which consists of LegalCQA-zh and LegalCQA-en in Chinese and English languages, respectively. Table 2 and Figure 3 show the detailed statistics of the dataset.

LegalCQA-zh: The Chinese sub-dataset is collected from 110 website[5]. We first remove duplicate questions and then further deeply explore these QA pairs to filter low-quality questions or answers. Specifically, for the QA pairs that questioners have labeled if they are satisfied with answers, we retain all of these questions and adopt the best answers chosen by questioners as the ground labels. For the remaining, we remove these QA pairs that the questions or all answers are too short to contain valuable information. For QA pairs that meet the length requirements, we choose the longest answers as the ground labels. For the legal knowledge base, we employ Chinese legal corpus released by (Zhong et al., 2020c). As illustrated in Figure 3(a), this sub-dataset comprises five main categories of legal questions (civil law, economic law, criminal administrative, corporate law, and foreign-related law), covering 70 subcategories such as resettlement, debt recovery, medical disputes, and traffic accidents.



(a) Chinese dataset (b) English dataset

Figure 3: The distribution of types of legal questions.

LegalCQA-en: We construct the English sub-dataset from the Law Stack Exchange[6]. Considering the differences in legal systems among countries, we first select the largest subset of American legal questions, amounting to a total of 8,899 pairs. We then filter questions with a vote count less than 1 and select the highest-voted answer as the best answer for them. We employ American legal corpus introduced by Chalkidis et al. (2023) as the legal knowledge base. This sub-dataset consists of a diverse range of question types. We have shown the detailed types of questions in Figure 3 (b).

We further divide the two sub-datasets randomly into training, validation, and test sets with a ratio of 8:1:1.

4.2 Baselines

We compare H-LegalKI with the following baselines:

- **BART** (Lewis et al., 2020) is a versatile sequence-to-sequence model, excelling in text generation and understanding tasks.

- **RFID** (Wang et al., 2023) employs encoders to differentiate between causal and spurious features from the context, producing more informed answers.

[5] https://www.110.com/

[6] https://law.stackexchange.com

- **EAR-RI** (Chuang et al., 2023) enhances the connection between query expansion and the retriever, furthermore enhancing question answering quality.

- **WisdomInterrogatory**[7] covers a wider range of legal tasks in both English and Chinese training on huge legal datasets.

- **Fuzi-Mingcha**[8] is constructed based on massive Chinese unsupervised judicial corpus and supervised judicial fine-tuning data.

- **DISC-LawLLM** (Yue et al., 2023) uses large language model with legal knowledge to provide a wide range of legal services, especially in legalQA.

- **Chatlaw** (Cui et al., 2023) fine-tunes LLaMA with specific legal tasks in Chinese.

These baselines can be categorized into two groups. The first group consists of models designed for open domain QA: BART, RFID, and EAR-RI. And the other group includes some legal LLMs. For baselines RFID and EAR-RI which design special modules to integrate questions and extra knowledge like H-LegalKI, we add legal items to the model's inputs. For other baselines, we exploit user questions as inputs. We fine-tune three baselines: BART, RFID, and EAR-RI, which have a similar model scale to our method. On the other hand, we do not fine-tune these large language models with more than 7B parameters for the considering of computation cost.

4.3 Evaluation Metrics

Automatic Evaluation: Following previous studies of (Louis et al., 2024; Dai et al., 2023; Wang et al., 2023), we employ two metrics: BLEU and METEOR. BLEU is used to evaluate the quality of generated answers based on n-gram overlaps between generated answers and reference answers. We also consider a more advanced automated evaluation metric–METEOR (Banerjee and Lavie, 2005). METEOR is able to simultaneously measure accuracy, recall, and fluency, which is more flexible for evaluating word matching as well as word order and more similar to human evaluations.

Human Evaluation: We randomly select 100 samples from the test set. Three Chinese native law students are assigned the task of scoring answers on Professionalism (Pro), Completeness (Com), and Relevance (Rel), using a scale of 1 (very bad) to 3 (very good). We report the average scores across the three annotators as final results. The details of these metrics are described as follows:

- Professionalism (Pro) evaluates whether the answer shows a high degree of legal professionalism and has great reference value.

- Completeness (Com) measures if the answer covers all the user's concerns, and whether it puts forward corresponding practical suggestions.

- Relevance (Rel) measures whether the answer is closely related to the question, and whether it contains irrelevant content.

4.4 Implementation Details

We employ BART$_{\text{BASE}}$ as the backbone model in our method. We use 3 Transformer layers for fusion module. The number of retrieved legal items k is set to 3. For training, we employ Adam optimizer and set the learning rate to 5×10^{-6}. The experiments are conducted on one RTX 3090 GPU (24G) and the training approximately takes 7-9 hours.

4.5 Results

Table 3 shows the experimental results using automatic evaluation, we can obtain the following observations: (1) Except for Bart, other baselines have learned legal knowledge, so the performances of these models have been improved in all metrics, which proves that external knowledge is necessary for the field of LQA; (2) The current LLMs do not show satisfactory performances in all metrics. This may be due to that they fail to effectively capture the multiple circumstances in questions; (3) The proposed method shows a large advantage in all metrics. This is attributed to the fact that, using a hierarchical way, H-LegalKI retrieve comprehensive legal knowledge and effectively use different levels of factual details and legal knowledge. Thereby H-LegalKI can meet users' practical needs like an legal professional; (4) We use t-test with significance level 0.05 to test the significance of performance difference. Results show that our method significantly outperforms all the baselines.

As for human evaluations, we conduct the experiment on LegalCQA-zh for the proposed H-LegalKI and three representative baselines (the

[7] https://github.com/zhihaiLLM

[8] https://github.com/irlab-sdu/fuzi.mingcha

Language (→)	English			Chinese		
Model (↓)	BLEU$_1$	BLEU$_2$	METEOR	BLEU$_1$	BLEU$_2$	METEOR
BART (0.22B)	1.1	0.16	2.21	0.05	0	0.08
RFID (0.58B)	4.15	1.22	6.8	-	-	-
EAR-RI (0.58B)	6.39	1.94	8.67	-	-	-
WisdomInterrogatory (7B)	4.51	1.55	4.96	4.81	1.01	9.99
Fuzi-Mingcha (7B)	-	-	-	5.06	1.05	9.88
DISC-LawLLM (13B)	8.3	2.03	9.18	5.2	1.09	10.06
Chatlaw (13B)	-	-	-	4.75	0.9	9.51
H-LegalKI (0.22B)	**21.93**	**8.63**	**15.28**	**14.79**	**5.95**	**12.28**

Table 3: Evaluation results. “-” indicates that the model does not support the language. Higher values are considered desirable for all metrics.

Question	借款他人 15 万到期后一直没还，后出具还款计划按月偿还 5000 元，2 年内还清，但是月偿还 5000 元也一直没履行，可以起诉要求他全额一次性偿还全部本息吗？具体怎么做？ The borrower has not returned after the maturity of 150,000 yuan. After issuing a repayment plan to repay 5,000 yuan per month, pay off within 2 years, but the monthly repayment of 5,000 yuan has not been fulfilled, can we sue him to repay all the principal and interest in full? How exactly?
DISC-LawLLM	根据《中华人民共和国合同法》的规定，借款人应当按照约定的期限和方式归还借款。如果借款人未按照约定的方式和期限归还借款，出借人有权要求其归还全部借款本金和利息。 According to the Contract Law of the People's Republic of China, the borrower shall repay the loan within the agreed time limit and manner. If the borrower fails to repay the loan in accordance with the agreed manner and time limit, the lender shall have the right to require the borrower to repay all the principal and interest of the loan.
Ours	可以起诉要求偿还，协商不成，可以诉讼解决，要求对方支付余款及利息及违约金，如果对方拒绝支付的，你可以向法院起诉，由法院裁定好相关证据材料，如欠条、欠款协议、收条、还款计划、汇款单、转账单、银行流水记录、个人所得税、利息等可以按照借款合同约定向出借人支付。如果造成合同无法继续履行或存在违约的情况，可向法院起诉，由法院依法判决分割财产及其他可以归还的财产以及损失。 You can sue for repayment, negotiation fails, you can sue to settle, ask the other party to pay the balance and interest and liquidated damages, if the other party refuses to pay, you can sue the court, and the court will decide the relevant evidence materials. Such as IOU, arrears agreement, receipt, repayment plan, money order, transfer statement, bank records, personal income tax, interest, etc. can be paid to the lender in accordance with the loan contract. If the contract cannot be continued to perform or there is a breach of contract, it can be sued to the court, and the court will judge the division of property and other property that can be returned as well as losses.

Table 4: Generated answers using the best baseline DISC-LawLLM and proposed H-LegalKI. Different highlighted parts indicate different circumstances and corresponding answers.

Methods	Pro	Com	Rel
BART	1.24	1.07	1.68
WisdomInterrogatory	2.11	2.33	2.13
DISC-LawLLM	**2.23**	2.35	2.56
H-LegalKI	2.21	**2.54**	**2.59**

Table 5: Human evaluation results.

most relevant model BART, and the two best-performing baselines WisdomInterrogatory and DISC-LawLLM). We report the average scores in Table 5. As can be seen, our approach achieves the best performance in terms of Completeness and Relevance, demonstrating the ability to address users' circumstances. In terms of Professionalism, our approach is slightly inferior to DISC-LawLLM, which may be due to the fact that DISC-LawLLM is fine-tuned in a large corpus of professional legal knowledge.

4.6 Analysis

Case Study: Figure 4 shows the generated answers from our method and the strongest baseline, DISC-LawLLM. The question mentions a debt dispute and contract, with the user seeking guidance on how to resolve the issue. Both approaches provide correct and relevant answers. In terms of fluency, DISC-LawLLM performs better, while our method introduces some minor repetition and stuttering. However, there are notable differences in content. DISC-LawLLM offers a response grounded in official law texts but lacks practical advice. In contrast, our method addresses the multiple circumstances and actual needs of users by analyzing the problem and providing actionable solutions, including spe-

Methods	BLEU$_1$	BLEU$_2$	METEOR
H-LegalKI	**21.93**	**8.63**	**15.28**
- Question Fusion	5.87	2.45	12.55
- Sentence Fusion	5.5	2.31	12.22
- De-Redundancy	6.26	2.17	8.53

Table 6: Ablation study results.

cific materials to prepare, such as arrears agreement and receipt. More cases can be found in Appendix A.2.

Ablation Study: We conduct ablation experiments on the LegalCQA-en sub-dataset to validate the effectiveness of the core mechanisms in our proposed method. The results are shown in Table 6, where "-" indicates the removal of a specific mechanism. Removing these mechanisms led to a significant drop in model performance, demonstrating the critical roles of hierarchical knowledge integration and de-redundancy module.

5 Conclusion

In this work, we propose a Hierarchical Legal Knowledge Integration (H-LegalKI) framework to enhance generative LQA. The proposed framework effectively learns multiple factual details in user-expressed questions and integrates comprehensive legal knowledge at both the question and sentence levels. We collect QA pairs from legal forums and construct the first public legal community QA dataset called LegalCQA. We conduct plentiful experiments on LegalCQA, and the results confirm the effectiveness of H-LegalKI compared to competitive baselines.

6 Limitation

We have not yet evaluated the performance of H-LegalKI with larger pre-trained language models since they exceed the capacity of our available GPU resources.

Another point is that we use Bertscore to retrieve the external legal knowledge, which takes more time than BM25 and TF-IDF which only consider word matching. According to our manual observation of 100 random samples, we found that Bertscore can find more relevant legal items. Moreover, we have improved the Bertscore to make it applicable to our framework and the time has been greatly shortened.

7 Ethics Statement

We collect the dataset from public legal forums, all of the questions are publicly available and anonymously posted by users. After checking, there was no specific personal information, and all of them were replaced by words. However, text generation is likely to be used for malicious purposes, such as to create false information. We should carefully consider and study this in the future.

We hired three Chinese native speakers who studied in the legal domain as annotators to manually evaluate the performance of the proposed method and baselines. Considering the wage standards of China, annotators will get 2.0 yuan (RMB) for each sample.

8 Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC), under Grant Nos. 62133012, 61936006, 62472340, 62103314, 62425605 and 62303366.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi Nojehdeh, and Nigel Collier. 2020. Cometa: a corpus for medical entity linking in the social media. In *Empirical Methods in Natural Language Processing 2020*, pages 3122–3137. Association for Computational Linguistics (ACL).

Davide Buscaldi, Paolo Rosso, José Manuel Gómez-Soriano, and Emilio Sanchis. 2010. Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, 34:113–134.

Ilias Chalkidis, Nicolas Garneau, Cătălina Goanță, Daniel Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535.

Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. Equals: A real-world dataset for legal question answering via reading chinese laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 71–80.

Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih, and James Glass. 2023. Expand, rerank, and retrieve: Query reranking for open-domain question answering. *arXiv preprint arXiv:2305.17080*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Yi Dai, Hao Lang, Yinhe Zheng, Fei Huang, and Yongbin Li. 2023. Long-tailed question answering in an open world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6362–6382.

Anja Dieckmann and Jörg Rieskamp. 2007. The influence of information redundancy on probabilistic inferences. *Memory & cognition*, 35:1801–1813.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 439–451. Springer.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1156–1165.

Shen Gao, Xiuying Chen, Zhaochun Ren, Dongyan Zhao, and Rui Yan. 2021. Meaningful answer generation of e-commerce question-answering. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–26.

Bert F Green Jr, Alice K Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224.

Ryu Iida, Canasai Kruengkrai, Ryo Ishida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Exploiting background knowledge in compact answer generation for why-questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 142–151.

Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2019. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2018 Workshops, JURISIN, AI-Biz, SKL, LENLS, IDAA, Yokohama, Japan, November 12–14, 2018, Revised Selected Papers*, pages 177–192. Springer.

Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. 2020. Answering legal questions by learning neural attentive text representation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 988–998.

Mi-Young Kim and Randy Goebel. 2017. Two-step cascaded textual entailment for legal bar exam question answering. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 283–290.

Long T Le, Chirag Shah, and Erik Choi. 2016. Evaluating the quality of educational answers in community question-answering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 129–138.

Haejun Lee, Drew A. Hudson, Kangwook Lee, and Christopher D. Manning. 2020. SLM: Learning a discourse language representation with sentence unshuffling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1551–1562, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. Sailer: Structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1035–1044, New York, NY, USA. Association for Computing Machinery.

Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. Mlljp: Multi-law aware legal judgment prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1023–1034.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

Macedo Maia and Markus Endres. 2024. Investigating questioner's explicit information influences in transformer-based community question answering. In *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pages 93–100. IEEE.

Behrooz Mansouri and Ricardo Campos. 2023. Falqu: Finding answers to legal questions. *arXiv preprint arXiv:2304.05611*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Yeon Seonwoo, Ji-Hoon Kim, Jung-Woo Ha, and Alice Oh. 2020. Context-aware answer extraction in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2418–2428.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3501–3507. International Joint Conferences on Artificial Intelligence Organization. Main track.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Cunxiang Wang, Haofei Yu, and Yue Zhang. 2023. Rfid: Towards rational fusion-in-decoder for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2473–2481.

Adam Zachary Wyner, Biralatei James Fawei, and Jeff Z Pan. 2016. Passing a usa national bar exam: a first corpus for experimentation. In *LREC 2016, Tenth International Conference on Language Resources and Evaluation*. LREC.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3086–3095, Online. Association for Computational Linguistics.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jie Zhao, Ziyu Guan, Cai Xu, Wei Zhao, and Enze Chen. 2022. Charge prediction by constitutive elements matching of crimes. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4517–4523. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jie Zhao, Ziyu Guan, Wei Zhao, and Yue Jiang. 2024. Enhancing criminal case matching through diverse legal factors. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2379–2383, New York, NY, USA. Association for Computing Machinery.

Jie Zhao, Ziyu Guan, Wei Zhao, Yue Jiang, and Xiaofei He. 2023. Few-shot domain adaptation for charge prediction on unprofessional descriptions. *Preprint*, arXiv:2309.17313.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Botao Zhong, Wanlei He, Ziwei Huang, Peter ED Love, Junqing Tang, and Hanbin Luo. 2020a. A building regulation question answering system: A deep learning methodology. *Advanced Engineering Informatics*, 46:101195.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020c. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

Xuekai Zhu, Jian Guan, Minlie Huang, and Juan Liu. 2023. StoryTrans: Non-parallel story author-style transfer with discourse representations and content enhancing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14803–14819, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Efficiency Improvement of Bertscore

We have made some efforts to improve the efficiency since simply employing the original BERTscore consumes huge computing resources. Specifically, we first encode all questions/sentences and legal items, and store token representations. Then we take out corresponding representations for a specific relevance evaluation (question-legal item or sentence-legal item), and implement the importance weighting and baseline rescaling to obtain the BERTScore results. This design avoids the repeated encoding and greatly improves the efficiency.

A.2 More cases in experiment of LegalCQA

Here we present more examples of answers generated by H-LegalKI, one from the Chinese sub-dataset and one from the English sub-dataset.

- Question: 同事介绍一挂靠驾校，考了科目一。驾校拆迁。至今两年教练以各种理由推脱学习其它科目。只知道教练姓和手机号。能起诉退还学费吗？ (A colleague introduced me to a registered driving school where I passed the first exam. However, the driving school was demolished soon after, and for the past two years, the coach has been avoiding arranging further lessons for me under various excuses. I only know the coach's surname and mobile phone number. Is it possible to sue for a refund of tuition fees?)

- Generated Answer: 可以向监察大队投诉或者直接向法院起诉维权的。必经程序：劳动保障行政部门批准后，可以将其与原劳动合同关系的内容、资料等保存下来，作为证据，自行保存或提供原件供查验，争取最大限度的利益。(You can either lodge a complaint or sue in court for rights protection. Necessary procedure: After obtaining approval from the administrative department of labor security, you can preserve the relevant content or materials from the original labor contract, serving as evidence. These can be kept personally or provided in original form for verification, in order to maximize your benefits.)

- Question: Is it an act of discrimination when the employer rejects your job application, given that you a) belong to a protected group and b) can demonstrate experience in all the essential job functions (as described in the job description) using the previous projects you have done, as much as it's possible objectively? For instance, the job description says 'must have experience working with X, Y and Z' and you can prove that you have a lot of experience working with X, Y and Z. If not, why?

- Generated Answer: In the US, there are federal law prohibiting discrimination on the basis of race, color, religion, sex, national origin, and national origin, particularly in the field of employment. There is a federal law against discrimination based on sexual orientation, and there are state laws against discrimination as well.