LeDQA: A Chinese Legal Case Document-based Question Answering Dataset

Bulou Liu
DCST, Tsinghua University
Quan Cheng Laboratory
Institute for Internet Judiciary, Tsinghua University
lbl20@mails.tsinghua.edu.cn

Zhenhao Zhu
Weiyang College, Tsinghua University
zhuzhenh22@mails.tsinghua.edu.cn

Qingyao Ai*
Quan Cheng Laboratory
DCST, Tsinghua University
Institute for Internet Judiciary, Tsinghua University
aiqingyao@gmail.com

Yiqun Liu
DCST, Tsinghua University
Institute for Internet Judiciary, Tsinghua University
yiqunliu@tsinghua.edu.cn

Yueyue Wu
DCST, Tsinghua University
Institute for Internet Judiciary, Tsinghua University
wuyueyue@mail.tsinghua.edu.cn

ABSTRACT

Legal question answering based on case documents is a pivotal legal AI application and helps extract key elements from the legal case documents to promote downstream tasks. Intuitively, the form of this task is similar to legal machine reading comprehension. However, in existing legal machine reading comprehension datasets, the background information is much shorter than the legal case documents, and the questions are not designed from the perspective of legal knowledge. In this paper, we present LeDQA[1], the first Chinese legal case document-based question answering dataset to our best knowledge. Specifically, we build a comprehensive question schema (including 48 element-based questions) for the Chinese civil law by legal professionals. And considering the cost of human annotations are too expensive, we use one of the SOTA LLMs (i.e., GPT-4) to annotate the relevant sentences to these questions in each case document. The constructed dataset originates from Chinese civil cases and contains 100 case documents, 4,800 case-question pairs and 132,048 sentence-level relevance annotations. We implement several text matching algorithms for relevant sentence selection and various Large Language Models(LLMs) for legal question answering on LeDQA. The experimental results indicate that incorporating relevant sentences can benefit the performance of question answering models, but further efforts are still required to address the remaining challenges such as retrieving irrelevant sentences and incorrect reasoning between retrieved sentences.

CCS CONCEPTS

• **Applied computing** → **Law**.

[1]https://github.com/BulouLiu/LeDQA

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10.
https://doi.org/10.1145/3627673.3679154

KEYWORDS

Legal question answering, datasets, Large Language Models

ACM Reference Format:
Bulou Liu, Zhenhao Zhu, Qingyao Ai*, Yiqun Liu, and Yueyue Wu. 2024. LeDQA: A Chinese Legal Case Document-based Question Answering Dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679154

1 INTRODUCTION

Legal disputes are an inevitable part of everyday life, with many individuals finding themselves entangled in issues related to marriage, debts, or employment [8, 21]. However, most people have little to no knowledge about their rights and fundamental legal processes [2]. The rapid progress in natural language processing and the growing availability of digitized legal data present unprecedented opportunities to bridge the gap between people and the law [14, 16, 17, 19, 25]. Especially recently, revolutionary Large Language Models (LLMs) techniques have shown strong zero-shot and few-shot generalization ability in many natural language processing tasks. However, to further improve their legal question answering performances, it is necessary to incorporate legal knowledge into the LLMs because they are usually built with open-domain data and do not possess enough legal knowledge [13].

Legal case documents are official records which present arguments, evidence, and decisions in court cases. They are primary legal materials in various law systems along with statutes [10]. And they also serve as legal knowledge to enhance downstream legal applications [18]. Therefore, it is promising to improve the legal question answering systems based on legal case documents. Furthermore, legal question answering based on legal case documents can help extract key elements from the legal cases to promote downstream tasks [27]. So we propose the legal case document-based question answering task: generating accurate answers given a legal case document and corresponding questions. Intuitively, the form of this task is similar to legal machine reading comprehension. However, on the one hand, the background information is

*Corresponding author

Table 1: An example in LeDQA: a legal case document, a question and the corresponding answer and relevant sentences.

Legal Case Document: ...Appeal Request by Lu Xiudi and Chen Chen: To revoke the original judgment and reject Zou Guoqing's claim for the appellant to repay a loan of 1.4 million yuan and corresponding interest. Facts and Reasons: On the day Zou Guoqing handed over the loan principal of 1.5 million yuan, the appellants immediately paid interest of 177,000 yuan, risk control fees of 118,000 yuan, intermediary service fees of 90,000 yuan, and paid 20,000 yuan to a third party and 57,000 yuan to Zou Guoqing's representative. Later, the appellants sold a house and had 100,000 yuan earnest money taken away, indicating the nature of a loan trap, which should not be protected by law ... **On the same day, Zou Guoqing transferred 1,500,000 yuan to Lu Xiudi.** ... Zou Guoqing claimed that Lu Xiudi and Chen Chen had returned 100,000 yuan of the principal on May 29, 2018, but had not paid any interest thereafter, and he did not acknowledge having engaged an outsider named Shi Lei. **The first-instance court held that the evidence provided by Zou Guoqing was sufficient to prove the existence of a loan agreement between the parties and that the loan had been actually delivered**...	
Question: Was the loan delivered?	**Answer:** Yes.
Relevant sentence 1: On the same day, Zou Guoqing transferred 1,500,000 yuan to Lu Xiudi. **Relevant sentence 2:** The first-instance court held that the evidence provided by Zou Guoqing was sufficient to prove the existence of a loan agreement between the parties and that the loan had been actually delivered.	

much shorter (mostly 200-500 words) than the legal case documents (mostly 2000-3000 words) in existing legal machine reading comprehension datasets [7]. On the other hand, most of the questions in existing datasets are too general for legal applications, such as the defendant name and the incident location, which are not designed from the perspective of legal knowledge.

To overcome these problems, we present **LeDQA**, the first Chinese **Le**gal case **D**ocument-based **Q**uestion **A**nswering dataset to our best knowledge. Specifically, we focus on private lending cases, as they are the most complex and voluminous among all civil cases. And we employ a legal expert team to design a comprehensive question schema which includes 10 categories with 48 legal element-based questions. These questions are designed from the legal perspective and have a higher practical legal applicability. And we collect 100 representative private lending case documents as background information from the legal cases published by the Supreme People's Court of China. The average number of words in these legal case documents are 2440, ensuring a certain level of difficulty for the task. And considering the cost of human annotations are too expensive, we use one of the SOTA LLMs (i.e., GPT-4) to annotate the relevant sentences to these questions in each case document. Totally, the LeDQA dataset contains 100 case documents, 4,800 case-question-answer triplets and 132,048 sentence-level relevance annotations.

We implement several text matching algorithms for relevant sentence selection, and legal pre-training machine reading comprehension models and various LLMs for legal question answering on LeDQA. The experimental results show that incorporating relevant sentences can benefit the performance of question answering models. Error analysis shows that there are remaining challenges such as retrieving irrelevant sentences and incorrect reasoning between retrieved sentences.

2 RELATED WORK

Addressing legal questions has long posed intricate challenges within the legal NLP community, stemming from the inherent complexities of legal texts [4, 12, 14–16, 19, 25]. [11] introduced a dataset for statutory reasoning in tax law and [29] presented a multi-choice question answering dataset designed to assess professional legal expertise. Recently, researchers focused on utilizing legal knowledge to enhance the question answering models. [7] crafted a judicial reading comprehension dataset in the Chinese language and [3, 18] offered a corpus featuring question-answer pairs as well as a pool of law articles in Chinese and French, respectively. Compared to these works, LeDQA utilizes the legal case documents as background information and design a question schema from the legal perspective.

3 DATASET CONSTRUCTION

In this section, we introduce our dataset construction methods. Our goal is to build a legal case document-based question answering dataset with sentence-level relevance annotation. Therefore, our task is to define a question schema, select the legal case documents and annotate the relevant sentences and answers to the questions. In this paper, we focus on private lending cases, as they are the most complex and voluminous among all civil cases. An example of our dataset is shown in Table 1.

3.1 Question Schema Construction

To comprehensively construct the legal question schema that describes private lending cases, we employed a legal expert team including 5 legal professionals such as judges and prosecutors. First, they read the law articles and regulations related to private lending and list the key legal elements they consider important. They then engaged in group discussions, merged their collections of legal elements, and added or removed some items. They divided these elements into 10 categories: guarantee situations; loan disbursement and repayment; collateral or security; loan contracts and evidence; contract validity; marital and economic relationships; funding sources and legal fees; litigation procedures and disputes; joint debt and shared usage; and company borrowing and identity confusion. And they rephrased each element into a question. Details can be found in our github link.

3.2 Legal Case Document Selection

To select representative legal case documents as background information for question answering, we first screened 462 authoritative cases from 7000 private lending cases. They have been examined by the Supreme People's Court of China and considered to be of reference and demonstrative value for similar cases. And to comprehensively cover the various categories of the question schema in Section 3.1, the legal expert team read the 462 authoritative cases and selected the most important element for each case. Then they classified the authoritative cases into 10 categories in the question schema based on the selected most important elements. Finally, we selected 100 legal case documents by randomly choosing 10 authoritative cases in each category.

3.3 Relevant Sentence and Answer Annotation

To evaluate the legal question answering models, we need to assign answers for all the document-question pairs. Here, all answers are selected from "yes/no/unknown," indicating whether the corresponding legal element exists in/does not exist in/cannot be determined to exist in the legal case document. Additionally, due to the length of legal documents, retrieving sentences related to the questions in the document intuitively can optimize the performance of the legal question answering model. Therefore, for each question, we need to assign relevance labels (1:relevant, 0:irrelevant) for all the sentences in the case documents. Considering the cost of human annotations are too expensive, we use one of the SOTA LLMs (i.e., GPT-4) to annotate the relevant sentences to these questions in each case document. And to ensure the annotation quality, we recruited another three PhD students majoring in Chinese civil law as annotators to assign answers for one of the document-question pairs. The Fleiss's κ scores of answer annotations and relevant sentence annotations were 0.853 and 0.827, respectively, indicating almost perfect agreement [9]. If there were disagreements, we took the result of the majority vote. And for the answers, if all three annotated scores are inconsistent, they engaged in group discussions to make the final decision. Based on the human anntations, the F1 scores of GPT-4 answer and relevance annotations are 0.923 and 0.891, respectively, indicating that GPT-4 can serve as a good annotator. Table 2 shows the statistics of our LeDQA dataset. We can find that the background information in LeDQA (i.e., legal case documents) is much longer than that of existing legal machine reading comprehension datasets (usually not exceeding 500 words such as average 441 words in CJRC) [3, 7]. And each question is only relevant to few sentences in the legal documents. This indicates that the tasks in LeDQA are more challenging compared to previous tasks because they contain more noisy information.

4 EXPERIMENTS

4.1 Relevant Sentence Retrieval

It is important to retrieve relevant sentences for legal case document-based question answering because it can exclude the noisy information. Therefore, we first evaluate the several baseline models for relevant sentence retrieval following a 5-fold cross-validation (i.e., 80 training documents and 20 testing documents each time).

Table 2: The statistics of the LeDQA dataset.

Statistic	Number
Total legal case documents	100
Total document-question pairs	4,800
Avg. sentences in each document	27.51
Avg. relevant sentences for questions	3.963
Avg. words in each document	2,440
Avg. words in relevant sentences for questions	210.0
"Yes/No/Unknown" ratio in the answers	28%/42%/30%

Table 3: Evaluation of the relevant sentence retrieval task.

Methods	R@3	R@5	MRR
BM25	0.2407	0.3787	0.3822
LMIR	0.1376	0.2463	0.2575
TF-IDF	**0.2511**	**0.3829**	**0.4091**
Chinese-Bert-WWT	0.1681	0.2817	0.3213
Chinese-Roberta-WWT	0.1862	0.3110	0.3577

Specifically, as for bag-of-words IR methods, we select three representative models BM25 [23], LMIR [22] and TF-IDF [24]. And we also utilize some pre-trained dense retrieval models including Chinese-BERT-WWM and Chinese-RoBERTa-WWM [6]. The two models are trained with Whole Word Mask (WWM) in Chinese. They generate the representations of the question and the candidate sentence, respectively and use the cosine similarity scores as the relevance scores.

It is essential that the retriever returns as many relevant sentences as possible within the first top-k results, which implies a primary interest in recall at small cutoffs (R@3/5). Additionally, we report the mean reciprocal rank (MRR) which offers valuable insights into the position of the first relevant result. The results are shown in Table 3. We can find that most bag-of-words models perform better than pre-training models and the TF-IDF model achieves the best performance, indicating that they can better find the important words in the questions. However, the top-5 sentences returned by TF-IDF only cover 40% of the relevant sentences. This demonstrates that the current retriever is not sufficient to accurately extract content relevant to specific legal questions from case documents.

4.2 Legal Case Document-based QA

We evaluate several baseline models on the legal case document-based question answering task. Specifically, we feed the background information and the question into LLMs to generate the answers. We evaluate five LLMs on this task totally, including three Chinese widely-adopted LLMs (i.e.,Baichuan2-13B-Chat [26], Qwen-7B-Chat [1] and ChatGLM3-6B [28]), one Chinese legal-specific LLMs (i.e., ChatLaw [5]) and one of the most widely used LLMs (i.e., GPT3.5-turbo [20]). And we set four variations of the baseline models by using different content as background information. **"Direct"** utilizes the whole legal case document as background information for the inputs of baseline models. **"CoT"** (i.e., Chain-of-Thought)

Table 4: Evaluation of the legal case document-based question answering task ("yes" vs. "no and unknown").

Method	Baichuan2-13B-Chat		ChatGLM3-6B		Qwen-7B-Chat		ChatLaw		GPT3.5-turbo	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Direct	0.6763	0.5203	0.4390	0.4369	0.7381	0.2451	0.6923	0.1203	0.6960	0.5250
CoT	0.7242	0.5281	0.6206	0.2989	0.7283	0.5034	0.4394	0.3146	0.7000	0.4752
Retrieve	0.7369	0.5543	0.6083	0.3821	0.7623	0.5605	0.5600	0.4015	0.6785	0.4764
Oracle	0.7213	0.5784	0.6240	0.3460	0.7404	0.5609	0.5129	0.3592	0.6642	0.4983

Table 5: Evaluation of the legal case document-based question answering task ("yes" vs. "no" vs. "unknown").

Method	Baichuan2-13B-Chat		ChatGLM3-6B		Qwen-7B-Chat		ChatLaw		GPT3.5-turbo	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Direct	0.4431	0.4139	0.3313	0.2799	0.3529	0.3272	0.2927	0.1497	0.4604	0.4263
CoT	0.4535	0.4567	0.3840	0.3134	0.4273	0.4302	0.2488	0.2082	0.4315	0.4360
Retrieve	0.4627	0.4648	0.4115	0.3413	0.4485	0.4598	0.2781	0.2207	0.4352	0.4247
Oracle	0.4792	0.4777	0.4106	0.3323	0.4554	0.4644	0.2919	0.2487	0.4358	0.4316

designs the prompt to make LLMs select relevant sentences by themselves and then generate the answers. **"Retrieve"** utilizes the top-5 retrieved sentences from the best baseline retrieval model TF-IDF in Section 4.1 and **"Oracle"** utilizes the human-annotated relevant sentences. The details of these prompts can be found in the github link. Additionally, apart from treating this problem as a three-class problem, we are more concerned whether the model can accurately identify "yes". Therefore, we consider "no" and "unknown" as one category, and evaluate the performance of the binary classification.

We select accuracy and Macro-F1 as the evaluation metrics. The results of the binary-class problem (i.e., "yes" vs. "no and unknown") and the three-class problem (i.e., "yes" vs. "no" vs. "unknown") are shown in Table 4 and Table 5, respectively. We can make the following observations. As for the three Chinese general LLMs, (1) Both **"CoT"** and **"Retrieve"** can improve the performances compared to using the whole documents. This shows that selecting relevant sentences from the case documents can exclude the noisy information. And **"Retrieve"** performs better than **"CoT"**, indicating that LLMs themselves cannot directly extract relevant sentences accurately from legal documents. (2) **"Oracle"** achieves the best performances among the four variations, demonstrating that the question answering performances can be further improved by designing more accurate relevant sentence retrieval models. In addition, we can find that the performances of the legal LLM ChatLaw are the worst, indicating that ChatLaw can only solve the specific legal tasks and can not generalize to others. Finally, we find that the differences between the four variations of GPT3.5-turbo are not significant. This shows that GPT3.5-turbo can solve the long texts well and find the relevant parts in the legal documents without external retrievers.

4.3 Error Analysis

Then we conducted error analysis and identified the challenges on the LeDQA dataset. We find that the questions with "unknown" as the correct answer are the most challenging. Then we recruited a PhD student majoring in law to check 100 wrong cases by GPT-4(retrieve). She found that the most common errors are, as mentioned before, retrieved sentences that are not relevant to the question, posing a challenge for the future work. Another common error is that the model fails to make the correct inference based on the relevant sentences. For example, there are two relevant sentences for the question "Was the loan delivered?" : "A provided the transfer records for the loan delivery." and "Upon review, it was found that the transfer records provided by A were forged." The correct answer is "No" but sometimes the model generates the answer "Yes" because it only considers the first sentence instead of inferring between the two sentences, posing another challenge. Further research is needed to explore how to tackle these two challenges on LeDQA.

5 CONCLUSION

In this paper, we present the first Chinese legal case document-based question answering dataset LeDQA. Specifically, we build a comprehensive Chinese civil legal question schema, collect representative case documents and annotate the relevant sentences and corresponding answers to these questions in each case document. We find that incorporating relevant sentences can benefit the question answering models by implementing text matching algorithms for relevant sentence selection and various LLMs for question answering. We will design the models to tackle the remaining challenges: retrieving irrelevant sentences and incorrect reasoning in the future. In addition, we will apply the LeDQA-based element extraction models to promote downstream legal AI applications, like legal case retrieval and legal judgement prediction, which can further show the usefulness of the question schema in the LeDQA dataset.

ACKNOWLEDGEMENTS

This work is supported by Quan Cheng Laboratory (Grant No. QCLZD202301) and the Natural Science Foundation of China (Grant No. 62002194).

REFERENCES

[1] 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* (2023).

[2] Nigel J Balmer, Alexy Buck, Ash Patel, Catrina Denvir, and Pascoe Pleasence. 2010. Knowledge, capability and the experience of rights problems. *London: PLEnet* (2010).

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

[4] Yanjiao Chen, Yuxuan Xiong, Bulou Liu, and Xiaoyan Yin. 2019. TranGAN: Generative adversarial network based transfer learning for social tie prediction. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

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

[6] 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 (2021), 3504–3514.

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

[8] Trevor CW Farrow, Ab Currie, Nicole Aylwin, Lesley Jacobs, David Northrup, and Lisa Moore. 2016. Everyday legal problems and the cost of justice in Canada: Overview report. *Osgoode Legal Studies Research Paper* 57 (2016).

[9] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[10] Hanjo Hamann. 2019. The German Federal Courts Dataset 1950–2019: From Paper Archives to Linked Open Data. *Journal of Empirical Legal Studies* 16, 3 (2019), 671–688.

[11] Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257* (2020).

[12] Bulou Liu, Bing Bai, Weibang Xie, Yiwen Guo, and Hao Chen. 2022. Task-optimized User Clustering based on Mobile App Usage for Cold-start Recommendations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3347–3356.

[13] Bulou Liu, Yiran Hu, Qingyao Ai, Yiqun Liu, Yueyue Wu, Chenliang Li, and Weixing Shen. 2023. Leveraging Event Schema to Ask Clarifying Questions for Conversational Legal Case Retrieval. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1513–1522.

[14] Bulou Liu, Yiran Hu, Yueyue Wu, Yiqun Liu, Fan Zhang, Chenliang Li, Min Zhang, Shaoping Ma, and Weixing Shen. 2023. Investigating Conversational Agent Action in Legal Case Retrieval. In *European Conference on Information Retrieval*. Springer, 622–635.

[15] Bulou Liu, Chenliang Li, Wei Zhou, Feng Ji, Yu Duan, and Haiqing Chen. 2020. An attention-based deep relevance model for few-shot document filtering. *ACM Transactions on Information Systems (TOIS)* 39, 1 (2020), 1–35.

[16] Bulou Liu, Yueyue Wu, Yiqun Liu, Fan Zhang, Yunqiu Shao, Chenliang Li, Min Zhang, and Shaoping Ma. 2021. Conversational vs Traditional: Comparing Search Behavior and Outcome in Legal Case Retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1622–1626.

[17] Bulou Liu, Yueyue Wu, Fan Zhang, Yiqun Liu, Zhihong Wang, Chenliang Li, Min Zhang, and Shaoping Ma. 2022. Query Generation and Buffer Mechanism: Towards a better conversational agent for legal case retrieval. *Information Processing & Management* 59, 5 (2022), 103051.

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

[19] Yixiao Ma, Yunqiu Shao, Bulou Liu, Yiqun Liu, Min Zhang, and Shaoping Ma. 2021. Retrieving Legal Cases from a Large-scale Candidate Corpus. In *Proceedings of the 18th International conference on Artificial Intelligence and Law*.

[20] OpenAI. 2022. Introducing ChatGPT. (2022).

[21] Alejandro Ponce, Sarah Chamness Long, Elizabeth Andersen, Camilo Gutierrez Patino, Matthew Harman, Jorge A Morales, Ted Piccone, Natalia Rodriguez Cajamarca, Adriana Stephan, Kirssy Gonzalez, et al. 2019. Global Insights on Access to Justice 2019: Findings from the World Justice Project General Population Poll in 101 Countries. *World Justice Project* (2019), 1.

[22] Jay M Ponte and W Bruce Croft. 2017. A language modeling approach to information retrieval. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 202–208.

[23] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*. Springer, 232–241.

[24] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[25] Yunqiu Shao, Bulou Liu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. THUIR@ COLIEE-2020: Leveraging Semantic Understanding and Exact Matching for Legal Case Retrieval and Entailment. *arXiv preprint arXiv:2012.13102* (2020).

[26] 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* (2023).

[27] Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A Large-Scale Chinese Legal Event Detection Dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*. 183–201.

[28] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).

[29] 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*, Vol. 34. 9701–9708.