CMDL: A Large-Scale Chinese Multi-Defendant Legal Judgment Prediction Dataset

Wanhong Huang[1], Yi Feng[1*], Chuanyi Li[1], Honghan Wu[1], Jidong Ge[1†], Vincent Ng[2]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2]Human Language Technology Research Institute, University of Texas at Dallas, USA
hwh@smail.nju.edu.cn, {fy, lcy}@nju.edu.cn

Abstract

Legal Judgment Prediction (LJP) has attracted significant attention in recent years. However, previous studies have primarily focused on cases involving only a single defendant, skipping multi-defendant cases due to complexity and difficulty. To advance research, we introduce CMDL, a large-scale real-world **C**hinese **M**ulti-**D**efendant **L**JP dataset, which consists of over 393,945 cases with nearly 1.2 million defendants in total. For performance evaluation, we propose case-level evaluation metrics dedicated for the multi-defendant scenario. Experimental results on CMDL show existing SOTA approaches demonstrate weakness when applied to cases involving multiple defendants. We highlight several challenges that require attention and resolution. The dataset is available at https://github.com/littlebowlnju/CMDL.

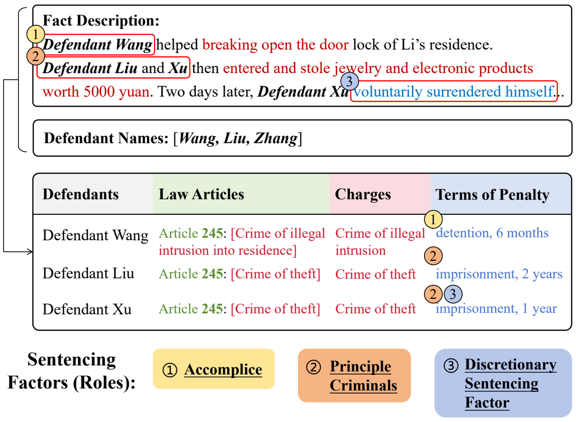

Fact Description:
① Defendant Wang helped breaking open the door lock of Li's residence.
② Defendant Liu and Xu then entered and stole jewelry and electronic products worth 5000 yuan. Two days later, Defendant Xu ③ voluntarily surrendered himself...
Defendant Names: [Wang, Liu, Zhang]
Defendants
Law Articles
Charges
Terms of Penalty
Defendant Wang
Article 245: [Crime of illegal intrusion into residence]
Crime of illegal intrusion
① detention, 6 months
Defendant Liu
Article 245: [Crime of theft]
Crime of theft
② imprisonment, 2 years
Defendant Xu
Article 245: [Crime of theft]
Crime of theft
② ③ imprisonment, 1 year
Sentencing Factors (Roles):
① Accomplice
② Principle Criminals
③ Discretionary Sentencing Factor

Figure 1: An illustration of multi-defendant LJP tasks. The numerical labels in Fact Description indicate the elements exhibiting related sentencing factors. Due to the different criminal roles of defendants in the joint crime and the discretionary sentencing factors, the corresponding outcomes vary.

1 Introduction

Legal Judgment Prediction (LJP) aims to predict judgment outcomes (e.g., law articles, charges) given the fact description of a case. LJP has been widely studied in various jurisdictions and languages (Liu et al., 2023; Jacob de Menezes-Neto and Clementino, 2022; Medvedeva et al., 2023; Valvoda et al., 2023; Feng et al., 2022a). English LJP (Chalkidis et al., 2019) focuses on violation prediction while Swiss LJP focuses on predicting plaintiff's claims (Semo et al., 2022; Niklaus et al., 2021). Chinese LJP primarily focuses on predicting charges, applicable law articles and terms of penalty in criminal cases. LJP can help both legal professionals in analyzing cases and nonprofessionals with low-cost consulting services.

In this paper, we focus on Chinese LJP. A variety of methods have been proposed to solve Chinese LJP. Some improve performance by exploring the relationships between subtasks and employ multi-task learning (Zhong et al., 2018; Yao et al., 2020; Huang et al., 2021). Some extract key information from facts for prediction, enhancing both accuracy and interpretability (Feng et al., 2022b; Wu et al., 2022). The importance of distinguishing between similar charges and legal provisions has also attracted attention (Zhang et al., 2023; Liu et al., 2022). With the development of pre-trained models, researchers begin to propose LJP solutions based on these models and achieve new state-of-the-art results (Xiao et al., 2021; Huang et al., 2021).

Nevertheless, existing methods are mostly designed for and tested on single-defendant cases, leaving out multi-defendant cases due to complexity (Xiao et al., 2018). While many researchers highlight the significance of exploring multi-defendant cases in future studies (Li et al., 2021; Xu et al., 2020), it remains a relatively underexplored area with limited satisfactory solutions.

Given the fact description of a case, multi-defendant LJP aims to predict judgment outcomes

[*] Corresponding author.
[†] Co-corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 5895–5906
August 11-16, 2024 ©2024 Association for Computational Linguistics

for each defendant. It's nontrivial to solve multi-defendant LJP. As shown in Figure 1, three defendants involved in a single case receive different judgments as they face different sentencing factors based on judicial theory (Liu, 2020). For instance, *Defendant Wang* involves in the case as the role of *accomplice* that does not directly participate in the theft activity, whereas *Defendant Liu* and *Xu* are *principle criminals* for stealing. Thus, *Defendant Wang* gets a lighter punishment compared to the others. However, there are several challenges for models to distinguish sentencing factors (e.g. identifying the roles in joint crimes) among defendants as follows.

Segmenting facts. Legal documents do not follow a sequential pattern where each defendant is individually described in facts. Instead, facts are typically presented in chronological order, with details concerning each defendant interspersed and not separated. Moreover, extracting details pertinent to each defendant solely by their names is impractical, as it may result in disjointed fragments.

Injecting relevant legal knowledge. Judging cases with multi-defendants necessitates adherence to specific legal logic. Accordingly, models should possess the ability to analyze facts and integrate applicable legal knowledge. As shown in Figure 1, while *Defendant Liu* and *Xu* are both principal criminals, they receive different penalties due to *Xu*'s mitigating circumstance of voluntary surrender. While there are numerous single-defendant LJP approaches, they are not applicable in the multi-defendant setting as they lack the necessary legal knowledge required for accurately predicting multi-defendant cases.

To facilitate multi-defendant LJP research, we present CMDL, a large-scale **C**hinese **M**ulti-**D**efendant **L**JP dataset. CMDL consists of 393,945 criminal cases involving multiple defendants, covering 321 charges and 275 law articles. Compared with existing multi-defendant LJP datasets (Lyu et al., 2023; Pan et al., 2019), CMDL contains over ten times the number of cases and the number of labels than others. CMDL also contains more comprehensive legal annotations, including (1) detailed fines (either specific amounts or being fined without a specified amount), (2) diverse term of penalty (e.g., imprisonment, surveillance, etc.) and (3) more fine-grained paragraph-level law articles. For evaluating models' performance under the multi-defendant setting, we propose case-level metrics, which will be described in 4.2.1. We conduct extensive experiments to uncover the challenges of multi-defendant LJP, and experimental results show multi-defendant LJP is still far from being resolved and requires further investigation.

In summary, our contributions are three-fold:

A large-scale Multi-Defendant LJP dataset. To the best of our knowledge, CMDL is the largest Chinese multi-defendant LJP dataset covering the widest range of charges, law articles and penalties. It can also be easily expanded and updated with our automatic annotation process. We believe our dataset can generate more interest in multi-defendant LJP and promote further research in the field of LJP.

A more comprehensive approach to evaluate model performance on multi-defendant cases. Given the varying number of defendants in multi-defendant cases, we believe it's important to assess model performance not only at the individual defendant level (by taking each defendant as a sample) but also from a holistic perspective of the case.

An empirical study on multi-defendant LJP. We implement and evaluate several SOTA models on CMDL. From the results of extensive experiments, we gain preliminary insights into the performance of existing models on multi-defendant cases and identify challenges. We hope that our findings can inspire further research in this area.

2 Related Work

LJP as an important application of AI in the legal field has been extensively studied. Researches in other languages and jurisdictions, especially in countries employing the common law system, typically address binary classification tasks such as predicting violation (Chalkidis et al., 2019) or whether the plaintiff's claims will be accepted/rejected (Malik et al., 2021; Semo et al., 2022; Niklaus et al., 2021). To the best of our knowledge, no LJP research from other countries involves multiple defendants. Therefore, our discussion here mainly focus on the context of Chinese LJP.

In recent years, deep learning and neural networks have emerged as the primary methods for solving various LJP subtasks. Researchers focus on improving accuracy and interpretability by extracting key elements to stimulate practical judicial process (Yue et al., 2021a; Feng et al., 2022b; Wu et al., 2022; Lyu et al., 2022), exploring relationships

	CAIL2018	MultiLJP	CMDL
Case Defendant	Single	Multiple	Multiple
# Cases	2,676,075	23,717	393,945
# Defendants	2,676,075	80,477	1,199,117
# Charges	202	23	321
# Law Articles	183	22	275
fines	✓	×	✓
paragraph-level law	×	×	✓
Non-imprisonment term	×	×	✓

Table 1: Comparison between CMDL and other open-source Chinese LJP datasets.

between subtasks for multi-task learning (Zhong et al., 2018; Yao et al., 2020), exploiting the semantics of article and charge labels (Zhang et al., 2023; Le et al., 2022; Liu et al., 2023) or injecting legal knowledge into networks (Luo et al., 2017; Gan et al., 2021). Pre-trained Language Models (PTMs) pretrained or finetuned on legal data also demonstrate promising performance (Xiao et al., 2021; Huang et al., 2021). With recent surge in Large Language Model (LLM) and generative AI, specialized LLMs targeted at legal field like ChatLaw (Cui et al., 2023) is proposed to solve multiple legal tasks, including LJP.

Basically all existing methods focus on single-defendant LJP using CAIL dataset (Xiao et al., 2018), whereas multi-defendant setting receives limited attention. Pan et al. (2019) use a multi-scale attention model to predict charges for multi-defendant cases with a self-constructed dataset (120 cases in total with only 4 common charge types). Another work predicts judgment results for each defendant using hierarchical reasoning chains, which relies entirely on the additional manually-annotated information (e.g. criminal relationships) in their dataset MultiLJP (Lyu et al., 2023). Unlike the above datasets that either focus on one subtask of LJP or require massive manual efforts, we propose a large-scale (both numbers of cases and charge types are over 10 times of those in MultiLJP) dataset with more comprehensive annotations which are annotated automatically. We compare our datasets with other open-source Chinese LJP datasets and show differences in Table 1.

3 CMDL

In this section, we first describe the definition of multi-defendant LJP task in our dataset. Then we elaborate on the dataset's annotation process and present the data analysis.

3.1 Task Definition

The typical subtasks of Chinese LJP include prediction for related law article, charge and terms of penalty. Specifically, given the fact description of a multi-defendant case as a word sequence $s^f = \{w_1^f, ..., w_{l_f}^f\}$ where l_f represents the number of words, and a set of defendant names $D = \{d_1, d_2, ..., d_n\}$, where n is the total number of defendants ($n \geq 2$) and each name is a sequence of words $s^{d_i} = \{w_1^{d_i}, ..., w_{l_d}^{d_i}\}$, the task is to predict law articles $Y_a^{d_i}$, charges $Y_c^{d_i}$ and corresponding terms of penalty for each defendant d_i. Note there are potentially multiple law articles and charges associated with each defendant.

Similar to previous works, we formalize law article prediction and charge prediction task as multi-label classification tasks. In Chinese criminal law, there are five main types of penalties: *death penalty*, *life imprisonment*, *fixed-term imprisonment*, *criminal detention*, and *public surveillance*. Previous LJP works focused only on the first three types of penalties in sentencing prediction, primarily predicting the length of fixed-term imprisonment. In our dataset, we extensively annotate the terms of detention and surveillance and include the types of penalties in prediction targets. For instance, when formalized as multi-class classification problem (as we do in Section 4), the class labels are composed of both penalty type and length (if applicable) such as "*surveillance less than 1 year*" and "*imprisonment less than 6 months*". It can also be formed as classification (penalty type) + regression (penalty length) problem.

Additionally, we annotate the amount of fine for each defendant. To our knowledge, no LJP work has specifically focused on the prediction of defendants' fines. We include such data in our dataset for future research.

3.2 Dataset Construction

Figure 2 illustrates the annotation process. We detail each annotation step as follows.

Corpus and pre-processing. We collect 699,263 criminal documents involving two or more defendants from the past 20 years as our raw corpus from China Judgments Online[1]. Following Xiao et al. (2018), we only keep *judgment* documents for LJP tasks requirements.

[1] https://wenshu.court.gov.cn/

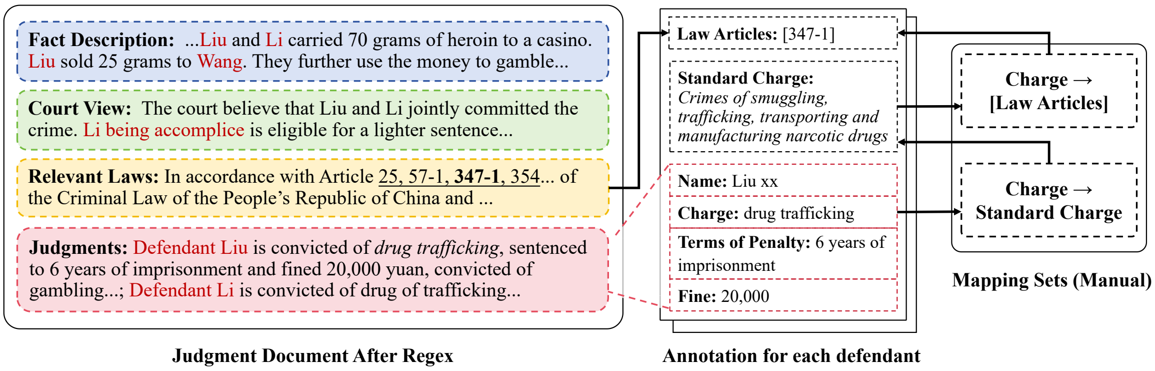

Fact Description: ...Liu and Li carried 70 grams of heroin to a casino. Liu sold 25 grams to Wang. They further use the money to gamble...
Court View: The court believe that Liu and Li jointly committed the crime. Li being accomplice is eligible for a lighter sentence...
Relevant Laws: In accordance with Article 25, 57-1, 347-1, 354... of the Criminal Law of the People's Republic of China and ...
Judgments: Defendant Liu is convicted of drug trafficking, sentenced to 6 years of imprisonment and fined 20,000 yuan, convicted of gambling...; Defendant Li is convicted of drug of trafficking...
Judgment Document After Regex
Law Articles: [347-1]
Standard Charge: Crimes of smuggling, trafficking, transporting and manufacturing narcotic drugs
Name: Liu xx
Charge: drug trafficking
Terms of Penalty: 6 years of imprisonment
Fine: 20,000
Annotation for each defendant
Charge → [Law Articles]
Charge → Standard Charge
Mapping Sets (Manual)

Figure 2: The data processing and annotation process of our dataset.

Given the fixed structure of most legal documents, as illustrated in Figure 2, we first use precisely designed regular expressions to divide the documents into four parts: *fact description*, *court view*, *relevant laws* and *judgments*. Cases with excessively short *fact description* (probably fail at regex) are filtered out. The judgment for each defendant can be further extracted from the *judgments* section, from which we double-check the number of defendants to filter cases with multiple defendants. We keep the *court view* part for potentially use in court view generation task (Ye et al., 2018; Yue et al., 2021b).

Standardized charge annotation. The Criminal Law of the People's Republic of China has undergone multiple revisions, resulting in the deletion or renaming of certain criminal charges. To obtain unified labels, we use the 2023 version of the criminal law as references, cases containing deleted charges are filtered out and we establish a mapping set for renamed charges. Besides, charges extracted via regex may not align precisely with the full standard charge names (often being partial or inaccurate), rendering them unsuitable for direct use as charge class labels. For example, there is a charge fully-named "*crime of luring, sheltering, and procuring prostitution*", which may only be referred as "crime of procuring prostitution". We also single out these charges and add their mappings to the mapping set through extensive manual checks.

Law annotation. The *relevant laws* section extracted corresponds to the laws cited for the entire case which are difficult to separate and attribute to each defendant. Also, the relationship between charges and law articles is not one-to-one (e.g. some articles add clarifications on the sentencing for specific charges, and certain paragraphs cover situations belong to charges in other articles). Therefore, with the assistance of a law student, we meticulously read through the Criminal Law and construct a mapping set from the standard charges to their relative legal provisions. In Chinese law, "*Article*" is the fundamental units where each potentially comprise multiple "*Paragraph*s" for more detailed subdivision and explanation. In legal documents, law articles are often cited down to the specific "Paragraph". Therefore, while previous LJP datasets annotate relevant laws at article-level(Xiao et al., 2018; Lyu et al., 2023), we provide paragraph-level annotations to be more accurate and meet a wider range of real-world needs. To annotate law label for each defendant, we map their charges to all possible relevant provisions through our mapping set, then include those that appear in the "*relevant laws*" section in their label list.

Term of penalty annotation. We retain the five principal punishments defined in the Criminal Law (introduced in Section 3.1) when labeling the penalty term, including terms of detention and surveillance, which were previously simplified as "fixed-term imprisonment of 0 months". When both life imprisonment and death penalty are labeled as "False", and the length of the other three penalty types are labeled as "0", it indicates that the defendant is "*exempted from criminal punishment*". We annotate the corresponding penalty terms for each charge. For a single defendant who commits several crimes, they typically receive a "*combined punishment*", which is also annotated as the "final penalty" in addition to the penalty terms for each crime.

Fine annotation. Apart from the above three LJP subtasks, we also extract the fines imposed on defendants using regular expressions and standardize them as integer types. For those unable to be directly converted into integer, we manually check

Dataset	CMDL-small	CMDL-big
# Training Set Cases	63,032	315,156
# Validation Set Cases	7,879	39,395
# Test Set Cases	7,879	39,394
# Charges	269	321
# Law articles(article-level)	239	275
# Law articles(paragraph-level)	462	564
Total Defendant	239,039	1,199,117
Max Defendant per Case	61	80
Average Defendant per Case	3.03	3.04
Max Charges per Defendant	15	15
Average Charges per Defendant	1.03	1.03
Max Charges per Case	14	19
Average Charges per Case	1.09	1.09
Average Input Length	1559.47	1564.61

Table 2: Statistics of CMDL.

and annotate them. Instances where fines are mentioned but without specifying exact amounts are also annotated.

Finally, we anonymize personal information (e.g. names and location) to protect privacy and avoid potential biases. It is worth emphasizing that the regular expressions and mapping sets we used have undergone repeated adjustments and manual checks to accommodate all possible descriptions, ensuring the quality of our dataset. Ultimately, the above steps were all executed through code, saving significant human and time effort while making the dataset more easily extendable. An example illustrating the final format and content of the dataset can be found in Appendix A.1.

The statistics of our dataset are listed in Table 2. We divide the dataset into training, validation, and testing subsets in an 8:1:1 ratio. Considering the time costs for training, we further extract a subset from the entire dataset. We denote the entire dataset as "CMDL" or "CMDL-big" and the subset as "CMDL-small".

3.3 Data Analysis

Case categorization. We further categorize and statistically analyze the cases based on the judgment results. Specifically, we use hierarchical classification criteria as follows: 1) single charge (**SC**) or multiple charges (**MC**) for each defendant in the case: if all defendants have only one charge each, it falls under SC, otherwise MC; 2) same charges (**SC**) or different charges (**DC**) for all defendants in a case: if all defendants have the same charges, it falls under SC, otherwise DC; 3) same penalties (**SP**) or different penalties (**DP**): under the premise that all defendants have the same charges, the case is classified as the former if all penalties are identical, otherwise as the latter. Notice that our dataset annotate penalties for each charge as well as a final penalty (if exists) for defendants with multiple charges, both penalties for each charge being different and final penalty different are all categorized as MC-SC-DP. Following these criteria, we categorize each case into six types, as shown in Figure 3. Detailed numbers of each category can be found in Appendix A.2.

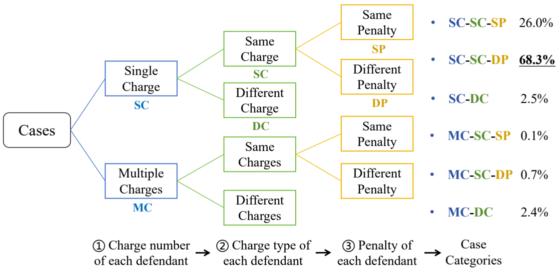


Figure 3: Hierarchical case categorization.

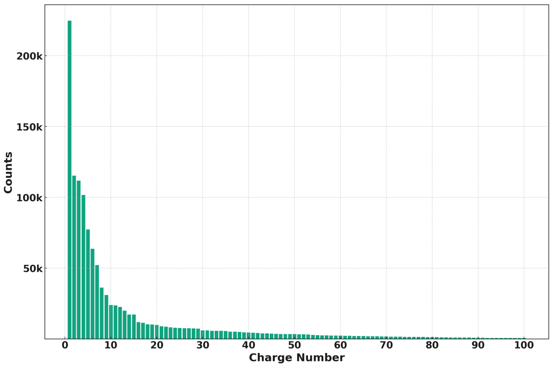


Figure 4: The distribution of top-100 frequent charges in our dataset.

According to statistics, only 5.7% involve multiple different charges while over 90% of the cases impose the same charge on multiple defendants. However 72.3% of the same-charge cases vary in the terms of penalty for different defendants. Cases where all defendants share one same charge but involve different terms of penalty (i.e. SC-SC-DP) are the most common type.

It is worth noting that there is a significant imbalance in the distribution of different categories within our dataset. As shown in Figure 4, the top 10 charge categories cover 66.63% of the defendants while the last 20 ones only appear less than 50 times in total. We retain low-frequency charge and law article categories to explore and demonstrate the challenges of few-shot issues.

4 Experiments

In this section, we apply state-of-the-art models to multi-defendant LJP based on CMDL, focusing on evaluating their performance and identifying the underlying challenges.

4.1 Baselines

We conduct experiments on seven SOTA models: (1) **TOPJUDGE** (Zhong et al., 2018), a multi-task learning framework for LJP by formalizing the dependencies among subtasks as a Directed Acyclic Graph; (2) **NeurJudge** (Yue et al., 2021a), a LJP framework separating fact to different circumstances for predictions; (3) **BERT** (Devlin et al., 2019), a pre-trained model based on Transformer architecture for Chinese; (4) **LEGAL-BERT-SC** (Chalkidis et al., 2020), a strategy of applying BERT in legal domain by pre-training BERT from scratch on legal corpora; (5) **Lawformer** (Xiao et al., 2021), a Longformer-based (Beltagy et al., 2020) PTM for Chinese legal long documents understanding. (6) **Dependant-T5** (Huang et al., 2021), a T5-based (Xue et al., 2021) model that exploits subtask dependencies. (7) **MAMD** (Pan et al., 2019), a multi-scale attention model for charge prediction in multi-defendant cases. Notice that models (1)-(2) are models designed specifically for LJP tasks while (3)-(6) are pre-trained models used in single-defendant LJP. Only MAMD is designed for multi-defendant cases but it solely predict charges. We exclude the recent multi-defendant LJP model HRN (Lyu et al., 2023) because it relies on manually-annotated data that are not included in our dataset.

4.2 Experimental Setting

Since most of the baselines are designed for single-defendant cases, we concatenate each defendant's name and the whole fact as the input to predict judgment results for each defendant as Lyu et al. (2023) do.

Law prediction are conducted at both article-level and paragraph-level. Penalty terms are divided into non-overlapping intervals with respective penalty type. As for defendants with multiple charges, the task is to predict the final term of penalty. All model settings basically follow their original papers. More details of baseline implementation can be found in Appendix B.

Due to limited computational resources and cost consideration, we train and test baselines on CMDL-small, which is derived from the complete dataset by maintaining the frequency of each charge (some charges with extremely low frequencies were thus excluded). Performance evaluation results are reported in terms of Accuracy (Acc.), Macro-Precision (MP), Macro-Recall (MR) and Macro-F1 (MF). All these metrics are evaluated at defendant level (metric scores are calculate in term of each label). To evaluate model performance more comprehensively, we extend these metrics to a case level as follows.

4.2.1 Case-level Evaluation Metrics

Concatenating fact description and one defendant name as input, and only predict judgments for that defendant each time is convenient for adopting existing models. However, simply using metrics that employed in single-defendant LJP works overlooks the fundamental difference between multi-defendant and single-defendant cases: *cases involving more defendants typically imply longer, more complex case facts and a higher difficulty for prediction* [2]. Therefore, we argue that it is necessary to distinguish between *defendant-level* and *case-level* metrics which are calculated as follows.

Take charge prediction as an example, the same applies to the other two tasks. Given a case c with n defendants, for defendant d_i, let there be m_1 labels present in its ground-truth, and during testing a model predicts m_2 labels out of which m_3 predictions are correct ($m_3 \leq m_2$ and $m_3 \leq m_1$). Then the accuracy for this defendant is 1 when prediction match the ground-truth exactly, the precision $p_c^{d_i}$ is m_3/m_2 and the recall $r_c^{d_i}$ will be m_3/m_1. From precision and recall scores we can compute F1 for this defendant, which is the harmonic mean of $p_c^{d_i}$ and $r_c^{d_i}$. The precision, recall, F1 and accuracy scores of case c is then computed by averaging the corresponding scores of all defendants. We obtain the final metric values by computing a weighted average of scores across all cases. For example specifically, the case-level precision score is

$$Precision_{case} = \frac{\sum_C w_c p_c}{\sum_C w_c}$$

where p_c is the precision score of case c:

$$p_c = \frac{\sum_{i=1}^{n} p_c^{d_i}}{n}$$

and w_c is the weight assigned to it. Here we calculate w_c as $\log_2 n$ where n is the number of defendants in c. The weight can also be considered as

[2] See Appendix C for more details.

Methods	Charges				Law Articles				Term of Penalty			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
TOPJUDGE	49.76	24.91	16.82	18.93	42.82	12.48	7.66	8.79	0.73	16.23	0.59	1.09
NeurJudge	18.36	35.91	36.08	35.84	18.34	34.58	34.75	34.49	20.55	4.97	8.00	4.08
BERT	75.53	51.83	47.84	48.41	76.90	52.00	48.38	48.73	22.42	15.67	14.41	14.74
LEGAL-BERT-SC	75.53	49.70	47.16	47.04	76.80	51.59	48.98	48.88	22.29	15.75	13.26	13.78
Lawformer	79.01	59.14	56.57	56.51	**80.20**	59.57	58.14	57.75	**29.16**	**30.15**	**21.19**	**23.96**
Dependant-T5	**81.72**	**70.90**	**68.30**	**67.60**	77.15	**76.90**	**75.80**	**75.90**	28.12	26.20	17.20	18.30
MAMD	35.40	19.64	7.07	9.53	–	–	–	–	–	–	–	–

(a) Article-level law prediction

Methods	Charges				Law Articles (Paragraph)				Term of Penalty			
	Acc.	MP	MR	F1	Acc.	MP	MR	F1	Acc.	MP	MR	F1
TOPJUDGE	49.96	25.88	19.72	20.54	37.26	7.60	4.19	4.79	1.29	23.35	1.53	2.75
NeurJudge	16.78	35.80	*	*	18.37	30.60	32.68	30.55	20.37	3.96	7.48	4.01
BERT	75.58	52.41	50.90	50.03	66.09	30.86	30.28	29.29	22.55	15.24	*	14.28
LEGAL-BERT-SC	76.07	54.97	50.46	50.84	65.37	30.49	29.55	28.48	22.97	19.40	14.43	14.62
Lawformer	*	61.10	56.80	57.25	69.98	39.48	37.47	36.55	28.33	26.85	**20.20**	**21.72**
Dependant-T5	**81.01**	**71.90**	*****	*****	**73.15**	**69.60**	**67.60**	**67.80**	**28.55**	**29.50**	18.30	20.40

(b) Paragraph-level law prediction

Table 3: Defendant-level judgment prediction results on CMDL-small. For law article prediction, we conduct experiments separately at the Article-level and the Paragraph-level law label set. Results of charge prediction and penalty prediction that show a minor discrepancy ($\delta < 0.05\%$) between two settings are marked with *. The best performance in each column is marked in **bold**.

"*prediction difficulty score*" indicating that the difficulty increase when defendant number increases. We use $\log_2$ function rather than other increasing functions because the difficulty does not dramatically increase with the number of defendants, as one legal expert suggests. The base 2 is chosen given the minimum number of defendants in multi-defendant cases is 2.

This evaluation approach treats each case as a whole and disregard the specific contents of the labels, providing a more intuitive reflection of the performance on cases of different scales. In addition to the originally calculated per-label metrics (i.e. defendant-level metrics), we evaluate models in a more comprehensive way. The higher the scores of case-level metrics, the better the model's prediction performance on complex cases.

During experiments and analysis we found that, the original test set predominantly contains cases with 2-4 defendants and very few with a larger number of defendants, making it hard to derive insights from case-level metrics. Thus, we additionally extract another test set not overlapping with the training set and containing a balanced amount of cases involving different number of defendants, to report case-level evaluation results[3]. Besides, we extract 50 cases of each type described in Section 3.3 also from data outside the training set to explore hard-to-predict case types. For these tests, we only employ Macro-F1 metric for comparison. Law prediction for case-level evaluation and case-type evaluation are both on article-level.

4.3 Results and Analysis

Defendant-level performance. The defendant-level performance of baselines on CMDL-small is shown in Table 3. It is evident that the pre-trained models significantly outperform other models in all tasks, demonstrating that *pre-trained models hold absolute advantage in multi-defendant LJP.* NeurJudge has the lowest accuracy on charge and law article prediction task, indicating it finds trouble perfectly predicting the results as we use exact-match accuracy. Some relatively lower macro scores are mainly due to the imbalance of label set. The extensive corpus used during pre-training can mitigate the impact of data imbalance to a certain extent even though LEGAL-BERT, which is pre-trained on legal copora, show little improvement comparing to original BERT. Additionally, pre-trained models' proficiency in handling long text is an important factor for multi-defendant LJP, where the input fact descriptions are generally longer. Specifically, Lawformer, known for its strong per-

[3] Refer to Appendix B for more details

Methods	Charges				Law Articles				Terms of Penalty			
	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
TOPJUDGE	64.17	69.00	69.29	68.50	75.08	28.27	28.26	28.13	0.72	1.21	0.97	1.02
NeurJudge	24.12	11.92	13.56	13.1	30.12	42.1	37.18	39.88	8.36	3.12	5.19	3.87
BERT	67.41	69.02	63.19	66.36	69.56	72.12	54.22	62.88	20.48	16.11	18.97	16.35
LegalBERT-SC	66.74	64.31	52.46	58.34	68.50	70.52	59.11	62.78	21.76	15.89	20.11	17.22
Lawformer	70.80	66.60	68.37	66.91	72.46	69.00	71.24	68.10	**28.12**	23.92	**28.01**	24.45
Dependant-T5	**79.08**	**81.39**	**79.89**	**80.35**	**83.70**	**87.84**	**86.22**	**86.66**	25.62	**26.87**	26.21	**26.41**

Table 4: Case-level judgment prediction results on CMDL-small. The law Articles prediction task is conducted on article-level label set. The best performance in each column is marked in **bold**.

formance in LegalAI tasks involving long legal documents (Xiao et al., 2021), outperform both BERT and LEGAL-BERT here. Model scale seems to be an important factor as well. The larger Dependant-T5, which is also the most time-consuming when training, performs the best overall.

Comparing model performance on article-level law prediction in Table 3a and paragraph-level prediction in Table 3b, as anticipated, we find that all models perform worse on paragraph-level, suggesting that *models struggle more with the prediction of finer legal provision labels.* We assume better performance on paragraph-level law prediction requires a much larger volume of training data to support the model learning to distinguish different paragraphs within the same article. The prediction of laws at different levels also impact the other two tasks. From the results, all pre-trained models and TOPJUDGE show slightly better performance on charge and penalty prediction when predicting law at paragraph-level. We hope finer-grained law categories will help models learn more detailed legal information, thus positively impacting the other two tasks.

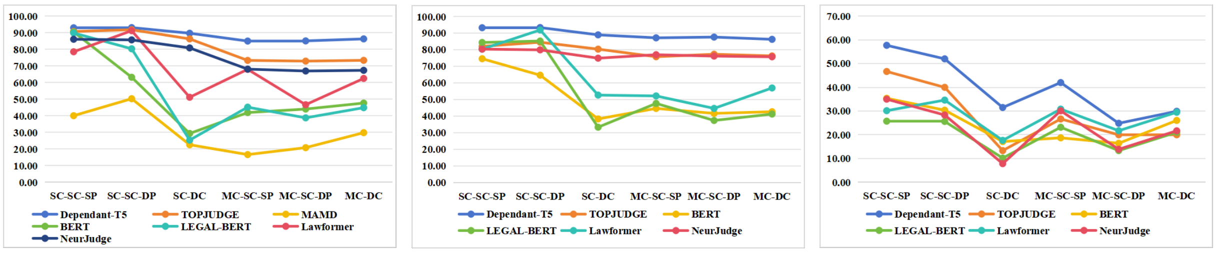
All models show unsatisfactory performance in terms of penalty prediction. Firstly, the task complexity significantly increases with the consideration of penalty type (e.g. detention) other than simply predicting a length. More importantly, many cases exhibit the same charges but different penalties for all defendants. In multi-defendant cases, intricate interactions and criminal relations (e.g. principal-accessory roles) exist among defendants, along with varying crime severity and sentencing factors (e.g. voluntary surrender), leading to considerable difference in the final penalties, as shown in the example from Figure 1. This information are difficult for models to accurately capture and understand. Lyu et al. (2023) address this problem by manually annotating these details and then training a hierarchical reasoning network, which we argue lacks scalability and transferability.

Case-level performance. As listed in Table 4, Dependant-T5 still achieves the best overall performance on cases of varying complexity. The noticeable better performance of Dependant-T5 on charge prediction task highlights its ability to remain consistently effective despite an increase in the number of defendants. In contrast, non-pretrained models like TOPJUDGE and NeurJudge that are specifically designed for single-defendant LJP perform poorly as the case scale increases.

Comparing Table 4 with Table 3, we observe that the case-level values are generally higher than the defendant-level ones, which is because each label contributes equally to these "*per-label*" metrics (e.g. macro-precision and macro-recall) in defendant-level evaluation, making it more susceptible to the performance on rare labels (Dutta et al., 2018). Besides, the relative performance of different models appears to be nearly consistent when evaluated at both defendant-level and case-level. This is primarily because the distribution of charge labels remains largely unchanged with the increase in the number of defendants, as analyzed in Appendix C. We believe that with a more complex dataset distribution and a greater variety of tested models, the differences between the two evaluation methods will be more pronounced, providing additional insights.

Performance on different case types. We evaluate model performance on different types of cases according to our categorization in Section 3.3. The results are shown in Figure 5.

Overall, *predicting cases involving defendants with multiple charges (MC) is more challenging*



(a) Charge prediction

(b) Law articles prediction

(c) Term of penalty prediction

Figure 5: Judgment prediction results on different types of cases. Macro-F1 scores are reported for each task.

than those where all defendants only have a single charge (SC), which is consistent with the distribution of different types of cases in the dataset. The performance of pre-trained models other than Dependant-T5 fluctuates markedly in the first two tasks, especially performing the worst when different defendants in a case have different charges (SC-DC). For term of penalty prediction, *predicting different penalties for defendants with the same charge is evidently more challenging.* since most models' performance drop significantly from predicting SC-SC-SP to SC-SC-DP and MC-SC-SP to MC-SC-DP.

5 Conclusion

This paper propose the first large-scale real-world Chinese multi-defendant dataset for LJP, which includes 393,945 multi-defendant cases, encompassing a total of nearly 1.2 million defendants and 321 charges. CMDL contains comprehensive annotations including fines, diverse terms of penalty, and paragraph-level law article labels. Experimental results indicate that multi-defendant LJP is challenging and requires further efforts for improvement.

Ethics Statement

The source files of our dataset are all from publicly available resource and personal information (e.g. name, location, etc.) is properly anonymized during dataset construction. Our motive is to inspire and facilitate LJP research with the purpose of assisting judges and laypersons, without any intention of developing techniques to replace judges.

Limitations

The label set in our dataset is quite unbalanced and lacks complete coverage of all charges and law articles in the criminal law. The primary reason is the inherent distribution of cases in real world. Given the scarcity of multi-defendant cases, we do not set a minimum limit for certain types of cases in order to make full use of these resources. More cases can be easily added with the automated annotation pipeline in the future. Speaking of which, the automated annotation method has its pros and cons. Though saving manual cost substantially, it cannot ensure perfectly proper data segmentation and extraction. To mitigate the issue, we conduct a certain degree of manual review with the manually constructed mapping sets and specific filtering mechanisms. We believe the final dataset is sufficiently reasonable and usable.

Acknowledgments

We thank all the reviewers for their comments on the earlier draft of this paper. This work was supported by the Overseas Project (KFKT2023A07, KFKT2024A07) and the General Innovation Project (ZZKT2024B02) of the National Key Laboratory for Novel Software Technology in China, the National Key R&D Program of China (2016YFC0800803) and the Cooperation Fund of NJU-Jinqiao Information Joint Research Center for Smart Justice and Artificial Intelligence.

References

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

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ayushi Dutta, Yashaswi Verma, and C. V. Jawahar. 2018. Automatic image annotation: The quirks and what works. *Multimedia Tools and Applications*, 77(24):31991–32011.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022a. Legal judgment prediction: A survey of the state of the art. In *Thirty-First International Joint Conference on Artificial Intelligence*, volume 6, pages 5461–5469.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022b. 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. Association for Computational Linguistics.

Leilei Gan, Kun Kuang, Yi Yang, and Fei Wu. 2021. Judgment prediction via injecting legal knowledge into neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12866–12874.

Yunyun Huang, Xiaoyu Shen, Chuanyi Li, Jidong Ge, and Bin Luo. 2021. Dependency learning for legal judgment prediction with a unified text-to-text transformer.

Elias Jacob de Menezes-Neto and Marco Bruno Miranda Clementino. 2022. Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from brazilian federal courts. *PLoS ONE*, 17:e0272287.

Yuquan Le, Yuming Zhao, Meng Chen, Zhe Quan, Xiaodong He, and Kenli Li. 2022. Legal charge prediction via bilinear attention network. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1024–1033, Atlanta GA USA. ACM.

Dapeng Li, Qihui Zhao, Jian Chen, and Dazhe Zhao. 2021. Adan: An intelligent approach based on attentive neural network and relevant law articles for charge prediction. *IEEE Access*, 9:90203–90211.

Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. 2022. Augmenting legal judgment prediction with contrastive case relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2658–2667, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xianquan Liu. 2020. *Criminal Jurisprudence*. Shanghai People's Publishing House.

Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. Ml-ljp: Multi-law aware legal judgment prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 1023–1034, New York, NY, USA. Association for Computing Machinery.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736.

Yougang Lyu, Jitai Hao, Zihan Wang, Kai Zhao, Shen Gao, Pengjie Ren, Zhumin Chen, Fang Wang, and Zhaochun Ren. 2023. Multi-defendant legal judgment prediction via hierarchical reasoning.

Yougang Lyu, Zihan Wang, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Xiaozhong Liu, Yujun Li, Hongsong Li, and Hongye Song. 2022. Improving legal judgment prediction through reinforced criminal element extraction. *Information Processing & Management*, 59(1):102780.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sicheng Pan, Tun Lu, Ning Gu, Huajuan Zhang, and Chunlin Xu. 2019. Charge prediction for multi-defendant cases with multi-scale attention. In *Computer Supported Cooperative Work and Social Computing*, Communications in Computer and Information Science, pages 766–777, Singapore. Springer.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. Classactionprediction: A challenging benchmark for legal judgment prediction of class action cases in the us.

Josef Valvoda, Ryan Cotterell, and Simone Teufel. 2023. On the role of negative precedent in legal outcome prediction. *Transactions of the Association for Computational Linguistics*, 11:34–48.

Yiquan Wu, Yifei Liu, Weiming Lu, Yating Zhang, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2022. Towards interactivity and interpretability: A rationale-based legal judgment prediction framework. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4787–4799, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Fanglong Yao, Xian Sun, Hongfeng Yu, Yang Yang, Wenkai Zhang, and Kun Fu. 2020. Gated hierarchical multi-task learning network for judicial decision prediction. *Neurocomputing*, 411:313–326.

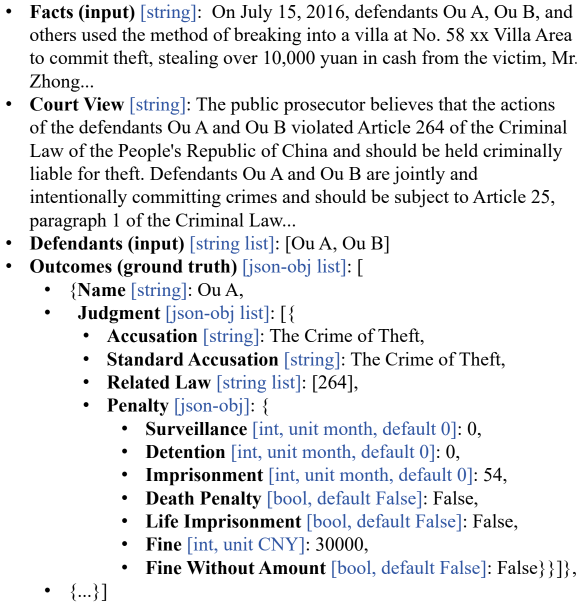
Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021a. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 973–982, New York, NY, USA. Association for Computing Machinery.

Linan Yue, Qi Liu, Han Wu, Yanqing An, Li Wang, Senchao Yuan, and Dayong Wu. 2021b. Circumstances enhanced criminal court view generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 1855–1859, New York, NY, USA. Association for Computing Machinery.

Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. Contrastive learning for legal judgment prediction. *ACM Transactions on Information Systems*.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.



- **Facts (input)** [string]: On July 15, 2016, defendants Ou A, Ou B, and others used the method of breaking into a villa at No. 58 xx Villa Area to commit theft, stealing over 10,000 yuan in cash from the victim, Mr. Zhong...
- **Court View** [string]: The public prosecutor believes that the actions of the defendants Ou A and Ou B violated Article 264 of the Criminal Law of the People's Republic of China and should be held criminally liable for theft. Defendants Ou A and Ou B are jointly and intentionally committing crimes and should be subject to Article 25, paragraph 1 of the Criminal Law...
- **Defendants (input)** [string list]: [Ou A, Ou B]
- **Outcomes (ground truth)** [json-obj list]: [
  - {**Name** [string]: Ou A,
  - **Judgment** [json-obj list]: [{
    - **Accusation** [string]: The Crime of Theft,
    - **Standard Accusation** [string]: The Crime of Theft,
    - **Related Law** [string list]: [264],
    - **Penalty** [json-obj]: {
      - **Surveillance** [int, unit month, default 0]: 0,
      - **Detention** [int, unit month, default 0]: 0,
      - **Imprisonment** [int, unit month, default 0]: 54,
      - **Death Penalty** [bool, default False]: False,
      - **Life Imprisonment** [bool, default False]: False,
      - **Fine** [int, unit CNY]: 30000,
      - **Fine Without Amount** [bool, default False]: False}}]},
  - {...}]

Figure 6: Example of CMDL annotation.

Type	Quantity	Percentage
SC-SC-SP	102,780	26.09%
SC-SC-DP	268,883	68.25%
SC-DC	9,826	2.49%
MC-SC-SP	224	0.06%
MC-SC-DP	2,875	0.73%
MC-DC	9,357	2.38%
TOTAL	393,945	-

Table 5: Number of different types of cases in CMDL.

A Dataset Details

A.1 Dataset Example

Figure 6 presents an example of our dataset.

A.2 Number of Cases

Table 5 lists the The specific numbers of cases for each type in CMDL, obtained based on the classification approach described in 3.3.

B Implementation Details

Baseline algorithms. All baseline are implemented using PyTorch[4] framework. We obtain pre-trained models from the Hugging Face Model Hub [5]. For BERT and LEGAL-BERT, we adopt bert-base-chinese (Devlin et al., 2019). For Dependant-T5, we adopt mT5-small (Xue et al.,

[4] https://pytorch.org/

[5] https://huggingface.co/models

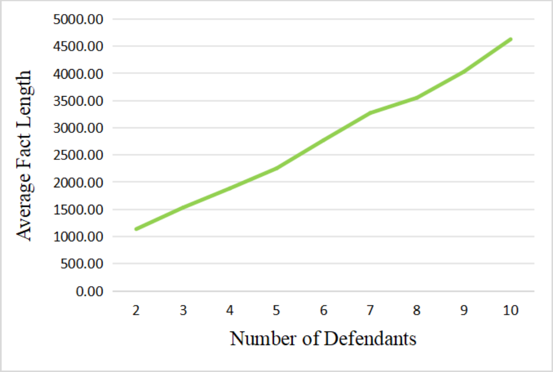


Figure 7: Average fact length of cases with different numbers (from 2 to 10) of defendants.

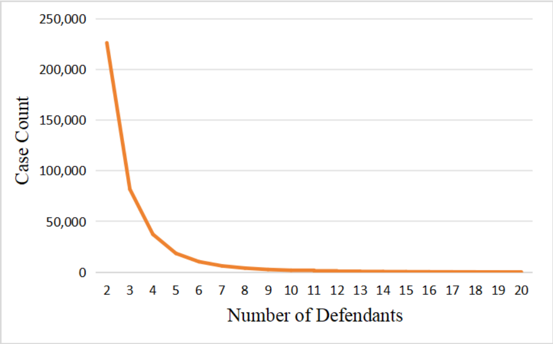


Figure 8: Number of cases with different numbers (from 2 to 20) of defendants.

2021), which has 300 million parameters, as the model base due to resource limitation. We set batch size as 128 and adopt the gradient accumulation strategy. For MAMD, we use one layer Bi-GRU and train for maximum 64 epochs. Other model structure and parameters are set according to the original paper. All experiments are conducted on a server with 2 RTX-3090-Ti GPU.

Metrics. For multi-label classification task, we adopt Exact Match for accuracy. Precision, recall and F1 score are calculated using scikit-learn metrics API [6] where zero-division is set to 1.

Case-level Evaluation. For case-level evaluation, we extract a dataset the from the part of CMDL that does not belong to CMDL-small, containing 200 cases for each defendant-number ranging from 2 to 16 (3000 cases in total).

C Prediction Difficulty according to Defendant-Num

First of all, Figure 7 illustrates the relationship between fact length and the number of defendants in a case, revealing a proportional relationship.

Furthermore, from the perspective of sample quantity, Figure 8 shows the relationship between the number of cases in CMDL and the number of defendants per case. It is evident that cases with 2-4 defendants constitute the majority. As the number of defendants increases, the quantity of cases decreases, resulting in the model lacking sufficient samples for learning, thereby increasing the difficulty of prediction. This can also be considered a form of few-shot challenge.

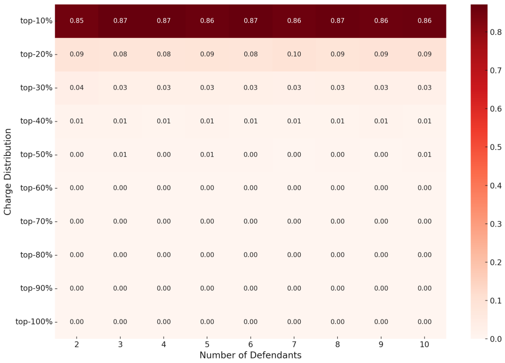


Figure 9: Charge distribution of cases with different numbers (from 2 to 10) of defendants.

However, there appears to be no specific relationship between the distribution of charges and the number of defendants. We arrange the charges in descending order or their frequency of occurrence in the entire dataset and divide them into ten groups (from top-10% to top-100%). Subsequently, we calculate the distribution of charges across these ten groups in cases with different numbers of defendants and depict this distribution in a heatmap, as shown in Figure 9. There is no trend of charges becoming rarer as the number of defendants in a case increases. Frequent charges remain frequent across cases with different numbers of defendants.

In summary, we believe that the complexity of case prediction exhibits a certain incremental relationship with the number of defendants in a case. However, the variation in difficulty primarily stems from the increase in fact length and the decrease in the number of cases, leading to a less steep increasing relationship.

[6] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics