Should multiple defendants and charges be treated separately in legal judgment prediction: An exploratory study and dataset

Anonymous ACL submission

Abstract

Legal judgment prediction (LJP) offers a compelling method to aid legal practitioners and researchers. However, the research question remains relatively under-explored: *Should multiple defendants and charges be treated separately in LJP?* To address this, we introduce a new dataset namely multi-person multi-charge prediction (MPMCP), and seek the answer by evaluating the performance of several prevailing legal large language models (LLMs) on four practical legal judgment scenarios: (S1) single defendant with a single charge, (S2) single defendant with multiple charges, (S3) multiple defendants with a single charge, and (S4) multiple defendants with multiple charges. We evaluate the dataset across two LJP tasks, i.e., charge prediction and penalty term prediction. We have conducted extensive experiments and found that the scenario involving multiple defendants and multiple charges (S4) poses the greatest challenges, followed by S2, S3, and S1. The impact varies significantly depending on the model. For example, in S4 compared to S1, InternLM2 achieves approximately 4.5% lower F1-score and 2.8% higher LogD, while Lawformer demonstrates around 19.7% lower F1-score and 19.0% higher LogD.

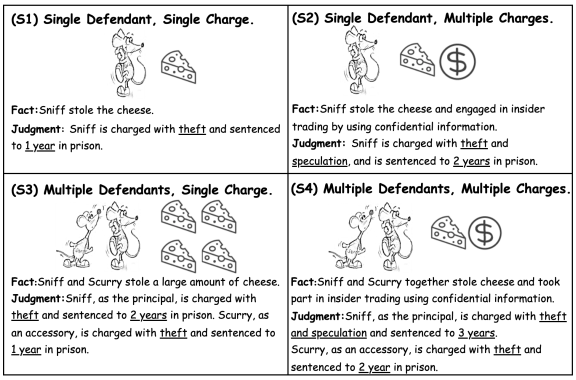

Figure 1: An illustration of the various charges and terms of penalty in four practical legal judgment scenarios: (S1) single defendant with a single charge, (S2) single defendant with multiple charges, (S3) multiple defendants with a single charge, and (S4) multiple defendants with multiple charges.

1 Introduction

Legal judgment prediction (LJP) is a crucial task for intelligent legal assistants, which aims to predict case outcomes based on factual descriptions (Cui et al., 2022). These outcomes typically encompass the types of charges and terms of penalty in the study of China's criminal law. The emergence of LLMs has significantly advanced research in this field. For instance, DISC-LawLLM (Yue et al., 2023) excels in providing comprehensive legal consultation, and LawBench (Fei et al., 2023) attracts an increasing number of LLMs for evaluation of legal tasks.

However, complex judgment prediction involving multiple defendants and multiple charges is common but highly challenging in real-world scenarios: In TOPJUDGE (Zhong et al., 2018), these complex cases are fully neglected to explore relationales between various subtasks. In MAMD (Pan et al., 2019), there are approximately 30.32% of cases involve multiple defendants. In MultiLJP (Lyu et al., 2023), 89.58 % of the cases the defendants received different judgments for at least one of the subtasks in the multi-defendant LJP task.

To address this gap, we introduce MPMCP dataset with four practical scenarios, as illustrated in Figure 1. For example, in (S4), the two defendants (i.e., Sniff and Scurry) should receive different outcomes (i.e., charges and penalty terms) based on the description of a fact involving two charges (i.e., theft and speculation). Unlike (S1), the factual description in (S4) involves more defendants and charges and provides more details, such as activities (i.e., stealing cheese and insider trading) and methods (i.e., using confidential information). As the number of defendants and charges increases, the complexity of the factual description also escalates, presenting greater challenges for prediction models. With an exploratory study of the proposed dataset, we seek to answer the main research question:

Should multiple defendants and charges be treated separately in LJP?

We use five prevailing open-source LLMs (i.e., MT5, MBERT, RoBERTa, LawFormer, and InternLM2) as benchmark models for generating charges and penalty terms across four scenarios in Chinese LJP. We also analyze the performance of InternLM2 variants under multiple settings to provide empirical insights into how these settings influence different scenarios. The main findings are that scenarios involving multiple defendants and multiple charges (S4) pose the greatest challenges, followed by S2, S3, and S1; The overall performance drops dramatically as the complexity of the scenario increases, although the relative impact varies significantly depending on the model. Our contributions include:

- MPMCP dataset, which encompasses four practical legal judgment scenarios involving multiple defendants and multiple charges.
- An exploratory study on benchmark models and the variant settings in different scenarios.

2 Related Work

Legal judgment prediction (LJP) is a critical task for smart legal assistants, which aims to predict the outcomes of legal cases given the description of facts (Cui et al., 2022). These outcomes usually include the types of the charge(s) and terms of penalty. Different countries have distinct legal systems (Sznycer and Patrick, 2020). Specifically, we focus on criminal legal cases in China.

Most related works have introduced datasets and methods to advance this field, as shown in Table 1. CAIL2018 (Xiao et al., 2018) release a large-scale legal dataset for fundamental LJP research considering a single defendant with a single charge. They implement several conventional text classification models (i.e., TFIDF+SVM, FastText, CNN) to facilitate the development and benchmarking of LJP models. Zhong et al. (2018) highlight the challenge of complex judgment prediction involving multiple defendants and multiple charges in real-world scenarios. However, their study focuses on exploring topological dependencies between various subtasks, without handling these complex cases. Pan et al. (2019); Lyu et al. (2023) focus on multi-defendant legal judgment prediction, without distinguishing whether the charges are single or multiple. CAIL-Long (Xiao et al., 2021) introduces Lawformer, a pre-trained language model specifically designed for Chinese legal long documents. This model addresses the challenges associated with processing lengthy legal texts, improving the accuracy of judgment predictions by leveraging a hierarchical transformer architecture. RLJP (Wu et al., 2022) generate rationales and outcomes separately to enhance the interactivity and interpretability of legal judgment. SLJA (Deng et al., 2023) present a method for syllogistic reasoning in legal judgment analysis and provide several LLMs as benchmarks. To sum up, these works address challenges such as handling long documents, and multi-defendant cases and enhancing logical reasoning with rationales. However, none of those works can fairly compare the difference between the four practical scenarios proposed in this study.

3 Dataset Construction

3.1 Raw data collection

We constructed the MPMCP dataset using first-instance documents collected from China Judgments Online[1], covering the period from 1998 to 2021. We exclusively obtain criminal cases with judgment outcomes and retain documents that clearly identify defendants, provide factual descriptions, and include charges, penalty terms, and applicable legal articles.

3.2 Data Extraction

We utilize regular expressions to directly extract relevant facts, applicable legal articles, charges, and penalty terms from four sections in a document, identified by inherent keyphrases, e.g., "Upon trial, it was found", "This court believes", and "The judgment is as follows". The first section provides a basic introduction to the case, which we do not consider relevant for dataset construction. The second section summarizes the facts of the case as determined by the court, based on statements from the parties involved, evidence presented, and court inquiries. This section is typically used as input for the LJP models. The third section contains the judge's explanation of the applicability of the law, including the legal articles referenced throughout the judgment process. The final section

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

Dataset	Defendant		Charge		#Case	#Charge	#Term	#Article
	Single	Multiple	Single	Multiple				
CAIL2018 (Xiao et al., 2018)	✓	✗	✓	✗	2,676,075	202	3	183
TOPJUDGE-CAIL (Zhong et al., 2018)	✓	✗	✓	✗	113,536	99	3	98
MAMD (Pan et al., 2019)	✓	✓	✓	NA	164,997	NA	NA	NA
CAIL-Long (Xiao et al., 2021)	✓	✗	✓	✗	229,505	201	5	244
RLJP (Wu et al., 2022)	✓	✗	✓	✗	89,768	48	1	95
SLJA-COR (Deng et al., 2023)	✓	✗	✓	NA	11,239	80	5	124
MultiLJP (Lyu et al., 2023)	✓	✓	✓	NA	23,717	23	11	22
MPMCP (Ours)	✓	✓	✓	✓	20,000	306	1	234

Table 1: Comparable public datasets for legal judgment prediction involving single vs. multiple defendants and charges. The symbol "✓" indicates that a characteristic is explicitly covered in a dataset, "✗" indicates that it is explicitly not covered, and "NA" denotes "not applicable" as it is not explicitly concerned in the reference work.

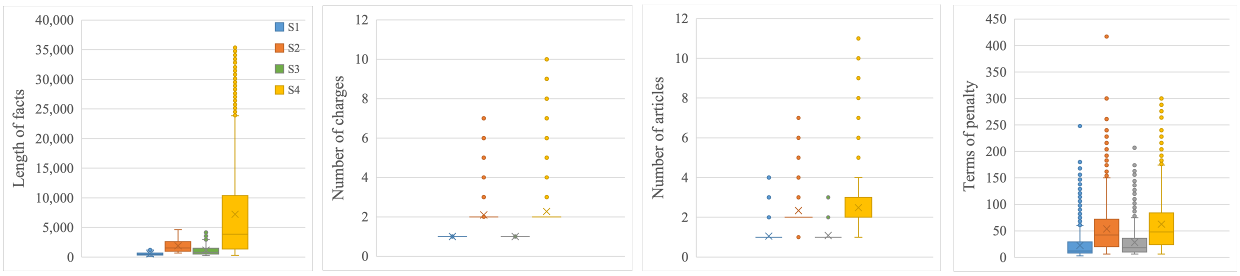

Figure 2: Box plots over the MPMCP dataset depict variations across four scenarios (S1, S2, S3, S4) for (a) number of facts, (b) number of charges, (c) number of legal articles, and (d) terms of penalty. In each box plot, the median is denoted by a line, and the mean value is marked by an "×".

details the judgments for each defendant, including the charges and the corresponding prison terms. Notably, we preserve all defendants and their corresponding judgments for each case to ensure the dataset accurately reflects the actual conditions of judicial rulings.

To ensure data quality, we mask any content within the extracted factual texts that precisely matched the names of the charges to prevent information leakage. We randomly select 5,000 cases for each of the 4 scenarios and manually assess approximately 5% of the data to ensure the inclusion of 20,000 qualified cases in the final dataset.

3.3 Data statistics

Figure 2 depicts the statistics of the proposed dataset. We observe that: First, the length of facts exhibits significantly higher median and mean values in (S4) compared to (S1, S2, S3), with the largest interquartile range (IQR) indicating diverse lengths. Similarly, this trend is observed in "terms of penalty" and "number of articles", where (S4) exhibits greater variability and higher median, mean, IQR values compared to (S1, S2, S3). This suggests that in (S4), the legal cases are more complex. Second, the number of charges is predominantly concentrated on 1-2 charges. Compared to (S1, S3) involving only 1 charge per case, scenarios (S2, S4) exhibit an average of 2 charges per case, with several outliers ranging from 3 to 10 charges.

4 Experimental Setup

4.1 Benchmark Models

We leverage the following five prevailing open-source LLMs for Chinese LJP as benchmark models to generate outputs in four scenarios.

MT5 (Xue et al., 2021), a T5 variant with multilingual capabilities, pre-trained on a novel dataset derived from Common Crawl, encompassing 101 languages.

MBERT (Devlin et al., 2019), a BERT model pre-trained on 104 of the most resource-rich languages in Wikipedia, supporting multilingual functionality.

RoBERTa (Liu et al., 2019), a variant of the BERT (Kenton and Toutanova, 2019) with modifications to training dynamics.

Lawformer (Xiao et al., 2021), a longfomer-based model pre-trained using extensive Chinese legal long case documents on a large scale

InternLM2 (Cai et al., 2024), built upon internlm2-base and additionally pre-trained on domain-

Model	Charge																Penalty Term			
	Accuracy (%) ↑				Precision (%) ↑				Recall (%) ↑				F1-Score (%) ↑				LogD (%) ↓			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
MT5	75.2	45.4	68.8	30.0	77.7	77.6	73.2	72.7	70.0	67.2	68.8	57.7	76.4	72.0	70.9	64.3	60.7	62.0	79.8	68.3
BERT	78.6	44.6	77.8	29.8	78.6	67.8	77.8	62.8	78.6	64.9	77.8	57.7	78.6	66.3	77.8	60.1	45.8	51.5	53.6	56.8
RoBERTa	81.0	47.0	75.2	30.8	81.0	71.9	75.2	64.1	81.0	69.1	75.2	60.3	81.0	70.5	75.2	62.1	43.3	49.3	51.7	57.1
Lawformer	81.4	52.0	78.0	34.8	81.4	73.8	78.0	64.1	81.4	71.0	78.0	59.4	81.4	72.4	78.0	61.7	**39.5**	**46.4**	**48.7**	58.5
InternLM2	**84.6**	**80.2**	**81.4**	**56.2**	**85.8**	**92.1**	**81.7**	**84.1**	**84.8**	**91.6**	**80.4**	**77.7**	**85.3**	**91.8**	**81.0**	**80.8**	59.3	54.1	61.3	**56.5**

Table 2: Main results of benchmark models in scenarios S1, S2, S3, S4. Bold font indicates the highest value in each column. "↑" denotes higher values are better, while "↓" denotes lower values are better.

Setting	Charge																Penalty Term			
	Accuracy (%) ↑				Precision (%) ↑				Recall (%) ↑				F1-Score (%) ↑				LogD (%) ↓			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Fine-tuning	**84.6**	**80.2**	**79.6**	56.2	**85.8**	**92.1**	**81.7**	**84.1**	**84.8**	**91.6**	**80.4**	77.7	**85.3**	**91.8**	**81.0**	**80.8**	59.3	54.1	61.3	56.5
Multi-task	79.0	79.6	68.0	**56.4**	84.7	91.0	77.3	80.9	80.0	91.6	68.6	**80.6**	82.3	91.3	72.7	80.7	**50.9**	**35.8**	58.2	**53.4**
/wo example	55.4	37.2	55.6	26.0	64.1	62.1	70.2	54.0	55.8	73.6	56.0	62.1	59.7	67.3	62.3	57.8	105.6	83.8	103.1	84.0
/w example	61.8	58.6	69.2	37.2	67.8	81.9	77.7	64.7	62.4	82.9	69.6	72.6	65.0	82.4	73.4	68.4	56.3	61.7	**56.9**	70.8

Table 3: Analysis study of variant settings for InternLM2 in scenarios S1, S2, S3, S4. Bold font indicates the highest value in each column. "↑" denotes higher values are better, while "↓" denotes lower values are better.

specific corpora, excels in its designated field evaluations while retaining strong general language capabilities.

4.2 Evaluation Metrics

We evaluate the generated legal judgment results in terms of charge prediction and penalty terms, following recent works (Deng et al., 2023; Pan et al., 2019), across four scenarios. Charge prediction is assessed as a standard classification task, and we utilize commonly used metrics, i.e., *Accuracy*, *Precision*, *Recall*, and *F1-score*, to evaluate its performance. The penalty term prediction is assessed by commonly used *LogD* (Cui et al., 2022), which measures the logarithmic difference between the predicted penalty term and the ground truth value.

5 Outcomes

We conduct massive experiments on several benchmark models in different scenarios, as shown in Table 2. First, scenario (S4) involving multiple defendants and multiple charges shows a significant drop in all evaluation metrics across most models, followed by S2, S3, and S1. For example, in S4 compared to S1, InternLM2 achieves approximately 4.5% lower F1-score and 2.8 higher LogD, while Lawformer demonstrates around 19.7% lower F1-score and 19.0 higher LogD. This demonstrates that scenarios involving multiple defendants and charges are still challenging and cannot be treated as simply as the single defendant and/or charge scenarios. Secondly, the impact of scenarios varies significantly depending on specific models. Compared with the top-performing model, InternLM2, the inferior models exhibit larger differences across the scenarios. For example, Lawformer decreases by 19.7% in F1-score from (S1) to (S4), while InternLM2 drops only 4.5%. Third, we analyze the variant settings for InternLM2 as shown in Table 3 and find that supervised fine-tuning on separate subtasks achieves the best overall performance. Learning in a multi-task setting increases the difficulty of task accomplishment, resulting in inferior values. Last, adding an example in a prompt yields better performance compared to prompts without examples.

6 Conclusion

In this paper, we introduce a dataset with four practical scenarios involving various numbers of defendants and charges in Chinese legal judgment prediction. We aim to answer whether multiple defendants and charges should be treated separately by comparing experimental results on several benchmark models across different scenarios. We find that scenarios involving multiple defendants and/or multiple charges pose great challenges. We call for future work in the research community to propose advanced models to facilitate smart legal assistants with real-world cases.

Limitations

While our study provides valuable insights, several limitations should be acknowledged. First, the dataset, sourced exclusively from Chinese criminal cases, may limit the generalizability of our findings to other legal systems. Second, the complexity of our dataset, especially with multiple defendants and charges, might affect how well models perform. Using a more balanced dataset with different types of cases could help. Potential biases in the training data could also affect model fairness, and despite anonymization efforts, data privacy risks remain, necessitating robust techniques and compliance with privacy regulations. Third, we notice that model performance varies, with some models struggling in complex scenarios, and the evaluation metrics used may not fully capture the nuances of legal judgments. Improving models through extra fine-tuning or combining different models might reduce this issue. Lastly, the black-box nature of LLMs limits their interpretability for understanding how they make decisions, posing challenges for practical use in the legal domain where decision transparency is critical. Developing methods for better transparency and decision justification could address this issue, making the models more usable in practice. Addressing these limitations is essential for advancing legal judgment prediction and ensuring the ethical and practical deployment of LLMs in the legal field.

Reproducibility

To support the development of research and ensure the reproducibility of our work, we will make the dataset and code available at https://anonymous.4open.science/status/MPMCP-07F4.

Ethical Statement

Throughout this research, we strictly followed ethical guidelines to ensure the responsible use of AI use and protect human data. We closely monitored LLMs employed to avoid generating harmful or biased content, especially in sensitive areas such as legal judgments.

Data Anonymization and Privacy

Data privacy is a top priority in our research. Since part of the data comes from legal judgments and contains sensitive information. To protect the privacy of the individuals involved, we implement strict anonymization procedures for any human data. We carefully remove or replace all identifiable information, such as names, addresses, and specific personal details to ensure confidentiality and anonymity.

Ethical Concerns

We carefully consider the ethical implications throughout our research and strictly follow the ethical guidelines of the institute. We aim to minimize any potential harm or misuse of the data and individual information. Future researchers who wish to use the dataset and findings should also follow these ethical standards, ensuring the data is used responsibly and ethically to advance knowledge in the field.

Use of AI Tools

In this work, we utilize AI tools (e.g., ChatGPT and Grammarly), solely for checking grammatical errors.

References

Zheng Cai, Maosong Cao, Haojiong Chen, et al. 2024. Internlm2 technical report.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009.

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

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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: 14th CCF Conference, ChineseCSCW 2019, Kunming, China, August 16–18, 2019, Revised Selected Papers 14*, pages 766–777. Springer.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.

Daniel Sznycer and Carlton Patrick. 2020. The origins of criminal law. *Nature human behaviour*, 4(5):506–516.

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.

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

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

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services.

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.

A Implementation Details

We fine-tuned benchmark models using a training dataset comprising 4,000 cases for each setting, followed by validation on 500 cases. Subsequently, we evaluated the models using a testing dataset of 500 cases. Additionally, we used prompt templates with and without examples for InternLM2 generation on distinct subtasks, and we applied the same method in a multitask setting.

Following the conclusions in (Shui et al., 2023), we utilized the BM25 [2] retriever to select the most similar case from the test set in each setting, which we then added as an example in our LLM generation. The details of our prompt templates for each setting are provided in Appendix C.

B An Example of Data

See Table 4 for the examples of our dataset in four different settings.

C Prompt template

Prompt templates for LLMs to generate outcomes with an example or without an example are shown in Table 5

[2] https://pypi.org/project/rank-bm25/

S1 (Single Defendant Single Charge)
Defendant: A1 **Fact**: Between October 7 and 20, 2019, ***defendant A1 stole*** a total of 2,140 yuan from victims A2, A3, and A4 in Hanjiang District, Putian City. A was arrested on October 22, and the stolen cash was recovered and returned. On March 16, 2020, A signed a plea agreement. During the trial, A did not dispute the facts. The evidence was sufficient to confirm A's crimes. **Legal Judgment**: ①**Charges:**Theft ②**Penalty Term:** 8 Months
S2 (Single Defendant Multiple Charges)
Defendant: B1 **Fact**: On November ..., ***B1 sold drugs*** to B2 near a hospital in....B1 was caught with 200 yuan and one packet of heroin... B1 did not contest the facts; the evidence was sufficient. On March ..., ***B1 injured*** B4 during a dispute, ***causing minor injuries***... B1 did not contest the facts; the evidence was sufficient. **Legal Judgment**: ①**Charges:**Trafficking Drugs, Intentional Injury ②**Penalty Term:** 10 Years
S3 (Multiple Defendants Single Charge)
Defendant: C1 **Fact**:On May 20, 2019, ***C1 and C2*** conspired to ***steal*** at MingmenShijia Community... ***C1 stole*** 100 yuan and a gold pendant from C3's home... Items were not recovered. ***C1 and C2*** confessed; evidence was sufficient. **Legal Judgment**: ①**Charges:**Theft ②**Penalty Term:** 11 Months
S4 (Multiple Defendants Multiple Charges)
Defendant: D1 **Fact**:Between April and August, ***D1 and D2 placed 27 gambling machines*** in Taizhou, earning 68,323 yuan. ***D1 earned 20,000 yuan***, D2 ... D1 also placed one machine alone, paying a 2,100 yuan bribe. In August, ***D3 contacted D1 and D2 to sell over 40 gambling machines*** ..., earning 180,000 yuan...The evidence was sufficient, and all three had no objections. **Legal Judgment**: ①**Charges:**Operating a Gambling Den, Illegal Business Operations ②**Penalty Term:** 4 Years and 3 Months

Table 4: Examples of data in four scenarios of MPMCP dataset.

Charge Prediction

Instruction:
请你模拟法官依据下面事实和被告人预测被告的罪名（一/多个）。
只按照例如的格式回答，不用解释。例如：被告人A其行为构成XX罪。
Please simulate a judge and predict all the charges (single/multiple) of the defendant based on the following factual description.
Respond only in the format provided, without explanation. For example: Defendant A is charged with XX.
Example:
下面是一个预测被告罪名的例子 Here is an example:
被告人 Defendant: B
事实 Fact: [*Fill based on the retrieval results of BM25*]
预测 Prediction：被告人B其行为构成XX罪。/ Defendant B is charged with XX.
Input:
被告人 Defendant：[*Fill based on the incoming data*]
事实 Fact：[*Fill based on the incoming data*]

Penalty Term Prediction

Instruction:
请你模拟法官根据下列事实和被告人预测被告的判决刑期。
只按照例如的格式回答，不用解释。例如：判处被告 人A有期徒刑X年X个月。
Please simulate a judge and predict the penalty term of the defendant based on the following factual description.
Respond only in the format provided, without explanation. For example: Defendant B is sentenced to X Years X Months.
Example:
下面是一个预测被告刑期的例子 Here is an example:
被告人 Defendant: B
事实 Fact：[*Fill based on the retrieval results of BM25*]
预测 Prediction：判处被告人B有期徒刑X月。 / Defendant B is sentenced to X Months.
Input:
被告人 Defendant：[*Fill based on the incoming data*]
事实 Fact：[*Fill based on the incoming data*]

Multitask: Charge and Penalty Term Prediction

Instruction:
请你模拟法官根据下列事实和被告人预测被告的所有罪 名（多个）以及最终判决刑期。
只按照例如的格式回答， 不用解释。例如：被告人A其行为构成XX罪、XX罪，判处 有期徒刑X年X个月。
Please simulate a judge and predict all the charges (single/multiple) and terms penalty of the defendant based on the following factual description.
Respond only in the format provided, without explanation. For example: Defendant A is charged with XX, and sentenced to X Years X Months.
Example: 下面是一个预测被告罪名和刑期的例子 Here is an example:
被告人 Defendant: B
事实 Fact: [*Fill based on the retrieval results of BM25*]
预测 Prediction：被告人B其行为构成XX罪,被判处有期徒刑X月。 / Defendant A is charged with XX, and sentenced to X Months.
Input:
被告人 Defendant：[*Fill based on the incoming data*]
事实 Fact：[*Fill based on the incoming data*]

Table 5: Prompt templates used in this paper.