SLARD: A Chinese Superior Legal Article Retrieval Dataset

Zhe Chen[1]*, Pengjie Ren[2], Fuhui Sun[3], Xiaoyan Wang[3]†, Yunjun Li[1]‡, Siwen Zhao[1], Tengyi Yang[1],

[1] School of information Science and Engineering, Shandong University
[2]School of Computer Science and Technology, Shandong University
[3]Information Technology Service Center of People's Court

Abstract

Retrieving superior legal articles involves identifying relevant legal articles that hold higher legal effectiveness. This process is crucial in legislative work because superior legal articles form the legal basis for drafting new laws. However, most existing legal information retrieval research focuses on retrieving legal documents, with limited research on retrieving superior legal articles. This gap restricts the digitization of legislative work. To advance research in this area, we propose SLARD: A Chinese **S**uperior **L**egal **A**rticle **R**etrieval **D**ataset, which filters 2,627 queries and 9,184 candidates from over 4.3 million effective Chinese regulations, covering 32 categories, such as environment, agriculture, and water resources. Each query is manually annotated, and the candidates include superior articles at both the provincial and national levels. We conducted detailed experiments and analyses on the dataset and found that existing retrieval methods struggle to achieve ideal results. The best method achieved a R@1 of only 0.4719. Additionally, we found that existing large language models (LLMs) lack prior knowledge of the content of superior legal articles. This indicates the necessity for further exploration and research in this field.

1 Introduction

As society progresses, new regulations must be established to keep pace with rapid development (Dror, 1958; Donelan, 2022). During the drafting process, it is essential to retrieve relevant superior legal articles from existing documents as a legislative foundation. These articles, enacted by higher-ranking legislative bodies such as national or provincial legislatures, provide a guiding framework for subordinate regulations, ensuring alignment with overarching legal principles. This hierarchical structure is crucial for maintaining the integrity of the legal system, avoiding conflicts, and ensuring consistency in legal governance (Vinx, 2007; Posner, 1993). Lawmakers must ensure that proposed articles are consistent with existing superior legal articles, which are reviewed to prevent violations and promote coherence (Kealy, 2021). The retrieval of superior legal articles is also vital for legislative review, legal interpretation, and maintaining consistent legal frameworks (Kelsen, 2017).

Past research in digital legislative development has highlighted the importance of employing Natural Language Processing (NLP) and Information Retrieval (IR) technologies to enhance the accuracy of retrieving superior legal articles. This is crucial for maintaining legal coherence and preventing conflicts within the legal framework (Sansone and Sperlí, 2022; Van Gog and Van Engers, 2001; Curtotti et al., 2015; Opmane et al., 2019). In the past, legal information retrieval has predominantly centered on the retrieval of similar cases (Ma et al., 2021; Li et al., 2024) and on matching legal articles to specific legal issues (Sansone and Sperlí, 2022; Chalkidis et al., 2021; Su et al., 2024). Despite these advancements, several critical issues remain unresolved. First, while the focus on similar case retrieval has yielded significant insights, it often falls short in addressing the specific need for superior legal article retrieval. Second, the problem-legal article pair retrieval approach does not adequately cater to the nuanced requirements of retrieving superior legal articles. These limitations underscore a significant gap in existing research: the absence of a specialized dataset to facilitate the study of superior legal article retrieval. This gap hinders the development of more sophisticated retrieval systems capable of addressing the complexities inherent in legal hierarchies.

To bridge this gap, we present the Superior Legal Articles Retrieval Dataset (SLARD), designed

*Email: cz2021@mail.sdu.edu.cn
†Co-corresponding Author, Email: 428163395@139.com
‡Co-corresponding Author, Email: liyujun@sdu.edu.cn

Proceedings of the 31st International Conference on Computational Linguistics, pages 740–754
January 19–24, 2025.

to facilitate subsequent research in this domain. SLARD consists of 2,627 queries of municipal-level legal articles and 9,184 candidate articles, including 2,976 provincial-level and 6,208 national-level articles. This dataset is specifically geared towards article-level legal retrieval, characterized by higher information density and more abstract expressions than previous retrieval tasks. The development of SLARD involved a rigorous and systematic approach, eight workers with legal expertise were engaged to identify the relevant superior legal articles for each query. To ensure the dataset's quality, each annotation was conducted by one worker and subsequently verified by another. This thorough annotation process underscores the reliability of SLARD and provides a valuable resource for advancing research in superior legal article retrieval.

This study conducted extensive and detailed experiments to validate their effectiveness in retrieving superior legal articles. Multiple retrieval models were evaluated to establish a performance benchmark, including traditional IR methods and modern deep learning-based approaches. Notably, several LLMs were also assessed for their performance in superior legal article retrieval. The results highlight the strengths and weaknesses of different methods, providing insights for future research. In this work, our contributions include:

- We introduced the task of retrieving superior legal articles, a novel scenario in the legal information retrieval field that addresses a critical need in the legislative process.
- We created and released SLARD[1], the first publicly available dataset specifically designed for superior legal article retrieval and provides a valuable foundation for future research in legal information retrieval.
- We conducted extensive experiments using various retrieval models. This evaluation establishes a benchmark for the performance of these models on the superior legal articles retrieval task, offering insights into their strengths and limitations.

2 Related Work

The SLARD falls within the scope of legal information retrieval tasks. Existing legal information retrieval tasks can mainly be categorized into the following two types:

2.1 Similar case retrieval

This task requires analyzing the factual aspects of a query case and retrieving cases with similar content from a set of candidates. The Competition on Legal Information Extraction/Entailment (COLIEE) 2020 (Rabelo et al., 2022) released a similar case retrieval dataset containing 650 queries, each requiring the retrieval of similar cases from a corresponding set of 200 candidates. COLIEE 2021 (Rabelo et al., 2021) expanded the dataset size and did not provide specific candidate collections for each query, instead requiring retrieval from the entire set of candidate cases. Other work (Šavelka and Ashley, 2022) focuses on retrieving from case law with a legal article to argumentation about the meaning of the phrase. The LeCaRD series constructed a Chinese similar case retrieval dataset. LeCaRDv1 (Ma et al., 2021) contains 107 queries, each with 100 candidate cases. LeCaRDv2 (Li et al., 2024) refines the relevance criteria and expands the dataset.

2.2 Legal Articles Retrieval

Article retrieval focuses on finding relevant legal articles in response to specific queries, which are typically legal case documents or legal questions from the general public. From 2015 to 2017, the COLIEE competition focused on retrieving relevant articles from the Japanese Civil Code for given legal questions (Kim et al., 2015, 2016; Kano et al., 2017). A similar approach was applied in the French legal context, where a study (Louis and Spanakis, 2021) aimed to match 1,108 legal questions with the appropriate articles from a comprehensive collection of 22,633 articles. In the context of Chinese law, a study (Su et al., 2024) introduced a dataset that expanded the scope of such retrieval tasks, including 1,543 query cases and a large set of 55,348 candidate legal articles. The REG-IR (Chalkidis et al., 2021) involves retrieving relevant documents for UK/EU law queries, with both queries and candidates being lengthy and complex.

However, these tasks do not cover the specific scenario of retrieving superior legal articles relative to a given query article. The objective of superior legal article retrieval we proposed is to enhance the efficiency and accuracy of legal research, particularly in understanding the legislative hierarchy and the relationships between different legal articles.

[1]SLARD is publicly accessible at https://github.com/xiaobo-Chen/SLARD for further study.

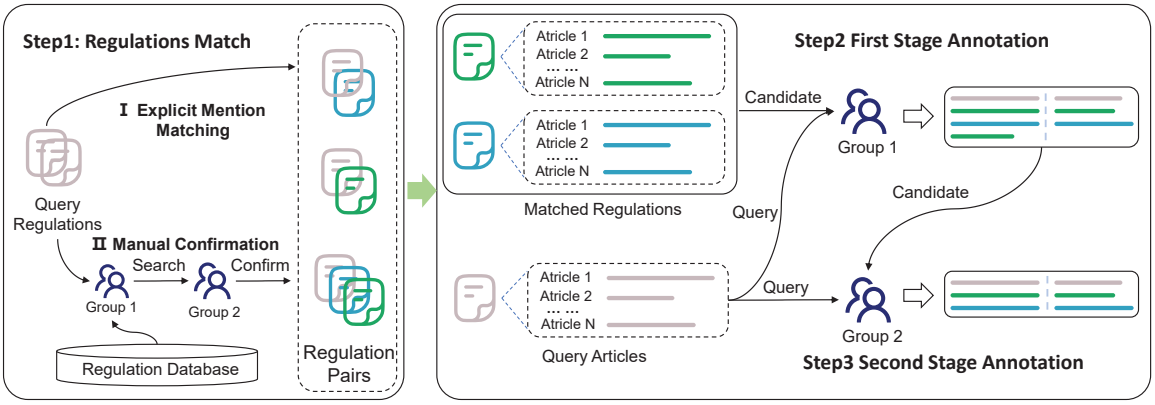


Figure 1: Overview of the Construction Process of the SLARD.

3 Dataset Construction

Eight undergraduate workers with a law background are hired to perform the annotations to build a high-quality and reliable SLARD. The construction process of SLARD is shown in Fig 1. Firstly, original superior legal regulation pairs should be collected at the regulation level. Secondly, manually annotate each pair of articles and identify the superior legal articles at the article level. Finally, recheck the annotation results from the last step to ensure the quality of the dataset.

3.1 Task Definition

The task of superior legal articles retrieval is to identify articles related to the query articles from a set of candidates with higher legal effectiveness, thus providing a legal basis for legislators drafting new articles. Specifically, given the query article q and candidate set of legal articles $D = \{d_1, d_2, \ldots, d_k, \ldots, d_{i-1}, d_i\}$, with i indicating the quantity of superior legal regulations, $d_k = \{a_{k1}, a_{k2}, \ldots, a_{kj}\}$, where j denotes the number of articles within regulations d_k, the task involves retrieving the top-k related articles $D_q = \{a_k | a_k \in D\}$ with the highest degree of relevance to the query q.

3.2 Superior Legal Regulations Collection

To construct the SLARD, we collected 150 municipal regulations from the China Law and Regulations Database, covering 32 categories of topics. Each municipal regulation was then matched to the corresponding provincial and national superior regulations. This matching process involved two specific methods:

Explicit Mention Matching: Superior regulations explicitly mentioned within the text of the municipal regulations were identified and extracted. This method ensures that any legal articles directly referred to by the municipal regulation are included in our dataset.

Manual Confirmation by Legal Experts: In cases where the municipal regulations did not explicitly mention superior regulations, annotators with a legal background manually retrieved and confirmed the relevant superior regulations. Initially, the first group of annotators conducted searches within the legal database to identify the superior regulations they deemed appropriate. Subsequently, the results from the first group were reviewed by a second group of annotators. If the second group agreed with the results, these were accepted as the final regulation matches. In cases of disagreement, the data was randomly assigned to a third annotator in group 2 for final confirmation.

After obtaining the superior regulations, we systematically extracted the individual legal articles from each regulation for subsequent annotation. Formally, let the set of municipal regulations be denoted as M = $\{m_1, m_2, \ldots, m_n\}$, where n represents the number of municipal regulations. Each municipal regulation m_i is associated with a set of articles $Q_i = \{q_{i1}, q_{i2}, \ldots, q_{ij}\}$, where j denotes the number of articles in m_i. The task is to identify a set of superior legal regulations $D_i = \{d_{i1}, d_{i2}, \ldots, d_{ik}\}$ for each municipal regulation m_i, where k indicates the number of superior regulations. The resulting dataset D consists of pairs (Q_i, D_i) for all municipal regulations.

3.3 Manually Annotating

After obtaining the regulation-level matches, the next step involved annotating at the article level. First, we extracted all the collected municipal reg-

ulations and identified the corresponding superior regulations. The scope of the superior regulations included the corresponding provincial and national regulations collected as described in Section 3.2. To ensure a thorough and accurate annotation process, we implemented a two-stage annotation procedure, dividing the eight undergraduate workers into two groups of five and three.

In the first stage, each worker independently reviewed the municipal articles and matched them to the relevant superior articles, annotating the articles they believed to be superior based on their content and relevance to the municipal articles. This stage is tasked with identifying as many correctly matching higher-level articles as possible while ensuring content relevance. The annotators were instructed to be inclusive, allowing suspected superior articles to be annotated as correct superior articles. The objective was to ensure that no potential superior legal articles were missed.

In the second stage, a team of three reviewers assessed the annotations made by the initial group. Their task was to verify whether each annotated article truly qualified as a superior article, refining the initial annotations to ensure the accuracy and reliability of the final dataset. The team systematically reviewed each annotation, evaluating the relevance and accuracy of the labeled superior articles and retaining only those that genuinely met the criteria for superior legal articles. In cases of inconsistencies, the disputed annotations were randomly reassigned to another worker from the second team for confirmation.

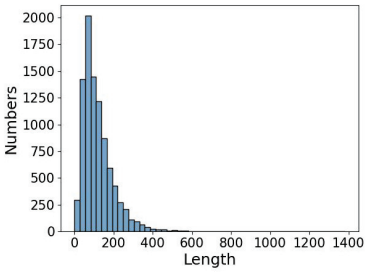
Through this rigorous two-stage annotation process, the SLARD was curated to ensure both breadth and accuracy. This structured approach, leveraging the expertise of workers with a legal background and a thorough verification mechanism, ensures the dataset's robustness and utility for legal analysis and retrieval tasks.

3.4 Quality Assurance

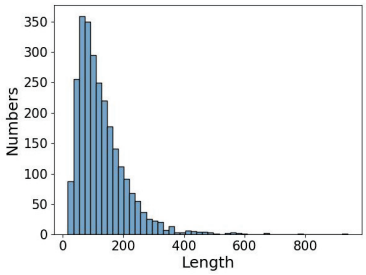
We implemented several measures to ensure the quality of the data. First, we provided comprehensive training for the annotators before the annotation process began. This training covered the task definition, the specifics of the two-stage annotation process, and the criteria for identifying superior legal articles. Additionally, we provided detailed guidelines[2] to ensure consistency in annotation standards. These guidelines included examples and counterexamples of superior legal articles, definitions of key legal terms, and instructions on how to handle ambiguous cases.

The annotation process itself was designed to minimize errors and ensure reliability. Each query legal article was annotated independently by two different annotators. This redundancy helps to reduce the errors that can arise from single-person annotation and ensures that the final dataset reflects a consensus among multiple legal experts, thereby increasing its reliability.

To further ensure the quality of the annotations, we sampled 5% of the data for a final inspection. This sample was reviewed by a team of senior legal experts who checked for errors and inconsistencies. If errors were found in the sampled data, the relevant sections were flagged, and the annotators were asked to review and correct them until no errors were found in the sampled data. This feedback loop helped to maintain high standards of accuracy throughout the dataset.



(a) Candidate length distribution



(b) Query length distribution

Figure 2: The length distribution of queries and candidates

3.5 Dataset Statistics

After the construction process described, we obtained a total of 2,627 manually annotated articles from 150 municipal regulations as query articles and 9184 articles as candidate articles which

[2]Detailed content can be found in our GitHub repository.

include 2,976 provincial-level articles and 6,208 national-level articles. The SLARD covers 32 distinct categories, ensuring broad coverage across various legal topics. As depicted in Fig 2, the SLARD reveals that the average length of a query article is 127 tokens, while the average length of articles within the candidate set is 119 tokens. This indicates a relatively balanced length distribution between the query and candidate articles.

4 Experimental Setups

4.1 Benchmark Settings

In our experiment, several models were fine-tuned to evaluate their performance on the SLARD. The dataset was partitioned into training and test sets with a 3:7 ratio for each regulation category, resulting in 1,978 samples for training and 649 for testing. From a practical application perspective, after consulting legal professionals, retrieving the top-5 results for reference was deemed acceptable. Therefore, retrieval performance was assessed using *Recall@K* $K \in (1, 3, 5)$ and Mean Reciprocal Rank (*MRR*) @5 as evaluation metrics.

For the implementation of the BM25 algorithm, we utilized Elasticsearch. The docT5query model was implemented using the mt5 model (Xue et al., 2020). General pre-trained models were directly loaded from the Hugging Face model hub, ensuring that state-of-the-art models were used for comparison. Retrieval-oriented pre-trained models were based on the Chinese-BERT-WWM model, following the official implementation guidelines.

For the HyDE method, we employed the BGE[3] (Xiao et al., 2023) model as the embedding model, and the LLMs indicated in Table 3 were represented by Qwen1.5-7B-Chat by default. During the training of the neural retrieval models, we set the maximum input length to 256 tokens and used a batch size of 16. To generate negative examples, we followed the approach of previous work (Kim et al., 2016; Wrzalik and Krechel, 2021), deriving these examples from incorrect search results produced by BM25. The ratio of positive to negative examples was set at 1:15.

4.2 Baselines

Four types of widely used retrieval models were used as baselines in this experiment: Sparse Retrieval Models, Generic Pre-trained Retrieval Models, Retrieval-oriented Pre-trained Models, and Retrieval Models Based on Large Language Models.

- **Sparse Retrieval Models**

  BM25(Robertson et al., 2009) is a traditional sparse retrieval model based on word frequency and document length.

  docT5query(Nogueira et al., 2019) enhances query robustness by generating related queries.

- **Generic Pre-trained Retrieval Models**

  Chinese-BERT-WWM(Cui et al., 2021) is the Chinese version of Bert trained with Whole Word Mask (WWM) and Next Sentence Prediction(NSP) tasks.

  Chinese-RoBERTa-WWM(Cui et al., 2021) is trained in enlarged datasets with only WWM tasks with the same architecture as Bert.

  Lawformer(Xiao et al., 2021) is pre-trained on a legal corpus and extends the maximum supported input length of the model and enhances its performance in scenarios involving long legal texts.

- **Retrieval-oriented Pre-trained Models**

  DPR(Karpukhin et al., 2020) proposed a bi-encoder architecture, which maps all text into a low-dimensional continuous space to achieve highly robust semantic retrieval performance.

  RetroMAE(Xiao et al., 2022) proposed a Masked Auto-Encoder pre-training strategy to enhance the model's representation capabilities at the sentence level.

  ColBERT(Khattab and Zaharia, 2020) performs late interaction at the token level to calculate the sentence similarity.

- **Retrieval Models Based on LLM**

  HyDE(Gao et al., 2022) uses pseudo documents generated by LLMs for semantic alignment, the pseudo documents are embedded and then vector retrieved to obtain relevant results.

  Query2Doc(Wang et al., 2023) uses LLMs for query expansion and concatenation with the query for subsequent sparse or dense retrieval.

[3] https://huggingface.co/BAAI/bge-base-zh

Model		Metrics			
		R@1	R@3	R@5	MRR@5
Sparse Retrieval Models	**BM25**	**44.62**	**70.17**	**76.65**	**57.69**
	docT5query	38.14	60.88	67.6	49.94
Generic Pre-trained Models	**Chinese-BERT-WWM**	25.55	33.01	33.99	29.12
	Chinese-RoBERTa-WWM	**27.63**	**35.82**	**38.51**	**31.94**
	Lawformer	26.16	34.6	37.41	30.57
Retrieval-oriented Models	**DPR**	**47.19**	74.33	81.3	61.07
	RetroMAE	**47.19**	**74.57**	**81.66**	**61.18**
	ColBERT	40.22	66.63	73.47	53.76
Large Language Model For Retrieval	**HyDE**	18.34	28	32.89	23.78
	Query2doc$_{+BM25}$	38.75	63.08	70.05	51.46
	Query2doc$_{+DPR}$	**41.2**	**65.04**	**72.37**	**53.51**

Table 1: Performance of different models on SLARD. The top-performing model for each method is highlighted in bold, while the second-best results are underlined.

5 Results and Analyses

In this section, we present and analyze the performance of various retrieval models on the SLARD. Through these experiments, we aim to provide a comprehensive assessment of the strengths and weaknesses of different retrieval approaches in the context of legal article retrieval and highlight areas for improvement.

5.1 Performance of existing methods on the SLARD

The performance of existing methods on the SLARD is presented in Table 1. Based on the experimental results, several insights can be drawn:

The sparse retrieval method, specifically BM25, demonstrates competitive performance in retrieving superior legal articles compared to other methods. Superior legal articles often form the basis for current articles, leading to significant overlap in vocabulary and phrasing between the query and the superior articles. This overlap enables BM25, which relies on term frequency and inverse document frequency, to achieve relatively good results by effectively matching similar terms. However, these results are only relatively good; with an R@5 of 76.65% and an R@1 of 44.62%, the performance remains less than satisfactory. This indicates that superior legal article retrieval continues to pose a challenge for the BM25 method.

In contrast, general pre-trained models significantly underperform compared to other methods. For example, Chinese-RoBERTa-WWM achieves only 33.85% in terms of R@5, which is inadequate for practical applications. This highlights that merely enhancing the representation capability of generic pre-trained models for legal texts does not yield satisfactory results. One key reason is the high degree of condensation and specificity in legal language, which poses challenges for general models to capture the necessary nuances. Additionally, Lawformer, despite being pre-trained on legal case data, does not perform optimally among general pre-trained models. This suboptimal performance can be attributed to the mismatch between its training data distribution and the actual content of the laws in the SLARD dataset.

Retrieval-oriented pre-trained models demonstrate the best overall performance. Fine-tuning these models with a substantial number of negative samples enables them to better differentiate between relevant and irrelevant articles, thereby improving accuracy in retrieval tasks. The top-performing method achieved 81.66% in R@5; however, R@1 is only 47.19%, indicating that approximately half of the superior legal articles are still missed. This highlights the ongoing need for improvements in recall rates for this task.

The performance of LLMs in retrieval tasks reflects their prior knowledge of superior legal articles. However, LLMs perform worse on the SLARD dataset, even compared to the traditional BM25 algorithm. This suggests that although LLMs are trained on a vast amount of general knowledge, the specificity and detailed nature of legal texts require more focused fine-tuning to enhance their performance in the legal domain.

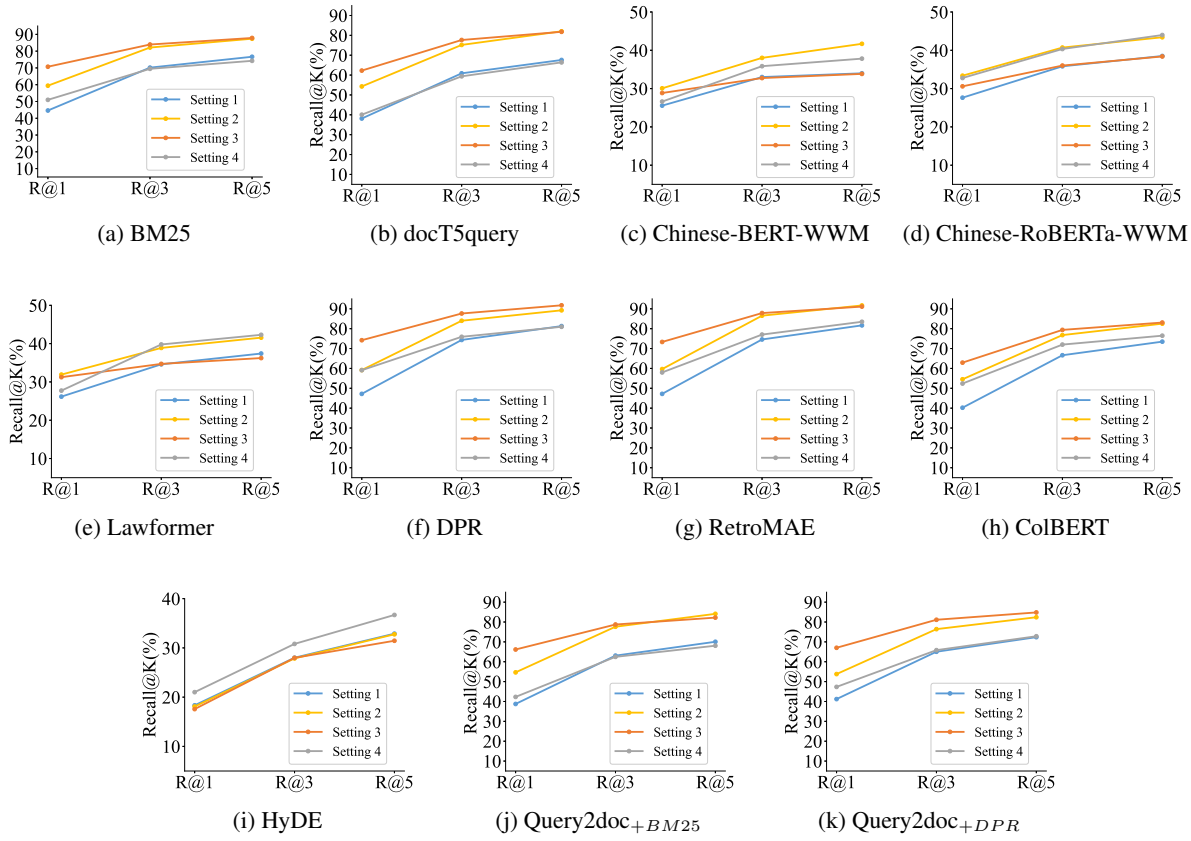


Figure 3: The performance of the models under different candidate collection settings.

5.2 Performance under different candidate collections

In this section, we classify the candidate collections into four settings based on their levels of legal effectiveness to evaluate the model's performance across various scenarios[4].

Setting 1 (Section 5.1): The candidate collection includes all 9,184 legal articles with higher effectiveness levels, encompassing both provincial and national articles.

Setting 2: The candidate collection is tailored to the given query article and consists of articles contained in the corresponding superior regulations. In this setting, the candidate collection includes an average of 73 articles for each query.

Setting 3: The candidate collection is restricted to provincial-level legal articles, totaling 2,976 articles. This setting assesses the model's performance within a specific jurisdictional scope, focusing on the retrieval of regional articles.

Setting 4: The candidate collection is limited to national-level legal articles, comprising a total of 6,208 articles. This setting evaluates the model's ability to identify relevant national articles.

Based on the results shown in Figure 3, we can draw the following conclusions:

In Setting 2, reducing the candidate set from 9,184 articles to an average of 62 significantly enhances all performance metrics. This enhancement is anticipated because a smaller candidate set simplifies the retrieval task, thereby facilitating the models' ability to pinpoint relevant articles. Typically, only one or two articles are pertinent to the query content. Nonetheless, the task of SLARD remains challenging. Under the R@1 metric, the best-performing method achieves only 59.66%, indicating that accurately retrieving the most relevant article continues to be a formidable challenge.

In Setting 3, where the candidate set comprises provincial articles, there is a marked improvement in performance metrics, with a 27% increase in R@1 for DPR compared to Setting 1. This boost is likely due to the narrowed query scope (from 9,184 to 2,976), which reduces interference from irrelevant results and the inherent similarities in context and lexicon between provincial and the queried mu-

[4]The accurate numerical results can be found in Appendix E.

Method	Model	Metrics			
		R@1	R@3	R@5	MRR@5
HyDE	Qwen	18.34	**28**	**32.89**	23.78
	ChatGLM	19.56	**28**	32.4	24.28
	Baichuan	**20.05**	27.75	31.78	**24.36**
Query2doc$_{+BM25}$	Qwen	38.75	63.08	70.05	51.46
	ChatGLM	39.98	**66.01**	71.03	52.64
	Baichuan	**41.44**	65.65	**72.25**	**53.96**
Query2doc$_{+DPR}$	Qwen	41.2	65.04	72.37	53.51
	ChatGLM	41.44	64.55	72.13	53.43
	Baichuan	**42.3**	**65.77**	**73.35**	**54.76**

Table 2: The retrieval performance of different Large Language Models on SLARD. The top-performing model for each method is highlighted in bold, while the second-best results are underlined.

nicipal articles. These factors ease the challenge of semantic representation, corroborating the hypothesis that documents sharing similar scopes and terminologies yield better retrieval outcomes.

In Setting 4, a notable decline in performance is observed across almost all models (except for ColBERT) relative to Setting 3. National articles, being more general, diverse, and abstract, increase the complexity for models to accurately discern relevant from irrelevant articles. Although the query scope is narrower than in Setting 1, enhancing retrieval metrics to a degree, the overall improvement is modest due to the heightened challenge of representing these documents accurately. This outcome highlights the critical role of document specificity and abstraction levels in optimizing legal document retrieval models.

5.3 Performance of different LLMs

In this section, we evaluate the performance of three widely used open-source LLMs for Chinese on the SLARD: Qwen1.5-7B-Chat (Bai et al., 2023), ChatGLM2-6B (GLM et al., 2024), Baichuan2-7B-Chat (Yang et al., 2023). Each of these models has been trained on extensive corpora.

Our experiment is grounded in the assumption that LLMs possess world knowledge about the content of superior legal articles acquired during training, which can assist in the retrieval task. We aim to comprehensively evaluate the memory and retrieval capabilities of existing open-source LLMs concerning superior legal articles. The experimental results are presented in Table 2.

The results demonstrate that Baichuan consistently outperformed the other LLMs across almost all metrics and retrieval methods. Although its R@3 and R@5 scores (27.75% and 31.78%, respectively) were slightly lower than those of Qwen and ChatGLM, Baichuan achieved the highest MRR@5 of 24.36%, indicating superior overall ranking quality. In the Query2doc$_{+BM25}$ and Query2doc$_{+DPR}$ methods, Baichuan's performance was markedly better across all metrics. These findings suggest that Baichuan2-7B-Chat possesses a higher degree of legal knowledge and superior retrieval capabilities for superior legal articles compared to the other models.

Despite these strengths, it is important to note that all LLMs, including Baichuan, underperformed compared to traditional retrieval methods and retrieval-oriented models. This underperformance underscores a critical limitation of current LLMs in specialized legal information retrieval tasks, highlighting the need for further fine-tuning and the incorporation of more domain-specific training data to enhance their performance.

Overall, our experimental findings suggest that while LLMs demonstrate promising potential in legal document retrieval, significant improvements are still necessary to effectively meet the specific demands of the legal domain.

6 Conclusion

In this paper, we introduce SLARD, a large-scale dataset for Chinese superior legal articles retrieval. SLARD includes 2,627 queries and 9,184 candidate articles across 32 categories, addressing a critical gap in legal information retrieval tasks. We evaluate several models on SLARD to establish a performance benchmark, with results indicating that it is a challenging dataset, particularly in retrieving national superior articles, where significant improvements are necessary. Moreover, experi-

ments reveal that existing open-source LLMs lack prior knowledge of superior legal articles. SLARD serves as a valuable benchmark for the development of advanced retrieval techniques and the fine-tuning of models specific to legal texts. It is anticipated that SLARD will become a foundational resource for legislative research and will advance the field of superior legal article retrieval, contributing to more efficient and coherent legal systems.

Limitations

We acknowledge two major limitations in this study that could be addressed in future research. The first limitation is the dynamic nature of regulations. Although we utilize the most up-to-date legal data, future modifications to existing regulations and the introduction of new ones remain a possibility. This inherent dynamism in regulatory frameworks presents challenges in maintaining the currency and accuracy of our analysis. The second limitation concerns the diversity of regulations. Regulations encompass a broad spectrum of social life, and while our dataset includes 32 common categories, some areas remain underrepresented. This diversity makes it challenging to ensure comprehensive coverage, underscoring the need for continuous updates and expansions of the dataset.

For future work, developing a more adaptive and scalable system that can automatically integrate new and updated regulations would be advantageous. Additionally, expanding the dataset to encompass a broader range of regulatory categories could further enhance the comprehensiveness and robustness of the analysis.

Ethics Statement

Ethical considerations have been a cornerstone of this research from the very beginning. Throughout the dataset construction process, we prioritized the well-being and rights of all annotators. We ensured transparency and consent by fully informing them about the nature of their tasks and the research objectives. To safeguard their welfare, we controlled working hours to prevent overwork and provided fair compensation. Annotators who demonstrated proficiency in their tasks received an average hourly wage of 40 yuan, which exceeds the local minimum wage (22 yuan) of in our area.

All data used in this study were sourced from publicly available information on official government websites, which can be freely accessed and downloaded by the public under the terms of service outlined on each site. These sources are intended for public use and do not require special permissions for data retrieval. No modifications were made to the original data to ensure data integrity and rigor. Annotators were explicitly instructed to refrain from copying or reproducing any copyrighted material without proper authorization and were reminded to cite sources appropriately when necessary.

We recognize the potential risks associated with automating legal articles retrieval. Our system is designed to complement rather than replace human expertise. It aims to enhance efficiency and accuracy in legal research, allowing professionals to focus on more complex and strategic aspects of their work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62472261,62102234).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. 2021. Regulatory compliance through doc2doc information retrieval: A case study in eu/uk legislation where text similarity has limitations. *arXiv preprint arXiv:2101.10726*.

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.

Michael Curtotti, Eric McCreath, Tom Bruce, Sara Frug, Wayne Weibel, and Nicolas Ceynowa. 2015. Machine learning for readability of legislative sentences. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 53–62.

Edward Donelan. 2022. *Regulatory governance: policy making, legislative drafting and law reform*. Springer Nature.

Yehezkel Dror. 1958. Law and social change. *Tul. L. Rev.*, 33:787.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Yoshinobu Kano, Mi-Young Kim, Randy Goebel, and Ken Satoh. 2017. Overview of coliee 2017. In *COLIEE@ ICAIL*, pages 1–8.

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*.

Sean J Kealy. 2021. Legislative scrutiny in the united states: dynamic, whole-stream revision. *The Theory and Practice of Legislation*, 9(2):227–249.

Hans Kelsen. 2017. *General theory of law and state*. Routledge.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Mi-Young Kim, Randy Goebel, Yoshinobu Kano, and Ken Satoh. 2016. Coliee-2016: evaluation of the competition on legal information extraction and entailment. In *International Workshop on Juris-informatics (JURISIN 2016)*.

Mi-Young Kim, Randy Goebel, and S Ken. 2015. Coliee-2015: evaluation of legal question answering. In *Ninth International Workshop on Juris-informatics (JURISIN 2015)*.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024. Lecardv2: A large-scale chinese legal case retrieval dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260.

Antoine Louis and Gerasimos Spanakis. 2021. A statutory article retrieval dataset in french. *arXiv preprint arXiv:2108.11792*.

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.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to doctttttquery. *Online preprint*, 6(2).

Inara Opmane, Juris Balodis, and Rihards Balodis. 2019. Governance of legislative requirements for the development of natural language processing tools. In *MIC 2019: Managing Geostrategic Issues; Proceedings of the Joint International Conference, Opatija, Croatia, 29 May–1 June 2019*, pages 13–27. University of Primorska Press.

Richard A Posner. 1993. *The problems of jurisprudence*. Harvard University Press.

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.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Jaromír Šavelka and Kevin D Ashley. 2022. Legal information retrieval for understanding statutory terms. *Artificial Intelligence and Law*, pages 1–45.

Weihang Su, Yiran Hu, Anzhe Xie, Qingyao Ai, Zibing Que, Ning Zheng, Yun Liu, Weixing Shen, and Yiqun Liu. 2024. Stard: A chinese statute retrieval dataset with real queries issued by non-professionals. *arXiv preprint arXiv:2406.15313*.

Ron Van Gog and Tom M Van Engers. 2001. Modeling legislation using natural language processing. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)*, volume 1, pages 561–566. IEEE.

Lars Vinx. 2007. *Hans Kelsen's pure theory of law: legality and legitimacy*. Oxford University Press, USA.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. corr abs/2303.07678 (2023). *arXiv preprint arXiv:2303.07678*.

Marco Wrzalik and Dirk Krechel. 2021. Gerdalir: A german dataset for legal information retrieval. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 123–128.

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.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. Retromae: Pre-training retrieval-oriented language models via masked auto-encoder. *arXiv preprint arXiv:2205.12035*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

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*.

A Example of SLARD

Table 3 presents an example of SLARD, including a municipal-level legal article and corresponding provincial and national superior legal articles.

B Prompt Template

The Table 4 presents the prompt templates for article generation used in HyDE and Query2doc.

C Category covered by SLARD

Table 5 presents the categories of legal articles included in SLARD, along with the number of instances in both the query and candidate sets.

D The results of LLMs' performance under different settings

Tabel 6 7 8 show the results of retrieval performance of different Large Language Models on SLARD in different candidate collection settings described on section 5.2.

Baichuan still performed the best overall. Compared to setting 1, the reduction in the number of candidate sets led to improvements in all metrics. However, the retrieval of national-level superior legal articles (setting 4) had relatively poor performance across all metrics, indicating that it remains a challenge.

E Numerical results under different

Tables 9 10 and 11 provide the numerical accuracy results shown in Figure 3.

Query Legal Article **Water and Soil Conservation Management Measures in Jining City, Article 4**: The municipal and county (or district) people's governments should strengthen the unified leadership of soil and water conservation efforts. They should incorporate soil and water conservation work into the local economic and social development plans, establish a target responsibility system and evaluation and reward-punishment mechanism for soil and water conservation, increase funding, and implement safeguard measures.
Superior Legal Articles **Water and Soil Conservation Regulations in Shandong Province, Article 4:** People's governments at or above the county level should strengthen the unified leadership of soil and water conservation efforts, incorporate soil and water conservation work into the local economic and social development plans, allocate special funds for the tasks determined by soil and water conservation plans, and organize their implementation. The state implements a target responsibility system and evaluation and reward-punishment mechanism for soil and water conservation at various local levels of government in key prevention and control areas for soil erosion. **Water and Soil Conservation Law of the People's Republic of China, Article 4:** People's governments at or above the county level should strengthen the unified leadership of soil and water conservation efforts, incorporate soil and water conservation work into local economic and social development plans, implement a soil and water conservation target responsibility system and evaluation and reward-punishment mechanism, and establish a coordination mechanism for soil and water conservation work to address major issues in this area. They should integrate the goals and tasks determined by soil and water conservation plans into the annual national economic and social development plans, allocate special funds in the fiscal budget, and organize their implementation.

Table 3: An example of superior legal article retrieval. The article 4 of the Water and Soil Conservation Management Measures in Jining City stipulates the responsibilities of local governments in aspects such as work planning, system construction, and financial investment for soil and water conservation. Its superior articles, Article 4 of the Water and Soil Conservation Regulations in Shandong Province and Article 4 of the Water and Soil Conservation Law of the People's Republic of China stipulate the responsibilities at the provincial and national levels respectively.

prompt template
You are a legal expert. Please provide the specific content of the superior legal article for the given current article. Only include the specific content of the superior legal article, without any additional information. ## Current Article: ## Superior Legal Article Content:

Table 4: Prompt templates for article generation used in experiments.

Category	Numbers	
	Query	**Candidate**
Forestry	148	293
Civil Affairs	313	721
Surveying	21	100
Energy	211	715
Agriculture	12	324
Real Estate	9	249
Environmental Protection	334	889
Tourism	133	288
Legal System	233	367
Animal Husbandry	18	425
National Security	37	243
Cultural Relics and History	0	246
Intellectual Property	97	302
Human Rights	171	484
Business Environment Optimization	96	139
Education	112	486
Land	104	313
Labor Unions	0	214
Water Conservancy	85	511
Sports	32	213
Constitution	18	86
Business Administration	38	207
Military	57	99
Advertising	40	119
Commerce and Trade	69	160
Industrial Management	57	405
Enterprise	30	120
Construction	65	55
Contracts	33	198
Fishing	20	92
Culture	22	48
Price	12	73

Table 5: Statistical of categories on SLARD

Model		Metrics			
		R@1	R@3	R@5	MRR@5
HyDE	Qwen	18.09	27.87	**32.76**	23.62
	ChatGLM	19.68	**28.24**	32.27	**24.34**
	Baichuan	**19.93**	27.87	31.91	24.33
Query2doc$_{+BM25}$	Qwen	54.65	77.63	84.11	66.52
	ChatGLM	54.89	78.85	84.23	66.89
	Baichuan	**55.75**	**79.34**	**84.96**	**67.71**
Query2doc$_{+DPR}$	Qwen	53.79	76.41	82.40	65.3
	ChatGLM	52.08	74.94	81.05	63.64
	Baichuan	**54.03**	**76.77**	**84.47**	**66.71**

Table 6: The retrieval performance of different Large Language Models on SLARD in Setting 2

Model		Metrics			
		R@1	R@3	R@5	MRR@5
HyDE	Qwen	17.57	27.98	**31.45**	23.08
	ChatGLM	**18.87**	**28.63**	31.24	**23.73**
	Baichuan	**18.87**	26.03	29.93	22.89
Query2doc$_{+BM25}$	Qwen	66.16	78.74	82.21	72.4
	ChatGLM	65.94	79.61	83.30	72.64
	Baichuan	**67.25**	**79.83**	**83.51**	**73.52**
Query2doc$_{+DPR}$	Qwen	67.03	81.13	84.82	74.15
	ChatGLM	65.51	**82.00**	**86.33**	73.5
	Baichuan	**67.68**	80.04	85.47	**74.27**

Table 7: The retrieval performance of different Large Language Models on SLARD in Setting 3

Model		Metrics			
		R@1	R@3	R@5	MRR@5
HyDE	Qwen	21.01	30.81	36.69	26.69
	ChatGLM	22.69	31.37	35.85	27.5
	Baichuan	**23.25**	**32.77**	**37.25**	**28.58**
Query2doc$_{+BM25}$	Qwen	42.30	62.46	68.07	52.68
	ChatGLM	45.38	65.27	70.87	55.63
	Baichuan	**47.90**	**65.55**	**72.27**	**57.45**
Query2doc$_{+DPR}$	Qwen	47.34	65.83	72.83	57.04
	ChatGLM	43.98	64.43	71.15	54.48
	Baichuan	**50.42**	**68.07**	**77.03**	**60.26**

Table 8: The retrieval performance of different Large Language Models on SLARD in Setting 4

Model		Metrics			
		R@1	R@3	R@5	MRR@5
Sparse Retrieval Models	**BM25**	**59.41**	**82.15**	**87.41**	**71.08**
	docT5query	54.28	75.18	82.03	65.34
Generic Pre-trained Models	**Chinese-BERT-WWM**	30.07	38.02	41.69	34.46
	Chinese-RoBERTa-WWM	**33.37**	**40.71**	**43.4**	**37.29**
	Lawformer	31.91	38.88	41.56	35.65
Retrieval-oriented Models	**DPR**	59.17	83.99	89.24	71.78
	RetroMAE	**59.66**	**86.55**	**91.69**	**73.29**
	ColBERT	54.52	76.77	82.52	66
Large Language Model For Retrieval	**HyDE**	18.09	27.87	32.76	23.62
	Query2doc$_{+BM25}$	**54.65**	**77.63**	**84.11**	**66.52**
	Query2doc$_{+DPR}$	53.79	76.41	82.40	65.30

Table 9: Numerical accuracy results in Setting 2

Model		Metrics			
		R@1	R@3	R@5	MRR@5
Sparse Retrieval Models	**BM25**	**70.72**	**83.95**	**87.85**	**77.44**
	docT5query	62.26	77.66	81.78	69.92
Generic Pre-trained Models	**Chinese-BERT-WWM**	28.85	32.75	33.84	30.91
	Chinese-RoBERTa-WWM	**30.59**	**36.01**	**38.39**	33.60
	Lawformer	31.24	34.71	36.23	**33.71**
Retrieval-oriented Models	**DPR**	**74.19**	87.64	**91.76**	**80.90**
	RetroMAE	73.32	**87.85**	91.11	80.33
	ColBERT	62.91	79.39	83.08	71.11
Large Language Model For Retrieval	**HyDE**	17.57	27.98	31.45	23.08
	Query2doc$_{+BM25}$	66.16	78.74	82.21	72.40
	Query2doc$_{+DPR}$	**67.03**	**81.13**	**84.82**	**74.15**

Table 10: Numerical accuracy results in Setting 3

Model		Metrics			
		R@1	R@3	R@5	MRR@5
Sparse Retrieval Models	**BM25**	**50.98**	**69.47**	**74.23**	**60.56**
	docT5query	40.06	59.38	66.39	50.05
Generic Pre-trained Models	**Chinese-BERT-WWM**	26.61	35.85	37.82	31.03
	Chinese-RoBERTa-WWM	**32.77**	**40.34**	**43.98**	**37**
	Lawformer	27.73	39.78	42.3	33.64
Retrieval-oriented Models	**DPR**	**59.1**	75.91	80.95	67.69
	RetroMAE	57.98	**77.03**	**83.47**	**68.1**
	ColBERT	52.38	71.99	76.47	61.95
Large Language Model For Retrieval	**HyDE**	21.01	30.81	36.69	26.69
	Query2doc$_{+BM25}$	42.3	62.46	68.07	52.68
	Query2doc$_{+DPR}$	**47.34**	**65.83**	**72.83**	**57.04**

Table 11: Numerical accuracy results in Setting 4