STARD: A Statute Retrieval Dataset for Layperson Queries

Anonymous ACL submission

Abstract

Statute retrieval aims to find relevant statutory articles for specific queries. This process is the basis of a wide range of legal applications such as legal advice, automated judicial decisions, and logical legal analysis. Existing statute retrieval benchmark emphasize formal legal queries from sources like bar exams and Supreme Court cases. This neglects layperson queries, which often lack precise legal terminology and ambiguously reference legal concepts. In this study, we introduce the STAtute Retrieval Dataset (STARD), an dataset derived from real-world legal consultation questions made by the general public. Unlike existing statute retrieval datasets that focus predominantly on professional legal queries, STARD captures the complexity and diversity of layperson queries. Through a comprehensive evaluation of various retrieval baselines, including conventional methods and those employing advanced techniques such as GPT-4, we reveal that existing retrieval approaches all fall short of achieving optimal results. Additionally, we show that employing STARD as a Retrieval-Augmented Generation (RAG) dataset markedly improves LLM's performance on legal tasks, which indicates that STARD is a pivotal resource for developing more accessible and effective legal systems.

1 Introduction

Statute is a written law formally created and approved by a legislative body, such as a parliament or congress. It sets out specific rules and guidelines that need to be followed within a certain area or jurisdiction. In civil law systems, which prioritize written laws, statutes are especially important and are considered to be the main source of legal authority, which is different from common law systems where past court decisions also play an important role. The focus on statutes helps to ensure that the legal rules are clear and predictable, which is essential for maintaining the order of the society and defining legal rights and responsibilities.

Statute retrieval involves finding relevant statutory articles or sections of laws for a specific query. This process is vital in the legal field and supports a wide range of applications including legal advice services, automated judicial decision-making systems, and logical legal analysis. However, this task is challenging for the following reasons:
(1) Statutes often use complex terminology and unique linguistic structures rarely found in general texts. As a result, traditional search models may struggle to accurately capture the exact meaning of these specialized legal terms due to their lack of knowledge on the legal domain.
(2) Identifying the appropriate statutory articles for a query involves intricate legal reasoning. This requires a thorough understanding of legal principles and their practical applications.
(3) The criteria for assessing information relevance in the legal domain differ greatly from those used in general search tasks. General search tasks mainly focus on textual similarity, while legal tasks require evaluating legal relevance and understanding the connections between different legal elements.

Due to the challenging nature of statute retrieval and its paramount importance in civil law systems, significant progress has been made in this field. For example, the annually COLIEE competitions introduce a series of statute retrieval datasets using the questions extracted from the Japanese legal bar exams (Goebel et al., 2023; Kim et al., 2022; Rabelo et al., 2022). The task of these competitions aims to retrieve relevant statute law from the Japanese Civil Code Article according to questions from bar exams. AILA (Bhattacharya et al., 2019) competitions also introduce a series of statute retrieval datasets. The queries from AILA are case documents that were judged in the Supreme Court of India. The candidate statutes are part of the set of statutes from Indian law.

Despite these advancements, a significant gap persists in addressing queries from laypeople, who represent a large portion of the users of legal advice services. The current statute retrieval benchmarks are primarily based on queries from formal legal documents, such as bar exam questions or Supreme Court case files, which differ significantly from the everyday language used by the general public. However, layperson queries often lack precise legal terminology and may include ambiguous references to legal concepts, complicating the task of accurately retrieving relevant statutes.

Thus, to address the limitations of existing benchmarks, we propose STAtute Retrieval Dataset (STARD), i.e., STARD, a statute retrieval dataset based on real legal consultation questions from the general public. The STARD dataset comprises 1,543 query cases collected from genuine legal consultations and 55,348 candidate statutory articles extracted from all official Chinese legal regulations and judicial interpretations. To the best of our knowledge, STARD is the first statute retrieval dataset where queries are from real-world legal consulting proposed by the general public.

We conduct experiments on a wide range of information retrieval (IR) baselines on the STARD dataset, including traditional lexical matching models, general domain neural retrieval models, legal domain neural retrieval models, and dense retriever distilled from GPT4. The experimental results show that all existing baselines fall short of accurately and comprehensively retrieving all the relevant statutes, suggesting significant room for future work. Additionally, we show that employing STARD as a Retrieval-Augmented Generation (RAG) dataset markedly improves Large Language Model's performance on legal tasks. Indicating STARD is a pivotal resource for developing more accessible and efficient legal systems, bridging the gap between advanced computational legal research and the everyday legal needs of individuals.

In conclusion, the contribution of this paper are as follows:

- We propose STARD, a statute retrieval dataset derived from real-world legal consultation questions posed by the general public, with 1.6K queries and their corresponding relevant statutes. All the codes and datasets are available at: https://anonymous.4open.science/r/STARD/.
- We propose a comprehensive framework of relevance judgment criteria specifically designed for the statute retrieval task, which provide references and insights for the annotation of retrieval tasks in the legal field.
- We conduct experiments on a wide range of retrieval baselines, showing that STARD tasks cannot be easily solved and deserves future study.
- We present experiments on Large Language Models (LLMs) solving legal tasks with and without the STARD dataset. Experiments show that STARD can notably enhance the performance of LLMs in legal tasks.

2 Related Work

2.1 Legal Evidence Retrieval

LER (Yao et al., 2023) has defined the legal evidence retrieval task, which aims to automatically retrieve the relevant evidence given a fact description within a case. The fact is the concise description of what happened in the case, formally written by the prosecutor, the evidence is the verbose record of oral statements by case participants. However, legal evidence retrieval is merely conducted from a single legal document, without conducting external knowledge retrieval for each query. As a result, it cannot serve as an external knowledge input to enhance the model's understanding of legal knowledge.

2.2 Similar Case Retrieval

Given a case, the goal of the Similar Case Retrieval (SCR) task is to retrieve similar cases from the candidate pool according to the judgment criteria. Existing works (Ma et al., 2021; Shao et al., 2023; Xiao et al., 2019) have proposed the SCR task and use Chinese court judgments to construct datasets. Subsequent works (Li et al., 2023c,b) have defined the relevance judgment of SCR more scientifically from a legal perspective. All the works target the construction of datasets based on court judgments under specific departmental laws, making it impossible to carry out SCR under the entire Chinese legal system. The COLIEE dataset (Rabelo et al., 2021) comes from Canadian court judgment documents. Due to its tradition of following precedents, the common law system has annotations of similar cases in judgment documents. When applying the SCR dataset to the process of injecting legal knowledge into LLMs, the extensive length of legal documents poses a challenge. It becomes arduous to condense the entire document into the prompt due

to input length restrictions during retrieval. Consequently, when dealing with the complexities of Retrieval-Augmented Generation (RAG) LLMs in the legal domain, the necessity of incorporating refined external legal knowledge becomes evident.

2.3 Statute Retrieval

Cail2018 (Xiao et al., 2018; Zhong et al., 2018) competitions conduct law statute retrieval work using court judgments in Chinese criminal law. The queries in the dataset originate from the 'court's findings' part of the judgments, and the candidates are statutes of Chinese Criminal Law. The ultimate goal of this dataset is to predict criminal charges through law statute retrieval, hence the incompleteness of the law statute retrieval in the dataset. For instance, some criminal cases come with civil litigation, but civil law statutes are not included in this dataset. The annually COLIEE competitions introduce a series of statute retrieval datasets using the questions extracted from the Japanese legal bar exams (Goebel et al., 2023; Kim et al., 2022; Rabelo et al., 2022). AILA (Bhattacharya et al., 2019) competitions also introduce a series of statute retrieval datasets. The queries from AILA are court judgments in the Supreme Court of India. The candidate statutes are part of the set of statutes from Indian law. All the previous works have used legal language to describe their queries, but we argue that in real-world legal consultation scenarios, queries from laypersons are more common. Moreover, the task of retrieving law statutes with everyday problems that do not include legal descriptions is more challenging. The STARD dataset we propose can promote law statute retrieval tasks and make significant contributions to the application of LLMs in legal aid.

3 Problem Formulation

3.1 Statute of Civil Law System

Civil law is a legal system that is primarily based on codified laws rather than case precedents, making written statutes the main source of legal authority. This contrasts with common law systems where judicial decisions also play a central role. In civil law, statutes are created and enacted by legislative bodies, such as parliaments, and are organized into systematic collections known as codes, which cover various areas of law like contracts, torts, and property. A statute is a formal written law that provides specific rules and guidelines to be followed within a jurisdiction. Within statutes, there are sections known as statutory articles, which detail individual provisions or clauses of the law, addressing particular aspects or requirements. These statutes and their articles are fundamental in civil law systems for ensuring that the legal framework is clear, predictable, and accessible, thereby facilitating order and defining rights and responsibilities within the society.

3.2 Definition of Statute Retrieval

The statute retrieval task aims to accurately retrieving relevant statutory articles in response to a query. To be specific, given a query q that describes a legal issue or situation, and a corpus of statutory articles $S = \{s_1, s_2, ..., s_n\}$, $n \in N^+$. For each statute s_i in the corpus, there is a Bernoulli variable r_i indicating whether s_i is relevant[1] to the query q. The goal of the statute retrieval task is to retrieve a set of statutes $R = \{s_j | r_j = 1\}$, which includes all statutes that are relevant to the query.

4 Annotation Framework

This section explains how annotators transform general life-related questions into specific legal questions and identify the most relevant legal statutes to support these questions. To be specific, annotators use a three-step method: recall, query decomposition, and filtering (illustrated in Figure 1). This method mirrors the structured approach commonly used in legal reasoning, which involves three logical steps: establishing a broad legal principle (major premise), applying it to the specific facts of a case (minor premise), and then reaching a conclusion. This section is organized by three subsections, each detailing a part of the annotation process that is designed to mirror these logical steps in legal reasoning.

4.1 Step 1: Recall

When annotating the legal statutes supporting the query, the annotators must first narrow down the scope of the legal statutes, that is, recall the relevant chapters of the departmental laws most related to the issue from the entire legal system. In this mode of major premise retrieval, the annotators follow the principle of going from macro departmental laws to micro behavioral types. The macro departmental laws include civil and commercial

[1]The definition of "relevant" is discussed in detail in Section 4.

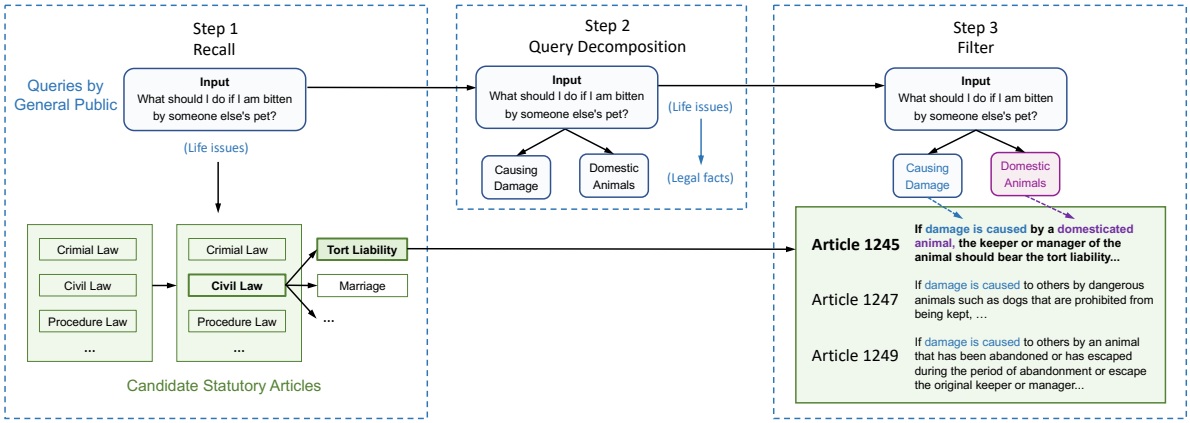

Figure 1: An illustration of our proposed three-step annotation framework.

law, criminal law, administrative law, etc. When encountering a problem, the annotators will first find out which field of departmental law it belongs to, and then gradually refine the problem into micro law. For example, if the problem is a civil law issue, the annotator judges whether it is a contract issue or a tort issue. Furthermore, if it is a contract issue, the annotator judges what type of contract it belongs to. Similarly, if it is a tort issue, the annotator judges what kind of specific tort it belongs to. After the step 1, the annotators narrow the scope of legal statute retrieval to specific chapters under the departmental law.

4.2 Step 2: Query Decomposition

Since the threshold of legal knowledge is high for those who have not received legal professional education, the legal questions raised by ordinary people are usually life issues, rather than legal issues. Life issues are simple semantic expressions that cannot directly form a meaning in legal norms. For example, when the questioner asks "What should I do if I am bitten by someone's pet?", "Pet bite" is a typical life fact. If you search for legal norms related to pets based on this, the major norms you find may not be accurate. Therefore, when annotators perform legal statute retrieval, they should transform the life facts described by the questioner into legal facts through interpretation in the step 2. This is the step to find the minor premise in the legal logic syllogism. In this transformation process, the annotator evaluates the life facts according to the provisions of the law, and selects the legal norms corresponding to these life facts. For example, for the aforementioned issue of a pet biting a person, the annotators will transform "pet bites a person" into the legal fact of "causing damage to others" and "domestic animals" according to the provisions of Chapter 9 of the Tort Liability Compilation of the Civil Code.

4.3 Step 3: Filter

During the process of filtering the recalled legal statuts, the annotators need to adopt the "subsumtion" method in the legal logic syllogism, that is, to encapsulate the legal facts transformed from life issues into the smallest range of legal statutes that support the answer to the query. The legal statutes that can cover all the legal facts in the query are defined as the most relevant set of legal statutes in this dataset. For example, the set of legal statutes obtained after recall is $S = \{S1, S2, S3\}$. The legal facts obtained from the transformation of the query are $F1, F2, F3$. The set of legal statutes that each legal fact can imply is $SF1 = \{S1, S4, S5\}, SF2 = \{S1, S5\}, SF3 = \{S3, S6\}$. The most relevant legal statutess in this dataset are $SG = (SF1 \cup SF2 \cup SF3) \cap S$. In this case, $SG = \{S1, S3\}$. The golden legal statutes selected after this filtering step can cover all the legal statutes that are implied by the legal facts in the query, which further support the answer to the query.

5 Dataset Construction

5.1 Data Sources

All queries in our dataset come from real legal consultations. Specifically, our legal team sourced legal questions from the 12348 China Legal Service Website[2], followed by a manual anonymization of

[2]This is the official website of the Chinese government for providing online legal services. The link is as follows:

each question, which involved removing any potential identifiers associated with entities, corporations, or individuals.

For the candidate statutory articles, our legal team first listed all national-level laws, regulations, and judicial interpretations of China, then manually downloaded the latest versions from official government sources. These were subsequently divided into the smallest searchable units based on Articles.

5.2 Annotation Details

5.2.1 Annotate Process

Annotators are tasked with identifying relevant articles of statutes in response to actual legal queries posed by laypersons. The specifics of the annotation framework are detailed in Section 4. Additionally, annotators are instructed not to use generative models, such as ChatGPT, for assistance. The annotation process commences with the manual anonymization of each question within the STARD dataset, involving the removal of any potential identifiers associated with entities, corporations, or individuals. Subsequently, annotators are required to locate relevant statutes for each question, following the three-step principle introduced in Section 4. We encountered a few cases (less than 0.5%) involving politically sensitive issues or scenarios without applicable statutes. These were designated as "special cases" and excluded from our final dataset. Each question was independently annotated by two different annotators. Only data with concordant annotations from both were included in the analysis.

5.2.2 Recruitment and Payment

For the recruitment process, we invite participants for annotation tasks from prestigious law schools. The remuneration scheme is designed to pay participants based on the number of completed annotations, with the payment rate varying according to the complexity of the annotation task. Given the specialized nature of legal knowledge and the intricate logic involved in reasoning, we have set an average payment of approximately 10 RMB per annotation. As per our data, on average, an individual can annotate four queries per hour, translating to an average hourly wage of 40 RMB. This wage is considerably higher than the minimum hourly wage in Beijing, exceeding it by 80%.

http://www.12348.gov.cn//homepage

5.2.3 Annotation Consistency

To evaluate the reliability of agreement among human annotators, we utilized Cohen's Kappa (Cohen, 1960) $\mathcal{K}$ coefficient in a binary classification context. This analysis, conducted on a dataset comprising 1543 annotated instances, yielded a $\mathcal{K}$ value of 0.5312. This indicates moderate agreement, highlighting the effectiveness and consistency of our annotation approach.

5.3 Ethics Discussion

In developing the STARD dataset, we have carefully addressed several ethical considerations, ensuring our research adheres to high standards of integrity and respects individual privacy.

- **Privacy and Anonymity:** Given the sensitive nature of legal consultations, we have rigorously anonymized all queries in the STARD dataset. This safeguards the privacy of individuals, preventing any disclosure of personal information and maintaining the confidentiality of legal advice seekers.

- **Transparency:** To promote reproducibility and transparency, we have made the dataset, associated models, and codebases publicly available[3]. This openness allows other researchers to verify, replicate, and expand upon our work, advancing the field of legal informatics.

- **Accountability:** Recognizing the dynamic nature of legal statutes, we commit to regularly updating the STARD dataset to reflect the latest changes in law. This ensures the dataset remains accurate and reliable for ongoing research and application.

- **Accessibility:** The STARD dataset is freely available for download from the official website under the MIT license, facilitating easy access for researchers and practitioners alike. This promotes broader usage and supports innovation across various fields.

These measures highlight our commitment to ethical research, ensuring the STARD dataset not only advances statute retrieval but also respects and promotes ethical standards across all aspects of our work.

[3]https://anonymous.4open.science/r/STARD/

Table 1: Basic statistics of our proposed STARD dataset.

Statistic	# Number
Total Candidate Statutory Articles	55,348
Total Queries	1,543
Avg. Relevant Articles per Query	1.76
Avg. Query Length	27.30
Avg. Article Length	119.93

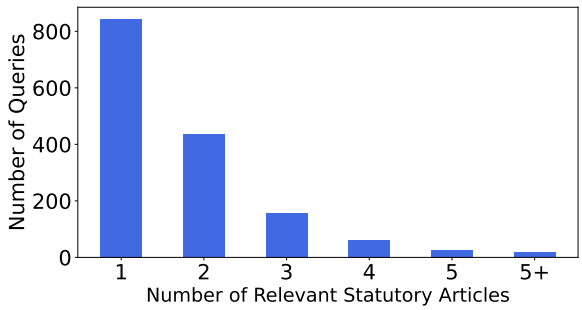

Figure 2: Distribution of relevant statutory article numbers for each query.

6 Dataset Statistics and Analysis

The basic statistics of our proposed dataset are shown in Table 1. STARD comprises a total of 1,543 queries and a large-scale corpus of 55,348 candidate statutory articles. The average query length is 27.3 words, and the average statute length is nearly 120 words.

Figure 2 presents the distribution of queries across the number of relevant statutory articles highlighting the varied complexity within the dataset. A substantial majority of the queries, 843 out of 1,543, correspond to just one relevant statutory article, indicating a significant number of queries can be addressed with a single, specific legal reference. This could suggest that many of the layperson queries are focused and pertain to specific legal issues that require straightforward statute retrieval. However, 45% of queries require multiple statutory articles which indicates some of the questions are more complex, involving multiple references of law. This diversity in query complexity demonstrates that our dataset is capable of accommodating a wide range of legal questions, from straightforward to highly intricate.

7 Statute Retrieval Experiment

7.1 Selected Retrieval Baselines

We consider four types of baselines for comparison, including traditional IR methods, pre-trained Language models on general domain data, PLMs tailored for IR, and pre-trained language models built with legal documents.

- **Traditional IR Methods**
  - **QL** (Zhai, 2008) is a language model based on Dirichlet smoothing and has good performance on retrieval tasks.
  - **BM25** (Robertson et al., 2009) is a highly effective retrieval model based on lexical matching that achieves good performance in retrieval tasks.

  For the implementation, we use the pyserini toolkit[4]. For the hyperparameter of BM25, we set $k1 = 3.8$ and $b = 0.87$[5]. Note that in our experiments, we use the scores of the BM25 and QL models to re-rank the candidate documents, rather than re-ranking the whole corpus.

- **General Domain Pre-trained Models**
  - **Chinese-RoBERTa-WWM** (Cui et al., 2021) is a RoBERTa-based model pre-trained with Whole Word Masking (WWM) strategy in Chinese corpora.
  - **SEED** (Lu et al., 2021) is a pre-trained text encoder for dense retrieval that achieves state-of-the-art performance.
  - **coCondenser** (Gao and Callan, 2021b). coCondenser is an enhanced version of Condenser(Gao and Callan, 2021a) that adds an unsupervised corpus-level contrastive loss to warm up the passage embedding space.

  For the implementation of Chinese-RoBERTa-WWM, we directly use their models released on Huggingface[6]. As SEED and Condenser have no available Chinese versions, we reproduce their work on the Chinese Wikipedia based on their open-source training code and follow all settings provided in their paper (Lu et al., 2021; Gao and Callan, 2021a).

- **Legal Domain Pre-trained Models**
  - **Lawformer** (Xiao et al., 2021) apply Longformer(Beltagy et al., 2020) to initialize and train with the MLM task on the legal domain.

[4]https://github.com/castorini/pyserini

[5]This is the best hyperparameter we got after parameter searching.

[6]https://huggingface.co/hfl/chinese-roberta-wwm-ext

– **SAILER** (Li et al., 2023a) is a structure-aware pre-trained language model for tailored legal document representation. It utilizes the logical connections between different sections within a legal document.

For the implementation of these models, we directly use the checkpoints released on the official GitHub[7].

- **Fine-tuning Retriever based on STARD**

  We initialize the model with Chinese-Roberta-WWM (Cui et al., 2021) and then employ a five-fold cross-validation technique on the STARD dataset to fine-tune the model. The dataset is randomly divided into five subsets, with one subset used as the test set and the remaining four as the training sets. The details of our fine-tuning process is introduced in Appendix C.

- **DR-GPT4** We distill a dense retrieval model from GPT-4. The details are introduced in Appendix D.

7.2 Evaluation Metrics

We use Mean Reciprocal Rank and Recall as evaluation metrics. By using both MRR and Recall, we can gain insights into both the accuracy of the top-ranked results and the comprehensiveness of the relevant statutory articles retrieved by the retrieval model. Detailed definitions of these metrics are provided in Appendix A.

7.3 Experimental Results

In this subsection, we provide a detailed analysis of the performance of various retrieval baselines evaluated on our proposed STARD dataset. The experimental results are shown in Table 2. For the fine-tuning process (detailed in Section 7.1), we initialize the model with Chinese-Roberta-WWM (Cui et al., 2021) and then employ a five-fold cross-validation technique on the STARD dataset. The dataset is randomly divided into five subsets, with one subset used as the test set and the remaining four as the training sets. Our results highlight several insights into the effectiveness of different retrieval methods. Under the zero-shot setting, traditional lexical matching techniques surpass both general and legal-domain pre-trained language models (PLMs). The performance of DR-GPT4 stands out, exceeding that of all unsupervised methods tested. Among all the approaches, fine-tuned by human annotation demonstrates the highest effectiveness. However, despite their superior performance, these models exhibit suboptimal recall rates. These findings highlight a substantial gap in existing retrieval methods for statutory tasks based on laypeople's queries, indicating the need for further exploration and additional research in this area.

[7] https://github.com/CSHaitao/SAILER/, https://github.com/thunlp/LegalPLMs

8 Retrieval Augmented Generation Experiment

8.1 Selected Benchmark

We select JEC-QA (Zhong et al., 2020), which stands as the most extensive multiple-choice dataset within the legal domain in the Chinese language. This dataset demands a high degree of reasoning ability to navigate the legal questions it contains. These questions are bifurcated into two categories: Knowledge-Driven Questions (KD-questions) and Case-Analysis Questions (CA-questions). The JEC-QA dataset encompasses a total of 26,365 questions, with 5,289 of them constituting the test set. It is crucial to highlight that the quantity of correct responses for each question within this dataset is not predetermined.

8.2 Selected LLMs

Our selected LLMs are listed as follows. The generation configuration are detailed in Appendix B.

- **Baichuan** (Yang et al., 2023) is a series of large-scale multilingual language model, trained from scratch on 2.6 trillion tokens. We choose the **Baichuan-2-Chat-13B** model which is widely used in bilingual Chinese-English scenarios.

- **ChatGLM** (Du et al., 2022) is a series of generative language models optimized for Chinese question answering and dialogue. We choose **ChatGLM3-6B** with 6.2 billion parameters.

- **ChatGPT** (Brown et al., 2020) is a series of large language models developed by OpenAI, includes several versions. Among these, we choose **GPT-3.5-turbo** which is identified as the most advanced GPT-3.5 model.

8.3 Experimental Results

In this subsection, we present the experiment of LLM's performance on legal multiple-choice

Table 2: The overall experimental results of multiple baselines on STARD. The best results are in bold and the second best results are underlined. “PLM" stands for Pre-trained Language Model, “R" stands for Recall, and “M" stands for MRR. DR-GPT4 is a dense retrieval model distilled from GPT-4. The results of General PLM and Legal PLM are all in zero-shot setting (without fine-tuning of human annotation).

		R@5	R@10	R@15	R@20	R@30	R@50	M@3	M@5	M@10
Lexical Matching	**QL**	0.3363	0.4020	0.4478	0.4651	0.4839	0.5537	0.3052	0.3167	0.3304
	BM25	0.3349	0.3943	0.4301	0.4504	0.4773	0.5240	0.3176	0.3251	0.3369
General PLM	**Roberta**	0.3216	0.3908	0.4338	0.4646	0.5042	0.5715	0.2766	0.2905	0.3010
	SEED	0.2897	0.3555	0.3997	0.4264	0.4589	0.4975	0.2607	0.2708	0.2816
	coCondenser	0.1120	0.1598	0.195	0.2223	0.2659	0.3288	0.0847	0.0922	0.1004
Legal PLM	**SAILER**	0.2330	0.3050	0.3488	0.3790	0.4286	0.4885	0.2006	0.2115	0.2234
	Lawformer	0.2411	0.2989	0.3414	0.3720	0.4137	0.4733	0.2205	0.2313	0.2412
Human Annotation	**Fine-tuned**	**0.5206**	**0.6061**	**0.6635**	**0.7064**	**0.7485**	**0.8107**	**0.4372**	**0.4543**	**0.4724**
GPT4 Annotation	**DR-GPT4**	0.4382	0.5174	0.5676	0.5961	0.6471	0.6810	0.3842	0.3948	0.4106

Table 3: The overall experimental results of three LLMs on the JecQA benchmark. We report accuracy as the evaluation metric. The best results are in bold and the second best results are underlined.

	Retriever	JecQA-CA	JecQA-KD
Baichuan-13B	**w/o RAG**	0.2307	0.2662
	BM25	0.2327	0.2878
	Fine-tuned	**0.2379**	**0.2905**
ChatGLM3-6B	**w/o RAG**	0.1852	0.1943
	BM25	0.1890	0.2235
	Fine-tuned	**0.1996**	**0.2367**
GPT-3.5-Turbo	**w/o RAG**	0.1870	0.2057
	BM25	**0.2330**	**0.2929**
	Fine-tuned	0.1926	0.2516

dataset Jec-QA. Table 3 presents the results of the LLM's performance with and without the use of Retrieval-Augmented Generation (RAG). In the scenario without RAG, the LLM directly selects the correct answer based on the questions in the JECQA. In the RAG scenario, we use the STARD corpus as the external knowledge base for RAG. For each question, the retrieval model (BM25 or Fine-tuned Dense Retriever) recalls the top 10 relevant statutory articles from the corpus based on the question. The retrieved statutory articles, along with the question, are input to the LLM, which then select the correct answer based on the relevant documents and question stem.

The experimental results reveal that using the STARD corpus as the external knowledge base for the RAG significantly enhances the performance of large language models (LLMs), underscores the value of our proposed dataset in improving the effectiveness of LLMs on legal tasks. Interestingly, the results diverge when comparing the performance of different retrieval models across specific LLM configurations. For the Baichuan and ChatGLM models, a fine-tuned dense retriever outperforms the BM25 algorithm, suggesting that these models may benefit from the high recall rate of dense retrievers. However, this advantage does not extend to the ChatGPT model, where the BM25 algorithm actually delivers superior performance compared to the fine-tuned dense retriever. This indicates that the effectiveness of retrieval models can vary significantly depending on the underlying characteristics of the LLMs.

9 Conclusion

We present STARD, a new benchmark consisting of 1,543 questions from the general public and their corresponding relevant statutes. To the best of our knowledge, STARD is the first Chinese statutes retrieval dataset tailored for the general public. The candidate corpus includes all the judicial interpretations and statutory provisions of the Chinese legal system. While not the focus of this paper, we also provide experimental results of using STARD as the RAG dataset for LLMs on multiple-choice benchmarks and question-answering benchmarks. The results demonstrate that our dataset can markedly improve LLM's performance on legal tasks.

10 License and Permissions

STARD are freely available under the MIT License. This permissive license was chosen to encourage the widespread use and adaptation of our resources, allowing for both academic and commercial applications without significant restrictions. For detailed terms and conditions, including how the dataset, code, and models can be used, modified, and shared, please refer to the documentation provided in our GitHub repository[8].

[8]https://anonymous.4open.science/r/STARD/

11 Limitations

We acknowledge the limitations of this paper. One of the primary limitations is that our dataset is specifically designed around the Chinese legal system, inherently limiting its direct applicability to legal systems outside of this context. Despite our discussions on potential methodologies for adapting STARD to other civil law systems, such an expansion necessitates creating and annotating new datasets tailored to those systems' distinct legal frameworks and statutes. Thus, our future work will be dedicated to developing additional datasets that encompass a broader range of civil law systems. This endeavor aims to extend the utility of our work and foster further research and development in the domain of legal statute retrieval, ensuring broader applicability and relevance across different legal landscapes.

12 Ethics Statement

In the framework of this research, ethical considerations have been paramount from the initial stages, underscoring our commitment to the responsible advancement and application of artificial intelligence technologies. Our adherence to the principles of open research and the critical importance of reproducibility have compelled us to make all associated models, datasets, and codebases publicly available on GitHub.

Moreover, in the development of our dataset, we have paid scrupulous attention to privacy and respect for individuals' rights. Given the inherently sensitive nature of legal consultations, we have diligently anonymized every query within the STARD dataset. This process involved the removal of any potential identifiers related to entities, corporations, or individuals, thereby safeguarding privacy and preempting the possibility of data misuse.

References

Qingyao Ai, Keping Bi, Jiafeng Guo, and W Bruce Croft. 2018. Learning a deep listwise context model for ranking refinement. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 135–144.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Fire 2019 aila track: Artificial intelligence for legal assistance. In *Proceedings of the 11th annual meeting of the forum for information retrieval evaluation*, pages 4–6.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Luyu Gao and Jamie Callan. 2021a. Condenser: a pre-training architecture for dense retrieval. *arXiv preprint arXiv:2104.08253*.

Luyu Gao and Jamie Callan. 2021b. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Rethink training of bert rerankers in multi-stage retrieval pipeline. In *European Conference on Information Retrieval*, pages 280–286. Springer.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the competition on legal information, extraction/entailment (coliee) 2023. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 472–480.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2022. Coliee 2022 summary: Methods for legal document retrieval and entailment. In *JSAI International Symposium on Artificial Intelligence*, pages 51–67. Springer.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023a. 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*, pages 1035–1044.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023b. Lecardv2: A large-scale chinese legal case retrieval dataset. *arXiv preprint arXiv:2310.17609*.

Qingquan Li, Yiran Hu, Feng Yao, Chaojun Xiao, Zhiyuan Liu, Maosong Sun, and Weixing Shen. 2023c. Muser: A multi-view similar case retrieval dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5336–5340.

Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. 2021. Less is more: Pretrain a strong siamese encoder for dense text retrieval using a weak decoder. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2791.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2342–2348.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2021. Coliee 2020: methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers 12*, pages 196–210. Springer.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, and Shaoping Ma. 2023. Understanding relevance judgments in legal case retrieval. 41(3).

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Heng Wang, Jianfeng Xu, et al. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Feng Yao, Jingyuan Zhang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Yun Liu, and Weixing Shen. 2023. Unsupervised legal evidence retrieval via contrastive learning with approximate aggregated positive. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4783–4791.

ChengXiang Zhai. 2008. Statistical language models for information retrieval. *Synthesis lectures on human language technologies*, 1(1):1–141.

Haoxi Zhong, Chaojun Xiao, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Overview of cail2018: Legal judgment prediction competition. *arXiv preprint arXiv:1810.05851*.

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

A Evaluation Metrics

We use Mean Reciprocal Rank and Recall as evaluation metrics.

The Mean Reciprocal Rank is a statistical measure used to evaluate the performance of a query-based system, where the primary goal is to retrieve the highest-ranked item. MRR calculates the average of the reciprocal ranks of results for a sample of queries. The reciprocal rank of a query response is the multiplicative inverse of the rank of the first correct answer:

$$\text{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\text{rank}_i}$$

where Q is the number of queries, and rank_i is the rank position of the first relevant document for the i-th query.

Recall measures the ability of a model to retrieve all relevant instances in a dataset. It is defined as

the ratio of the number of relevant items correctly retrieved to the total number of relevant items in the database, which is critical in scenarios where missing any relevant item could be costly:

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items}}$$

B Generation Configuration

For the selected LLMs, we directly download model parameters from the official Hugging Face repositories for each model, and use the code provided by Hugging Face to conduct text generation. For the generation configuration, we use the official default configurations provided by each model.

C Fine-tuning Process

We initialize the model with Chinese-Roberta-WWM (Cui et al., 2021). We use the dual-encoder architecture (Karpukhin et al., 2020) to compute the dot product between two embedding vectors as the relevance score:

$$X(c) = [CLS]q[SEP] \tag{1}$$

$$X(s) = [CLS]s[SEP] \tag{2}$$

$$Emb(X) = transformer_{[CLS]}(X) \tag{3}$$

$$S(q, s) = Emb(X(q))^{\top} \cdot Emb(X(s)) \tag{4}$$

where q is the query, s is the statute, $transformer_{[CLS]}(\cdot)$ outputs a contextualized vector for each token and we select the "[CLS]" vector as the embedding vector of the input. In Equation 4, we regard the inner products of embeddings as the relevance score S.

For the loss function, we use the Softmax Cross Entropy Loss (Cao et al., 2007; Ai et al., 2018; Gao et al., 2021) to optimize the re-ranking and retrieval model, which is defined as:

$$\mathcal{L}(Q, s^+, N) = -\log \frac{exp(S(Q, s^+))}{exp(S(Q, s^+) + \sum_{s^- \in N} exp(S(Q, s^-))} \tag{5}$$

where S is the relevance score function which is defined in Equation 4. Q is the query, s^+ is the relevant statute and N is the set of irrelevant statutes randomly sampled from the corpus.

D Dense Retrieval Model Distilled from GPT-4

For each article a_i in the STARD corpus, we let GPT-4 generate a legal question q_i based on the a_i using the following prompt:

Prompt 1

Given the following known statutory article:
[Content of the statutory article]
Imagine a scenario in which a person without legal knowledge is seeking legal advice. Please generate a question that this party might ask.
Note: The question must be fully explainable using the statutory article mentioned above, and remember that the person who propose this question has never read the legal articles mentioned before.

We initialize the model with Chinese-Roberta-WWM (Cui et al., 2021). Then we use the following loss function to train the dense retriever which is defined as:

$$\mathcal{L}(q_i, a_i, N) = -\log \frac{exp(S(q_i, a_i))}{exp(S(q_i, a_i) + \sum_{s^- \in N} exp(S(q_i, s^-))} \tag{6}$$

where S is the relevance score function which is defined in Equation 4, and N is the set of irrelevant statutes randomly sampled from the corpus.