Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on AI for Good

LEEC for Judicial Fairness: A Legal Element Extraction Dataset with Extensive Extra-Legal Labels

Zongyue Xue[1,2] , **Huanghai Liu**[1] , **Yiran Hu**[1] , **Yuliang Qian**[1] , **Yajing Wang**[1] ,
Kangle Kong[1] , **Chenlu Wang**[1] , **Yun Liu**[1] and **Weixing Shen**[1]
[1]School of Law, Tsinghua University
[2]Yale Law School
{zongyuexue, liuhh19, huyr17}@outlook.com,
{qianyl23,wangyj20, kkl22}@mails.tsinghua.edu.cn,
{wangchenlu, liuyun89, wxshen}@mail.tsinghua.edu.cn

Abstract

An extensive label system is pivotal to facilitate judicial fairness and social justice. Prior empirical research and our interview with legal professionals underscore the importance of extra-legal factors in criminal trials. To help identify sentencing biases and facilitate downstream applications, we introduce the **L**egal **E**lement **E**xtra**C**tion (LEEC) dataset comprising 15,919 judicial documents and 155 labels. This dataset was constructed through two main steps: **First**, designing the label system by legal experts based on prior empirical research which identified critical factors driving and processes generating sentencing outcomes in criminal cases; **Second**, employing legal knowledge to annotate judicial documents according to the label system and annotation guideline. LEEC represents the most extensive and domain-specific legal element extraction dataset for the Chinese legal system. Our experiments reveal that despite certain capabilities, both Document Event Extraction (DEE) models and Large Language Models (LLMs) face significant restrictions in legal element extraction tasks. Finally, our empirical analysis based on LEEC provides **evidence for judicial unfairness** in Chinese criminal sentencing and confirms the applicability of LEEC for future empirical research and other downstream applications.

1 Introduction

Extracting key elements and their relations from judicial documents is valuable to facilitate further research and promote judicial fairness. The divergence between the "law in books" and "law in action" is notable in judicial practice [Pound, 1910], which may undermine judicial fairness and social justice. Prior research has discerned that the practical application of the law is influenced significantly not only by legal factors but also by extra-legal ones. For instance, studies in Western jurisdictions have indicated that disparities in gender and ethnicity have significant impacts on criminal sentencing [Ulmer, 2012]. However, previous legal datasets have almost exclusively incorporated legal factors. Consequently, researchers who endeavor to empirically investigate the influence of extra-legal factors are impeded by the lack of comprehensive datasets. With the help of the extensive label system constructed on the legal knowledge graph by our team of legal experts, the Legal Element ExtraCtion (LEEC) dataset aims to provide comprehensive element mentions, trigger words, and values manually annotated from large-scale judicial documents with high quality[1]. This could facilitate the automatic extraction of elements, benefiting numerous LegalAI applications, such as Legal Judgement Prediction and Similar Case Retrieval, as well as the replication and innovation of empirical legal research.[2]

Inspired by the success of general-domain element extraction [Guo *et al.*, 2020; Hogenboom *et al.*, 2011; Liao and Grishman, 2010], previous studies [Feng *et al.*, 2022; Shen *et al.*, 2020; Sierra and others, 2018] attempted to construct an element extraction system in the legal domain, leveraging both hand-crafted features and neural networks. For instance, LeCaRD [Ma *et al.*, 2021], the first Legal Case Retrieval Dataset in China, contains 107 query cases and 10,700 candidate cases selected from over 43,000 Chinese criminal judgments. LEVEN is a large-scale Chinese Legal event detection dataset [Yao *et al.*, 2022], with 8,116 legal documents and 150,977 human-annotated event mentions in 108 event types. At present, the existing datasets in China also include CAIL [Xiao *et al.*, 2019], Criminal [Ma *et al.*, 2021], LERD [Yao *et al.*, 2023], CJO[3], and PKULAW[4], etc. However, there are several main challenges in the existing work:

(1) Incomprehensive Label System. Existing label systems [Li *et al.*, 2023a; Li *et al.*, 2023b; Li *et al.*, 2023c; Richards *et al.*, 2016; Liu *et al.*, 2023; Li *et al.*, 2023d] of prior studies mainly lay emphasis on a limited scope of

[1]The LEEC dataset sample is available at https://github.com/THUlawtech/LEEC. For access to the complete dataset, please contact us via computational_law@mail.tsinghua.edu.cn. The disclosed LEEC dataset is solely for non-commercial use.

[2]To access the full text of this paper including appendices, please refer to this link: https://github.com/THUlawtech/LEEC/blob/main/LEEC_IJCAI24.pdf.

[3]https://wenshu.court.gov.cn, accessed on April 5th, 2024.

[4]https://home.pkulaw.com, accessed on April 5th, 2024.

charge-oriented elements, which is far from enough. Existing studies predominantly focus on the legally prescribed factors in sentencing, overlooking extra-legal elements. However, a wealth of empirical research suggests that these elements, such as the defendant's and victim's age, gender, race/ethnicity, etc., may significantly influence trial and sentencing outcomes [Chen *et al.*, 2023; Doerner and Demuth, 2010; Richards *et al.*, 2016; Tran *et al.*, 2019; Ulmer, 2012]. The absence of these factors in the label system may compromise empirical studies on judicial fairness and the performance of downstream tasks.

(2) Lack of Domain Focus. An overwhelming majority of existing datasets [Nguyen *et al.*, 2016; Veyseh *et al.*, 2022] for element extraction mainly focus on the element or event extraction in the general domain. However, such datasets may not be well suited to applications in the legal domain. For example, Recidivist (*Leifan* in Chinese) and Previous Criminal Record (*Qianke* in Chinese) are closely connected yet distinct legal concepts in Chinese criminal law, which could be difficult to distinguish without adequate legal knowledge. Furthermore, judicial documents may depict different interpretations and perspectives on the same legal elements, such as whether the defendant voluntarily surrendered, confessed, or pled guilty, from various court participants. This can cause confusion without legal knowledge. Therefore, existing datasets in general domains are hardly applicable for comprehensive analysis based on legal texts owing to their lack of understanding of legal knowledge and contexts.

To provide a solid foundation for legal element extraction, LEEC makes the following improvements:

(1) Extensive Label System. Our team of legal experts not only expanded the coverage of legally prescribed factors that may significantly impact Chinese criminal trials, but also actively conducted interviews with legal professionals and utilized a comprehensive collection of empirical studies in Chinese contexts published in Chinese core legal journals and internationally. In this way, we were able to construct an extended and comprehensive label system in the legal domain, incorporating both legal and extra-legal key labels that may have a substantial impact on Chinese judicial practice.

(2) Large Scale. LEEC is annotated based on the publicly available cases of both LEVEN and LeCaRD, with a total of 15,919 cases. Therefore, the high coverage of cases could help alleviate the problem of the limited number of cases in few-shot charges, leading to an increased ability to meet needs in real-world court settings. Besides, the annotation of LEEC could be combined with the previous annotation from LEVEN and LeCaRD, providing more comprehensive information to facilitate the analysis of judicial documents.

(3) Broader Application. The knowledge graph for annotation encapsulates significant relationships among various elements. For instance, since it is common for a single Chinese judicial document to involve multiple defendants, crimes, and victims, our team of legal experts has linked defendant and victim characteristics to their respective individuals and affiliated crime characteristics to the corresponding offenses. This integration of crucial interrelations among labels enhances performance in various downstream applications, including the prediction of a specific defendant's crime and sentencing, and also expands LEEC's applicability in real-world court settings and future empirical research.

To validate the quality and applicability of LEEC, we implement various DEE models and LLMs and evaluate them on LEEC, which shows that these models exhibit insufficient accuracy in extracting elements from the LEEC dataset. Moreover, our preliminary empirical analysis based on LEEC uncovers several defendant demographic characteristics that significantly impact sentencing decisions, highlighting potential judicial unfairness in Chinese criminal trials. It is shown that this dataset has the potential to facilitate future empirical research and other downstream applications, enabling the identification and resolution of judicial unfairness and contributing to social justice.

2 Interview

To dig deeper into the value of a labeled legal dataset and how to build it, we interviewed 11 Chinese legal researchers, 3 officers in two legal aid agencies in Jiangsu Province and Shanghai, and one lawyer, who occupies a leadership position in the Shanghai Bar Association.

Among the 11 legal researchers, none of them has ever utilized legal resource research. This confirmed the severe limitations of these works as discussed in Introduction. However, they all speculated that at least some extra-legal factors, including demographic characteristics, may influence criminal sentencing. For example, one researcher witnessed a judge commenting on a female offender:"How could this happen to a little girl? She must have suffered a lot." Therefore, that researcher speculated that female offender may be treated more leniently sometimes as they may be perceived as vulnerable. Naturally, over 80% of them (9 in 11) welcome more high-quality empirical analyses to discover the potential sentencing disparities and contribute to judicial fairness.

Seven out of the nine researchers asked about the current situation of Chinese empirical legal studies believe that the proportion of these studies remains low. Several barriers were identified: 1) As Chinese legal education does not typically involve methods for empirical legal research, most legal researchers lack basic knowledge regarding such research; 2) All researchers identified significant barriers in processing and extracting labels from large judicial text data. For example, one researcher mentioned that a scholar may need to spend two to three months to extract labels for such analysis. Meanwhile, building algorithms like regular expression matching require considerable time and often fail to achieve high accuracy because of the complexity of judicial documents. Finally, all researchers confirm the value of a labeled judicial dataset with high quality based on criminal judicial documents, which could facilitate empirical legal research and discover valuable patterns regarding "law in action".

Three officers from legal aid agencies provide unique and valuable insights, highlighting potential judicial unfairness faced by socio-economically disadvantaged defendants. They unanimously affirm that the financial compensation provided to attorneys representing legal aid defendants is notably low. Concurrently, both agencies typically allocate legal aid cases to attorneys who have actively expressed a willingness to han-

dle such cases. This indicated that legal aid cases may be disproportionately handled by inexperienced attorneys who struggle to secure their own clients. One officer, when questioned about such concern, said: "We have regulations that prohibit treating legal aid cases as a means for training or gaining experience for attorneys as every case is important. However, we cannot control attorneys' minds." The interview with these officers underpins the necessity of conducting more empirical analyses regarding judicial unfairness.

The lawyer we interviewed claimed to have handled over 2,000 cases throughout her legal career. She confirmed that as she became a more sophisticated lawyer, she gradually ceased handling legal aid cases. By the time of the interview, she had neither read nor heard of anyone in the judicial practice citing or mentioning any empirical legal research or legal resource studies. This clearly highlights the dearth and limitations of these studies in China, which led to its lack of visibility in judicial practice. Based on her extensive experience, she believes that judges are influenced by a wide range of extra-legal factors, including the defendant's gender, ethnicity, age, employment status, etc. Judges often consider these factors comprehensively and may even emotionally resonate with the defendant in some cases. She welcomes insightful legal research regarding this issue.

Based on our interview, we confirm the value of a meticulously annotated legal dataset with a more extensive label system that covers both legal and extra-legal factors. Such a dataset could significantly facilitate empirical legal research and uncover the potential influence of extra-legal factors in Chinese criminal trials, thereby enhancing judicial fairness and social justice in China.

3 Data Analysis

This section summarizes the preprocessing of LEEC and discusses its detailed distribution in important dimensions.

3.1 Corpus and Preprocessing

Our case selection is based on the publicly available LEVEN [Yao *et al.*, 2022] and LeCaRD [Ma *et al.*, 2021] datasets. After deduplication, the number of total unique cases is reduced to 15,919 cases from a total of 17,352 cases. All documents within this collection represent criminal cases filed between 2001 and 2021. Among the complete dataset, we preserve 689 judicial documents as unpublished for future evaluation, and publish the annotated data from 15,230 judicial documents. We compare LEEC with two types of datasets: (1) General-domain ED datasets. Compared with ACE2005 [Grishman *et al.*, 2005] and MAVEN [Wang *et al.*, 2020], LEEC's label system is specifically designed for legal texts. Moreover, the judicial documents of LEEC were annotated by selected law school students with a deep understanding of legal knowledge and concepts. (2) Legal-domain datasets. Compared with LEVEN and LeCaRD, our dataset offers a comprehensive expansion of the label system with a finer level of granularity, encapsulating both legal and extra-legal labels. Furthermore, LEEC connects different characteristics to their corresponding entities, such as victims, defendants, and crimes. This methodology significantly enhances the dataset's utility, thereby substantially benefiting downstream applications and empirical research.

3.2 Data Distribution

Our dataset reveals the presence of multiple defendants in 33% of cases, multiple victims in 19% of cases, 49% of cases that contain at least one victim, and multiple crimes in 41% of cases.[5] This underscores the necessity and effectiveness of introducing a sophisticated, domain-specific label system to handle such complexity. The distribution of cases is displayed in Table 1. We measure the data quality by Kappa, with a value of 0.71. Specifically, 3,990 documents were annotated with mentions of elements, trigger words, and values for each element, while the remaining 11,240 documents underwent extended annotation without element mentions and trigger words. Detailed annotation process of the dataset can be found in Appendix B, and samples of the Annotation Guideline are available on our *GitHub repository*.

Case Characteristics	**Case Distribution**					
Defendant number	1	2	3	4	5	>5
	10162	2167	1122	600	378	801
Victim number	0	1	2	3	4	>4
(Missing Value=1297)	6483	4510	919	547	320	1154
Crime number	1	2	3	4	5	>5
	8936	2836	1258	705	428	1067

Table 1: Case distribution on defendants, victims, and crimes.

4 Label System

This section encompasses the compilation of the extensive labels and the crucial relationships among them.

4.1 Label Compilation

Our team of legal experts, led by law professors, incorporated a wide range of legal and extra-legal elements to build a comprehensive knowledge graph covering key elements within the Chinese legal domain. First, our team of legal experts compiled the crucial legal circumstances and factors stipulated by Chinese criminal law and judicial interpretations[6], such as whether the defendant confessed, pled guilty, voluntarily surrendered, conducted a justifiable defense, etc. Furthermore, It has widely been revealed that extra-legal factors may significantly impact judicial practice. Therefore, we utilized elements and theories developed and validated by empirical legal research to comprehensively capture the important factors in Chinese criminal trials.

Our team of legal experts systematically compiled 178 quantitative legal studies from 2018 to 2022, published across 22 journals in Chinese in the China Legal Science Citation Index (CLSCI). The CLSCI is curated by the Law Institute of

[5] A minor fraction of the victim number contains missing values due to instances where the precise number of victims cannot be ascertained. For a detailed explanation, please see Appendix B.

[6] Judicial interpretations of the Supreme People's Court have binding effects for courts of lower levels in China.

China Law Society (LICLS), which provides a list of core legal journals in China.[7] As published judicial documents are among the most commonly used data for empirical analyses on sentencing factors, the labels, theories, and results of these studies serve as valuable sources of potentially salient factors influencing judicial decisions. In addition, we drew upon a wide range of empirical legal studies published in SSCI journals, particularly those in Chinese contexts. Our team meticulously collected the core theories and labels used in these studies and incorporated them into our legal system.

For instance, the Group Threat Theory suggests that when majority groups feel threatened by minority populations, criminal justice systems may treat racial or ethnic minorities adversely [Ulmer and Johnson, 2004]. This theory has been validated and developed in the Chinese context by prior empirical research, which found that minorities perceived as "problem minorities" that might disrupt public order may face discrimination in Chinese criminal cases [Hou and Truex, 2022]. Therefore, we included the ethnic status of offenders in our knowledge graph. Moreover, the Focal Concerns Theory highlights four crucial factors influencing sentencing decisions: the defendant's culpability, redeemability, the risk posed to the community, and pragmatic considerations such as the court's workload [Ulmer *et al.*, 2023]. Research in Chinese contexts has shown that, in line with the Focal Concern Theory, the defendant's being a rural-to-urban migrant – measured by the registered permanent residence (*Hukou*) – significantly impacts sentencing outcomes [Jiang and Kuang, 2018]. As a result, we also included the *Hukou* information of defendants in our knowledge graph. Following this scheme, we effectively constructed an extended, multi-level knowledge graph to cover 155 important elements – both legal and extra-legal – in Chinese criminal sentencing. The elements in the knowledge graph are divided into four main categories: defendant characteristics, victim characteristics, case characteristics, and crime characteristics. [8]

4.2 Relation Construction

We integrated the relationships between elements into the knowledge graph, recognizing their significant impact on judicial decisions, as each defendant or victim in the document may have unique circumstances and characteristics. Therefore, all characteristics pertaining to a victim or defendant are linked directly to each individual. Besides, as a defendant may have over one defenders in Chinese criminal trials, the defender characteristics are connected to each individual defender of a specific defendant. Furthermore, as each defendant could be sentenced for multiple crimes in Chinese judicial documents, all characteristics of a crime are connected to the specific crime committed by a particular defendant.[9] The elements within the knowledge graph are depicted in Appendix D. For details regarding the annotation process based on this label system, please see Appendix B.

5 Experiments

In this section, we conduct experiments on representative Document-level Event Extraction (DEE) models and Large Language Models (LLMs). We then discuss the challenges identified in legal element extraction.

5.1 DEE Models

Experiment Settings

For traditional DEE models, we selected 21 labels about defendants' characteristics and sentencing in LEEC label system, which are shown in Table A1 in the Appendix. We used the LEEC dataset with triggers as the experimental data. For each label, the annotation content or trigger word corresponding to the original text was regarded as the entity. The dataset was split into the train, dev, and test sets at a ratio of 8:1:1. We used the same vocabulary as [Zheng *et al.*, 2019] and randomly initialized all the embeddings where dh=768 and dl=32. We employed the Adam optimizer with the learning rate 5e-4 and the batch size is 16. All models were trained for 100 epochs and the checkpoints with the best F1 scores on the dev set were selected for evaluation on the test set.

Baselines and Metrics

Baselines. We introduce the following models as baselines: 1) DCFEE [Yang *et al.*, 2018] is the first model that introduced Distance Supervision (DS) into the DEE task. there are two variants included: DCFEE-O only extracts one event record from one document while DCFEE-M tries to extract multiple possible event records; 2) Doc2EDAG [Zheng *et al.*, 2019] is an end-to-end DEE model that constructs event records in an auto-regressive way by generating entity-based Directed Acyclic Graphs (DAGs); 3) GreedyDec is a baseline proposed in Doc2EDAG [Zheng *et al.*, 2019] which fills one event table greedily; 4) PTPCG [Zhu *et al.*, 2021] is a lightweight model for end-to-end DEE task based on pruned complete graphs with pseudo triggers.

Metrics. We follow the same evaluation setting in the previous studies [Zheng *et al.*, 2019; Zhu *et al.*, 2021; Peng *et al.*, 2023]. For each prediction record, we select a golden record by matching records with the same defendant name and the most shared arguments, and calculate the F1 score by comparing the parameters between them.

[7] For details, please visit https://fzyjs.chinalaw.org.cn, accessed on May 2nd, 2024.

[8] Some elements, such as court name, judge name, case title, and year of judgment, can be easily and accurately extracted from Chinese verdicts using keyword identification or regular expression matching. For useful references, please visit *this GitHub program*. These elements do not require manual annotation, and thus, are not included in the knowledge graph of this study.

[9] Specifically, it is noteworthy that in Chinese criminal cases where an individual defendant committed multiple crimes, the court typically adjudicates a sentence for each individual crime, followed by an overall aggregated sentence. This final sentence, which is usually subject to a certain degree of the judge's discretion, may not necessarily align with the sum of the individual sentences. Consequently, in our knowledge graph, we deliberately included both the sentencing elements, linked to each distinct crime of a specific defendant, and the final, aggregated sentence, linked to each defendant.

Results

Table 2 shows the experimental results. we have the following observations: 1) Some baselines cannot converge well on LEEC, such as Doc2EDAG and similar structured Greedy-Dec. One reason is that legal documents are longer and the arguments are more dispersed, which is not conducive to Doc2EDAG's sequential path extension method for reasoning. 2) The DEE models can only extract the words in the document. For example, for the sentence of imprisonment, it can only extract "fixed-term imprisonment", indirectly deriving yes or no. This two-step approach does not sufficiently meet the requirements in real-world applications. 3) We selected 21 relatively simple elements for the DEE task. However, the LEEC label system includes labels that are either sparse[10] or in need of complex judgment and high-level reasoning capabilities[11]. Developing DEE models to extract these labels may present a greater challenge.

Model	Precision	Recall	F1 score
DCFEE-O	62.98	**83.13**	71.67
DCFEE-M	59.56	82.54	69.19
Greedy-Dec	80.27	56.00	65.97
Doc2EDAG	42.73	70.01	53.07
PTPCG	**86.82**	77.99	**82.17**

Table 2: Overall performance of DEE models

5.2 Large Language Models

Experiment Settings

We selected some advanced general LLMs and legal LLMs for our experiments.

General LLMs: 1) GPT-3.5[12]: An advanced LLM by OpenAI that excels in understanding and generating text; 2) ChatGLM3 [Zeng *et al.*, 2022]: A bilingual open-source LLM for the general domain. We used GLM3-6B-32K as our baseline for its larger context lengthl; 3) LLaMA3[13]: Meta's SOTA open-source LLM. For Chinese documents, we used Llama3-Chinese-8B-Instruct[14].

Legal LLMs: 1) Lawyer-LLaMA [Huang *et al.*, 2023]: A Chinese legal LLM based on LLaMA [Touvron *et al.*, 2023]. The model without a retrieval module was used in this experiment; 2) Tongyi Farui[15]: A legal LLM launched by Aliyun, capable of performing various legal tasks such as answering legal questions, assisting in case analysis, and generating legal documents.

LLM models make up for some inherent limitations of DEE. Compared to traditional DEE models that can only extract triggers of labels, LLMs possess more advanced reasoning abilities. Consequently, the outputs of LLMs could be closer to human annotation results.[16] Therefore, we provide prompts to LLMs asking them to produce final annotation results in the format of LEEC json data. The dataset used is the same as in the DEE task. For evaluation, we compare the accuracy of the predicted and annotated results of humans.

Results

Table 3 shows the experiment statistics and the accuracy of LLMs. Table 4 provides an example of how LLMs extract judgment documents. Although LLMs make up for some of the inherent shortcomings of DEE models, there are still many challenges as shown in Tables 3 and 4:

Model	Max length	Truncation	Unformatted output	Accuracy
GPT-3.5	16385	12.71%	0.13%	0.7070
GLM3-32k	32768	1.48%	4.49%	0.4920
LLaMA3	8192	20.70%	7.17%	0.6392
Lawyer-LLaMA	2048	80.20%	*	*
Tongyi Farui	12000	12.41%	0.08%	0.6456

Table 3: Overall performance of LLMs. For each model, Max length is the maximum context length, Truncation is the proportion of truncated data. Unformatted output is the proportion of the output not in the given format. Since most responses in Lawyer-LLaMA outputs don't follow the correct format, its result isn't included.

(1) Truncated Input and Unformatted Output. One notable discovery is that the limited context window of current LLMs leads to significant problems when analyzing the often lengthy judicial documents. Moreover, despite the instructions provided, it was found in all LLMs evaluated that in some cases the output could not be generated in the expected format. These issues include repeated outputs, incomplete responses, irrelevant answers, and other similar circumstances. Lawyer-LLaMA is the one that deviates the most from the correct format. This may be due to the small Chinese corpus, the limited context window, and a degraded general understanding ability after fine-tuning in the legal domain.

(2) Incomplete Defendant Coverage. LLMs may overlook certain defendants in element extraction. Except Lawyer-LLaMA, both GPT and GLM only extracted one of the three defendants. This shortcoming is important for legal research as the absence of defendants can significantly impact the nature and analysis of the case.

(3) Erroneous Legal Reasoning. LLMs excel at extracting basic labels such as "Name" and "Birth", but there are still

[10]Some elements do not occur frequently in Chinese criminal trials, such as the Not_guilty element or the Excluding_evidence_decision element. Extracting such labels may be more difficult for DEE models as they rely heavily on the quantity of effective training data.

[11]One such label is the Joint_crime element. The corresponding annotation guideline is available in Appendix B.2.

[12]https://openai.com, accessed on April 13th, 2024.

[13]https://github.com/LlamaFamily/Llama-Chinese, accessed on April 13th, 2024.

[14]https://llama.meta.com/docs/model-cards-and-prompt-formats/meta-llama-3, accessed on April 13th, 2024.

[15]https://tongyi.aliyun.com, accessed on April 13th, 2024.

[16]For example, LLMs can directly judge whether the sentencing of "Imprisonment" is delivered by the court with 0/1, and convert the sentencing length into numeric months, rather than only extracting the trigger as traditional DEE models do.

Prompt	**Please extract the defendants' information from the following judgment documents. If a label doesn't exist or can't be extracted, fill in the empty string "". For 0/1 labels such as "Control", if the judgment is yes, then fill in 1, otherwise fill in 0. For labels like [ControlTime (months)], you just need to fill in the number. Output example (in the case of one defendant, if there are multiple defendants, please select Defendant 1, Defendant 2...) :** "Defendant 1" : "Name" : "", "Gender" : "", "Birth" : "", "Place" : "", "Control" : "", "ControlTime (months)" : "", "Detention" : "", "DetentionTime (month)" : "", "Imprisonment" : "", "ImprisonmentTime (month)" : "", "PoliticalRights" : "", "PoliticalRightsTime (month/life)" : "", "Fine" : "", "FineNum (yuan)" : "", "PartofProperty" : "", "PartofPropertyNum (yuan)" : "", "AllProperty" : "", "AllPropertyNum (yuan)" : "", "EcoCompensation" : "", "EcoCompensationNum (yuan)" : ""
Input	"Judgment documents: ....Defendant Xie, male, born on September 27, 1992, Han nationality, farmer, civilian, junior high school education, registered in Gaoyang, Hebei Province ... Xie collided with Zhang 1, who was standing on the road, and then collided with a small van of Hebei Fxxxxx parked on the road, causing damage to both vehicles and the death of Zhang 1. After the accident, Ge called Fu to drive to the scene of the accident and pull Xie away to help him escape. Xie takes full responsibility for this accident, and Zhang 1 and Zhang 2 have no responsibility. **After the accident, the relatives of the defendant Xie took the initiative to compensate the economic loss of the victim Zhang 1 family of 460,000 yuan, and obtained the understanding of the victim's family**.... The public prosecution organ provided the corresponding evidence that the **defendant Xie**'s behavior violated the provisions of Article 133 of the Criminal Law of the People's Republic of China, and he should be investigated for the crime of causing traffic accidents. The actions of the defendants, **Ge and Fu**, violated the provisions of Article 310 of the Criminal Law of the People's Republic of China and should be investigated for criminal responsibility for harboring. The defendant Xie fled and surrendered himself, so it is recommended that he be sentenced to between three and five years in prison. The defendant Ge, Fu surrendered the plot, please the court on the defendant Ge, Fu's judgment according to law. The defendants Xie, Ge and Fu have no objection to the criminal facts, charges and sentencing recommendations charged by the public prosecution. ... By the provisions of Article 133, Article 301, Article 67, paragraph 1, and Article 72, paragraph 1 of the Criminal Law of the People's Republic of China, the verdict is as follows: 1. **The defendant Xie, who committed the crime of causing a traffic accident, is sentenced to fixed-term imprisonment of three years,** with a suspension of four years...."

LEEC	DEE Model	GPT-3.5	GLM3-32k	Lawyer-LLaMA

{**"Defendant 1"** : {
"Name" : "Xie",
"Gender" : "Male",
"Birth" : "1992/9/27",
"Nation" : "Han",
"Place" : "Gaoyang, Hebei province",
"Control" : "0",
"ControlTime (months)" : "",
"Detention" : "0",
"DetentionTime (month)" : "",
"Imprisonment" : "1",
"ImprisonmentTime (month)" : "36",
"PoliticalRights" : "0",
"PoliticalRightsTime (month/life)" : "",
"Fine" : "0",
"FineNum (yuan)" : "",
"PartofProperty" : "0",
"PartofPropertyNum (yuan)" : "",
"AllProperty" : "0",
"AllPropertyNum (yuan)" : "",
"EcoCompensation" : "0",
"EcoCompensationNum (yuan)" : "",
}
"Defendant 2" : {
"Name" : "Ge",
...}
"Defendant 3" : {
"Name" : "Fu",
...}}
	{**"Defendant 1"** : {
	"Name" : "Xie",
	"Gender" : "Male",
	"Birth" : "September 27, 1992",
	"Nation" : "Han",
	"Place" : "Gaoyang, Hebei Province",
	"Control" : "",
	"ControlTime (months)" : "",
	"Detention" : "",
	"DetentionTime (month)" : "",
	"I..." : "fixed-term imprisonment",
	"I...Time (month)" : "three years",
	"PoliticalRights" : "",
	"PoliticalRightsTime (month/life)" : "",
	"Fine" : "",
	"FineNum (yuan)" : "",
	"PartofProperty" : "",
	"PartofPropertyNum (yuan)" : "",
	"AllProperty" : "",
	"AllPropertyNum (yuan)" : "",
	"EcoCompensation" : "",
	"EcoCompensationNum (yuan)" : "",
	}
	"Defendant 2" : {
	"Name" : "Ge",
	...}
	"Defendant 3" : {
	"Name" : "Fu",
	...}}
		"Defendant 1" : {
		"Name" : "Xie",
		"Gender" : "Male",
		"Birth" : "1992/9/27",
		"Nation" : "Han",
		"Place" : "Gaoyang, Hebei province",
		"Control" : "0",
		"ControlTime (months)" : "",
		"Detention" : "",
		"DetentionTime (month)" : "",
		"Imprisonment" : "1",
		"ImprisonmentTime (month)" : "36",
		"PoliticalRights" : "",
		"PoliticalRightsTime (month/life)" : "",
		"Fine" : "",
		"FineNum (yuan)" : "",
		"PartofProperty" : "",
		"PartofPropertyNum (yuan)" : "",
		"AllProperty" : "",
		"AllPropertyNum (yuan)" : "",
		"EcoCompensation" : "1",
		"EcoCom...Num (yuan)" : "460000"
		}}
			"Defendant 1" : {
			"Name" : "Xie",
			"Gender" : "Male",
			"Birth" : "1992/9/27",
			"Nation" : "Han",
			"Place" : "Gaoyang, Hebei province",
			"Control" : "0",
			"ControlTime (months)" : "",
			"Detention" : "0",
			"DetentionTime (month)" : "",
			"Imprisonment" : "0",
			"ImprisonmentTime (month)" : "",
			"PoliticalRights" : "0",
			"PoliticalRightsTime (month/life)" : "",
			"Fine" : "0",
			"FineNum (yuan)" : "",
			"PartofProperty" : "0",
			"PartofPropertyNum (yuan)" : "",
			"AllProperty" : "0",
			"AllPropertyNum (yuan)" : "",
			"EcoCompensation" : "0",
			"EcoCompensationNum (yuan)" : ""
			}
				According to your description, I have extracted the information of the **defendant Xie** for you. The defendant Xie, male, Han nationality, farmer, civilian, junior high school education, registered in Gaoyang County, Hebei Province, now lives in Gaoyang County. ... Defendant Xie's behavior violated the provisions of Article 133 of the Criminal Law, and he should be investigated for criminal responsibility for the crime of causing a traffic accident. Notably, defendant Xie surrendered to the plot, **compensated the victim's economic losses, and requested the court for a lighter punishment.**In addition, you mentioned that the **defendant Ge, Fu** also surrendered, please the court on the defendant Ge, Fu's judgment by the law. Finally, it is recommended that you provide more facts and evidence to better analyze the situation.

Table 4: An example of extraction. For LLMs, the input message is *[prompt, input]*, and then it is truncated according to the context length of each model. *LEEC* is the manual annotation in LEEC dataset. *DEE Model* is the golden record based on LEEC for traditional DEE models, *GPT-3.5, GLM3-32K and Lawyer-LLaMA* represents the output of LLMs. Due to page limitations, We illustrate only representative outputs, omitting some unimportant information and the outputs of LLaMA3 and Tongyi Farui.

challenges in extracting labels that require the understanding and knowledge of legal concepts, which make it hard to meet application requirements. For instance, GPT correctly identified Imprisonment labels, but mistakenly categorized the voluntary compensation by Xie's family as "FineNum", the amount of fine in the court ruling. GLM incorrectly identified *0* for all sentencing labels. Lawyer-LLaMA observed that Defendant Xie had compensated the victim's family for economic losses, but it failed to distinguish it from the judge's ruling.

Overall, current LLMs cannot achieve the accuracy required for empirical research without introducing substantial bias and errors.[17] The LEEC dataset annotated by legal experts can serve as a benchmark to identify existing challenges, evaluate extraction accuracy, and function as the training data for LLMs so that they can be more effectively applied to downstream applications in the future.

6 Empirical Analysis

Utilizing our LEEC dataset, we conducted empirical legal analyses with a threefold objective: 1) to verify the suitability and applicability of the LEEC dataset for empirical analysis; 2) to determine whether it yields patterns that are reasonable or coherent with related findings from prior empirical studies, thereby attesting to the quality and robustness of this dataset; 3) to provide preliminary evidence regarding the judicial (in)equality within Chinese criminal trials.

Stratification and inequality in criminal sentencing have garnered considerable attention from scholars across the social sciences [Ulmer, 2012]. To investigate such issues in Chinese contexts, we aim to explore the sentencing impact of defendant demographic characteristics, including gender, ethnicity, and age, based on labels in the "Demographic Characteristics" section (refer to Figure A1 in the Appendix) within

[17]It is shown in Table 3 that the highest accuracy of LLMs is a little over 70%, which is far from satisfactory for legal research.

the LEEC dataset. The dependent variable is the length of limited imprisonment. In line with the predominant approach in empirical legal research for investigating causal effects in sentencing [Peng and Cheng, 2022; Liu *et al.*, 2021; Ulmer, 2012], we used the Ordinary Least Squares (OLS) regression model as our methodological tool. For details regarding our regression model, please refer to Appendix C.

Figure 1 presents the forest plot displaying the estimation coefficients of the defendant demographic variables and their respective 95% confidence intervals. Several interesting results emerge. Firstly, we found that female defendants are likely to be sentenced more leniently in our dataset. This finding is consistent with a series of studies in Western jurisdictions [Embry and Lyons Jr, 2012; Fernando Rodriguez *et al.*, 2006]. These studies provide evidence supporting the chivalry hypothesis, suggesting that due to gender patriarchy, women may be perceived as vulnerable, less blameworthy, and in need of extra protection in criminal sentencing. Secondly, we discovered that as defendants age, their sentences may become more lenient, aligning with prior studies in the U.S. [Ryon *et al.*, 2017; Steffensmeier *et al.*, 1995]. Thirdly, unemployed defendants tend to receive harsher sentences, illustrating a concern for Chinese courts to maintain social stability [Trevaskes *et al.*, 2014]. Additionally, those who are less educated are likely to be sentenced more leniently, possibly because the courts may perceive these defendants as disadvantaged and less blameworthy for their wrongdoings.

Overall, our empirical analysis reveals multiple extra-legal factors that may contribute to sentencing disparities and judicial unfairness in Chinese criminal trials. However, it should be noted that our investigation is exploratory and preliminary. Deeply investigating the impact of each of these defendant demographic characteristics may require conducting individual studies or even a series of studies in the social sciences, with detailed theoretical construction, robustness tests, further analyses, etc., far beyond the scope of this paper. Nevertheless, our results provide reasonable and interesting findings that validate the applicability and quality of the LEEC dataset, while offering insightful direction and evidence for future researchers' deeper and broader investigations.

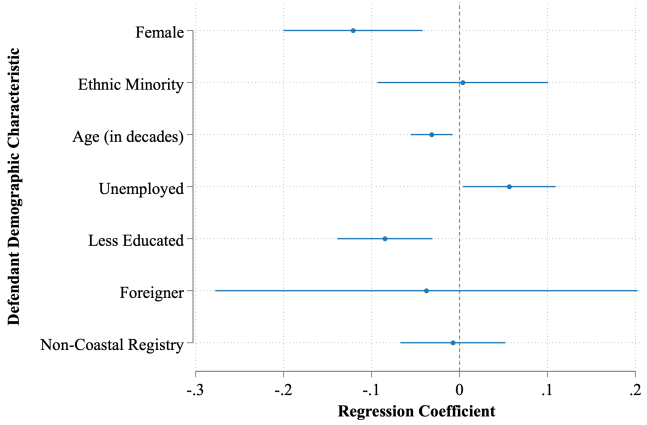

Figure 1: The impact of defendant demographic characteristics on the length of fixed-term imprisonment.

7 Discussion and Conclusion

In this study, we introduce LEEC, a unique dataset designed for legal elements extraction in Chinese legal system. LEEC stands out because its label system is enriched with both legal and extra-legal labels, integrating crucial legal knowledge drawn from Chinese law, empirical legal studies, our interview, and legal experts' understanding of Chinese legal contexts and practices. Each of the 15,919 cases in the dataset is annotated by law school students. Experimental results underline the challenges for traditional models and LLMs in element extraction and the biases in Chinese sentencing, signifying areas of focus for future research.

This study has several limitations that we hope will be addressed by future research: 1) Potential Selection Bias: About 75% of all judicial verdicts in China have been disclosed for cases not processed through mediation in recent years [Tang and Liu, 2019]. Consequently, cases in LEEC may not fully represent those in actual courts due to potential selection bias. Future studies should consider this when utilizing LEEC; 2) Context-Specificity: The data, label system, and annotation methods are inherently embedded in Chinese contexts. However, although laws vary across jurisdictions, the underlying logic, concepts, and biases of the criminal system have much in common in modern societies. Therefore, we believe that the majority of the labels and methods of data curation could serve as useful references for future legal resource research in other jurisdictions; 3) Biases of LLMs: We identified several types of biases in Chinese real criminal sentencing using LLEC, which are typically embedded in social structure, stratification, and ideologies. Whether the biases we identified in LLMs are randomly distributed or similar to those in the real world requires further examination to ensure that AI could better serve judicial fairness and social good.

Ethical Statement

To support downstream applications, the knowledge graph and annotation in this study exhibit high granularity. It is of paramount importance that users of LEEC exercise due caution. We strongly oppose the use of LEEC for any purposes that could lead to discrimination or violations of the rule of law. The personal information included in the published judicial documents was collected and processed in strict compliance with Chinese law. Any future utilization of LEEC must also adhere to applicable laws and commit to responsible, ethical handling of the data.

Acknowledgments

We extend our deep gratitude to Huaiyu Hu, Senyu Li, Bifan Zhao, Jin LI, Shuchen Tang, Yige Fan, Zhiwei Zhang, Caixuan Huang, Dikun Zhu, Kun Liu, and other anonymous interviewees for their insightful opinions and invaluable help. We also wish to express our sincere appreciation to Zexia Yang, Huihan Li, and Zhijie He for their dedicated assistance. This work is supported by the National Key Research and Development Program of China (No.2022YFC3301504).

Contribution Statement

Zongyue Xue, Huanghai Liu, and Yiran Hu have contributed equally to this work. Zongyue Xue: Methodology, Label System, Annotation Guideline, Experiment, Paper Writing, Organizer; **Huanghai Liu**: Implementation, Experiment; **Yiran Hu**: Methodlogy, Label System, Experiment, Paper Writing, Organizer; **Yuliang Qian**: Data Processing, Experiment; **Yajing Wang**: Experiment; **Kangle Kong**: Annotation Guideline, Organizer; **Chenlu Wang**: Organizer; **Yun Liu**: Superviser, Funding Provider; **Weixing Shen**: Superviser, Funding Provider.

References

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

[Doerner and Demuth, 2010] Jill K Doerner and Stephen Demuth. The independent and joint effects of race/ethnicity, gender, and age on sentencing outcomes in us federal courts. *Justice Quarterly*, 27(1):1–27, 2010.

[Embry and Lyons Jr, 2012] Randa Embry and Phillip M Lyons Jr. Sex-based sentencing: Sentencing discrepancies between male and female sex offenders. *Feminist Criminology*, 7(2):146–162, 2012.

[Feng *et al.*, 2022] Yi Feng, Chuanyi Li, and Vincent Ng. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Fernando Rodriguez *et al.*, 2006] S Fernando Rodriguez, Theodore R Curry, and Gang Lee. Gender differences in criminal sentencing: Do effects vary across violent, property, and drug offenses? *Social Science Quarterly*, 87(2):318–339, 2006.

[Grishman *et al.*, 2005] Ralph Grishman, David Westbrook, and Adam Meyers. Nyu's english ace 2005 system description. *ACE*, 5:2, 2005.

[Guo *et al.*, 2020] Kaihao Guo, Tianpei Jiang, and Haipeng Zhang. Knowledge graph enhanced event extraction in financial documents. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1322–1329. IEEE, 2020.

[Hogenboom *et al.*, 2011] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57, 2011.

[Hou and Truex, 2022] Yue Hou and Rory Truex. Ethnic discrimination in criminal sentencing in china. *The Journal of Politics*, 84(4):2294–2299, 2022.

[Huang *et al.*, 2023] Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*, 2023.

[Jiang and Kuang, 2018] Jize Jiang and Kai Kuang. Hukou status and sentencing in the wake of internal migration: The penalty effect of being rural-to-urban migrants in china. *Law & Policy*, 40(2):196–215, 2018.

[Li *et al.*, 2023a] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. Sailer: Structure-aware pre-trained language model for legal case retrieval, 2023.

[Li *et al.*, 2023b] Haitao Li, Weihang Su, Changyue Wang, Yueyue Wu, Qingyao Ai, and Yiqun Liu. Thuir@coliee 2023: Incorporating structural knowledge into pre-trained language models for legal case retrieval, 2023.

[Li *et al.*, 2023c] Haitao Li, Changyue Wang, Weihang Su, Yueyue Wu, Qingyao Ai, and Yiqun Liu. Thuir@coliee 2023: More parameters and legal knowledge for legal case entailment, 2023.

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

[Liao and Grishman, 2010] Shasha Liao and Ralph Grishman. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 789–797, 2010.

[Liu *et al.*, 2021] Lin Liu, Christy A Visher, and Daniel J O'Connell. Strain during reentry: A test of general strain theory using a sample of adult former prisoners. *The Prison Journal*, 101(4):420–442, 2021.

[Liu *et al.*, 2023] Bulou Liu, Yiran Hu, Yueyue Wu, Yiqun Liu, Fan Zhang, Chenliang Li, Min Zhang, Shaoping Ma, and Weixing Shen. Investigating conversational agent action in legal case retrieval. In *European Conference on Information Retrieval*, pages 622–635. Springer, 2023.

[Ma *et al.*, 2021] Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 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*, SIGIR '21, page 2342–2348, New York, NY, USA, 2021. Association for Computing Machinery.

[Nguyen *et al.*, 2016] Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. A dataset for open event extraction in english. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1939–1943, 2016.

[Peng and Cheng, 2022] Yali Peng and Jinhua Cheng. Ethnic disparity in chinese theft sentencing. *China Review*, 22(3):47–71, 2022.

[Peng *et al.*, 2023] Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu, and

Weixing Shen. The devil is in the details: On the pitfalls of event extraction evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9206–9227, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Pound, 1910] Roscoe Pound. Law in books and law in action. *American Law Review*, 44:12, 1910.

[Richards *et al.*, 2016] Tara N Richards, Wesley G Jennings, M Dwayne Smith, Christine S Sellers, Sondra J Fogel, and Beth Bjerregaard. Explaining the "female victim effect" in capital punishment: An examination of victim sex–specific models of juror sentence decision-making. *Crime & Delinquency*, 62(7):875–898, 2016.

[Ryon *et al.*, 2017] Stephanie Bontrager Ryon, Ted Chiricos, Sonja E Siennick, Kelle Barrick, and William Bales. Sentencing in light of collateral consequences: Does age matter? *Journal of Criminal Justice*, 53:1–11, 2017.

[Shen *et al.*, 2020] Shirong Shen, Guilin Qi, Zhen Li, Sheng Bi, and Lusheng Wang. Hierarchical Chinese legal event extraction via pedal attention mechanism. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 100–113, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[Sierra and others, 2018] G Sierra et al. Event extraction from legal documents in spanish. In *1st Workshop on Language Resources and Technologies for the Legal Knowledge Graph*, page 36, 2018.

[Steffensmeier *et al.*, 1995] Darrell Steffensmeier, John Kramer, and Jeffery Ulmer. Age differences in sentencing. *Justice Quarterly*, 12(3):583–602, 1995.

[Tang and Liu, 2019] Yingmao Tang and John Zhuang Liu. Mass publicity of chinese court decisions. *China Review*, 19(2):15–40, 2019.

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

[Tran *et al.*, 2019] Vu Tran, Minh Le Nguyen, and Ken Satoh. Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 275–282, 2019.

[Trevaskes *et al.*, 2014] Susan Trevaskes, Elisa Nesossi, Flora Sapio, and Sarah Biddulph. *The politics of law and stability in China*. Edward Elgar Publishing, 2014.

[Ulmer and Johnson, 2004] Jeffery T Ulmer and Brian Johnson. Sentencing in context: A multilevel analysis. *Criminology*, 42(1):137–178, 2004.

[Ulmer *et al.*, 2023] Jeffery T Ulmer, Eric Silver, and Lily S Hanrath. Back to basics: A critical examination of the focal concerns framework from the perspective of judges. *Justice Quarterly*, 40(6):813–836, 2023.

[Ulmer, 2012] Jeffery T Ulmer. Recent developments and new directions in sentencing research. *Justice Quarterly*, 29(1):1–40, 2012.

[Veyseh *et al.*, 2022] Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Huu Nguyen. Mee: A novel multilingual event extraction dataset. *arXiv preprint arXiv:2211.05955*, 2022.

[Wang *et al.*, 2020] Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. MAVEN: A massive general domain event detection dataset. In *Proceedings of EMNLP*, pages 1652–1671, 2020.

[Xiao *et al.*, 2019] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. CAIL2019-SCM: A dataset of similar case matching in legal domain. *CoRR*, abs/1911.08962, 2019.

[Yang *et al.*, 2018] Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55, 2018.

[Yao *et al.*, 2022] Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. LEVEN: A large-scale chinese legal event detection dataset. In *Findings of ACL*, pages 183–201, 2022.

[Yao *et al.*, 2023] Feng Yao, Jingyuan Zhang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Yun Liu, and Weixing Shen. 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, 2023.

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

[Zheng *et al.*, 2019] Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. *arXiv preprint arXiv:1904.07535*, 2019.

[Zhu *et al.*, 2021] Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Min Zhang. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. *arXiv preprint arXiv:2112.06013*, 2021.