Divide and Conquer: Legal Concept-guided Criminal Court View Generation

Qi Xu[1], **Xiao Wei**[1], **Hang Yu**[1]*, **Qian Liu**[2], **Hao Fei**[3]
[1]School of Computer Engineering and Science, Shanghai University, China
[2]School of Computer Science, University of Auckland, New Zealand
[3]School of Computing, National University of Singapore, Singapore
{welch,xwei,yuhang}@shu.edu.cn, Liu.Qian@auckland.ac.nz, haofei37@nus.edu.sg

Abstract

The Criminal Court View Generation task aims to produce explanations that inform judicial decisions. This necessitates a nuanced understanding of diverse legal concepts, such as `Recidivism`, `Confess`, and `Robbery`, which often coexist within cases, complicating holistic analysis. However, existing methods mainly rely on the generation capability of language models, without paying enough attention to the important legal concepts. To enhance the precision and depth of such explanations, we introduce Legal Concept-guided Criminal Court Views Generation (LeGen), a three-stage approach designed for iterative reasoning tailored to individual legal concepts. Specifically, in the first stage, we design a decomposer to divide the court views into focused sub-views, each anchored around a distinct legal concept. Next, a concept reasoning module generates targeted rationales by intertwining the deconstructed facts with their corresponding legal frameworks, ensuring contextually relevant interpretations. Finally, a verifier and a generator are employed to align the rationale with the case fact and obtain synthesized comprehensive and legally sound final court views, respectively. We evaluate LeGen by conducting extensive experiments on a real-world dataset and experimental results validate the effectiveness of our proposed model. Our codes and dataset are available at `https://github.com/xuqi220/LeGen`.

1 Introduction

The criminal court view generation task aims to automatically generate rationales supporting final legal judgments (Wu et al., 2020; Huang et al., 2020; Deroy et al., 2023; Wu et al., 2022; Yue et al., 2021, 2024). Essentially, this involves automatically generating accurate, clear, and logical texts that summarize criminal case facts and justify legal verdicts, such as charges and sentencing. It is pivotal in the development of legal artificial intelligence (LegalAI), which not only assists legal professionals but also contributes to transparency in the judicial process by making legal decisions understandable and justifiable to the public.

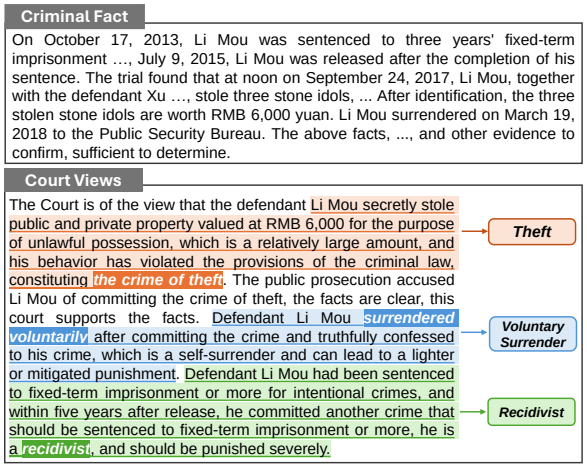

Figure 1: Given the criminal fact, the ideal court views should contain applicable legal concepts, such as `Theft`, `Voluntary Surrender`, and `Recidivist`, and corresponding rationales.

Recently, Large Language Models (LLMs) have exhibited impressive performance on various tasks due to their powerful contextual learning capabilities (OpenAI, 2023a; Taori et al., 2023; Touvron et al., 2023; Wu et al., 2024). Unfortunately, as illustrated in Fig.2 (a), commonly used LLMs (*e.g.*, ChatGPT (Ouyang et al., 2022), ChatGLM (Du et al., 2022)) show inferior performance on the court view generation task with the naïve prompt method. To tackle this issue, some researchers explore fine-tuning LLMs on legal datasets, such as Legal Question Answering (Cui et al., 2023; Huang et al., 2023; Yue et al., 2023) and Court View Generation (Li, 2023). However, these efforts have not yielded satisfactory results on the court view generation task (Yue et al., 2024).

Most existing works formulate the court view generation as a text summarization task, however,

* corresponding author

Findings of the Association for Computational Linguistics: EACL 2024, pages 3395–3410

it presents greater challenges (Li and Zhang, 2021), particularly in adhering strictly to the crime facts and legal provisions. As shown in Fig.1, it not only needs to summarize the criminal fact but also precisely reason *legal concepts* from the fact, *e.g.*, `Theft`, `Voluntary Surrender`, and `Recidivist`. Moreover, the court view is centered around these legal concepts to provide explanations, like details of the case, legal reasoning, and decisions made by the court.

As a fundamental part of the reasoning process of law, legal concepts play an important role in the legal system (Frandberg, 1998; Savelka et al., 2023). In this work, the legal concept refers to legal principles defined in the Criminal Law of China. Within the Chinese legal system, the criminal court view formally contains several sub-views, each dedicated to the rationale for a specific legal concept. The rationale could be seen as a summary of the case facts from the legal concept aspect.

Inspired by this, we propose a new Legal concept-guided court view Generation framework named LeGen in divide-conquer strategies[1]. Specifically, based on the predicted applicable legal concepts of a case, our framework first decomposes the complex court view into several more manageable sub-views that only require fewer legal reasoning steps. We then employ LLM to generate multiple rationales for each sub-view based on the given legal concept independently, and a verifier is subsequently employed to select the rationale that obtains the maximum alignment score with the case fact. Finally, the court view is obtained by summarizing these sub-views. We first implement the LeGen by prompting LLMs (denoted as LeGen-PT), as illustrated in Fig.2 (b), which surpasses the naïve prompt-based method by improving 14.6% in terms of average rouge score, which confirms the efficacy of legal concepts in improving the quality of the generated court view. Besides, we further implement the LeGen by fine-tuning LLMs (denoted as LeGen-FT), which achieves state-of-the-art performance. Our contributions are as follows:

- We present LeGen, a streamlined and potent framework designed to synthesize criminal court views underpinned by legal concepts, innovatively integrating them as prompts for large language models (LLMs).
- To rigorously assess the efficacy of LeGen, we introduce a novel dataset, meticulously crafted with linked legal concepts and exhaustive rationales, specifically tailored for criminal court view generation tasks. This benchmark resource fills a critical gap in the field and enables comprehensive evaluation of our model and future comparative studies.
- Extensive experimentation reveals LeGen's superior capabilities. Specifically, the LeGen-PT achieves a 14.6% increase in Average Rouge Score over the naïve prompt-based LLMs. Moreover, the LeGen-FT outperforms the prior state-of-the-art EGG (Yue et al., 2024), affirming its advancement in generating coherent, contextually relevant court narratives. Our contributions thus offer a robust foundation for advancing AI's role in legal text synthesis and setting a new benchmark.

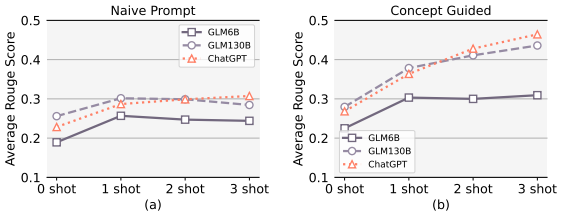

Figure 2: The naïve prompt-based LLMs (*e.g.*, ChatGPT) underperform on the criminal court view generation task in terms of Average Rouge Score (a). In contrast, we observe that their performance is significantly improved by incorporating the legal concepts (b).

2 Related Works

Court Views Generation. Recently, generation has gained significant attention (Liu et al., 2024; Zeng et al., 2024; Suo et al., 2024). As the fundamental task of LegalAI (Zhong et al., 2020a), the criminal court view generation task has attracted substantial attention (Yue et al., 2024; He et al., 2023a; Wu et al., 2022). As far as we know, Ye et al. (2018) is the first to explore the court view generation task and propose a charge label-enhanced method. Yue et al. (2021) propose a two-stage framework splitting the court views into two parts that are generated independently. He et al. (2023a); Wu et al. (2022) jointly modeling the legal judgment prediction task and court views generation task, and the generated court views can serve as explanations for prediction results. Yue et al. (2024) propose a model that generates the court view by

[1]Divide and Conquer algorithm is a problem-solving strategy that involves breaking down a complex problem into smaller, more manageable parts, solving each part individually, and then combining the solutions to solve the original problem.

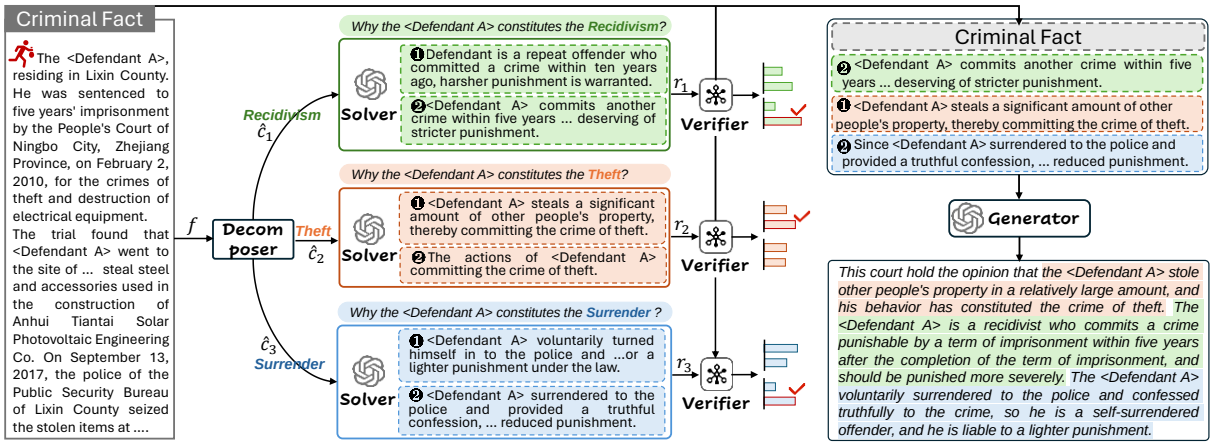

Figure 3: Overall architecture of our framework LeGen, which consists of four components: the *Decomposer*, the *Solver*, the *Verifier*, and the *Generator*. In particular, given the criminal fact, we first apply the *decomposer* to split the court views into several sub-views based on the predicted legal concept. Then the *solver* followed by the *verifier* is applied to generate and filter rationales. Finally, the *Generator* aggregates sub-views to craft the criminal court views with the rationales guidance.

injecting legal events into the LLMs. Despite these efforts achieving promising progress, they overlook that the generation process could be seen as applicable legal concept reasoning based on the given facts. This inspires us to enhance the precision and depth of the court views, guided by legal concepts.

LLMs in Legal AI. Recently, LLMs such as ChatGPT (Ouyang et al., 2022), ChatGLM (Du et al., 2022), and LLama (Touvron et al., 2023) have exhibited impressive performance across various complex tasks, such as knowledge graph completion (Li et al., 2024a), spatial relation matching (Chu et al., 2024), and logical reasoning (Xu et al., 2024). In LegalAI, some researchers explore combining LLMs with legal-related tasks. For example, He et al. (2023b) fully pre-trained a LLM with 7 billion parameters with the legal instruction dataset. Huang et al. (2023); Liu et al. (2023) fine-tuning LLMs on the Chinese legal datasets containing various LegalAI tasks, such as legal question answering (Zhong et al., 2020b) and case analysis (Deng et al., 2023). Particularly, Li (2023) fine-tune LLMs based on the criminal court view generation, and Cui et al. (2023) further handle the model hallucination issues. Despite these works achieving promising results, they overlook the ability of legal concept reasoning, which is a key step in the court view generation process. To fill this gap, in this work, we focus on improving the legal concept reasoning ability of LLMs.

3 Methodology

Task Definition. We first introduce the task definition and the used terms in the criminal court view generation. **Fact** denotes the description of a criminal case involving the defendant's background information (*e.g.*, previous convictions), which is denoted as $f = \{w_1, w_2, ..., w_{|f|}\}$ where w_i is the i-th word. **Legal Concept Set** consists of n applicable legal concepts of a fact f, denoted as $\mathcal{C} = \{c_1, c_2, ..., c_n\}$. **Court View** is the summary of the criminal case with a sequence of words, denoted as v. In this work, given the fact f, our goal is to automatically generate the court view v. It typically comprises multiple sub-views, each one centered around a specific legal concept $c_i \in \mathcal{C}$, accompanied by its corresponding rationale r_i.

Overview. The overall architecture of our LeGen framework is shown in Fig. 3. The basic idea is to generate the court view step-by-step with the guidance of the legal concepts. Our method consists of four modules, in a *divid and conquer* way. The *Decomposer* is to *divide* the court view into multiple sub-views generation, by predicting a set of related legal concepts. Then, for each sub-view, the *Solver* followed by the *Verifier* is used to generate and select the rationales with the highest align score with the fact. Finally, the *Generator* is used to integrate selected rationales for all legal concepts and the fact to generate the final court view. We detail the four modules below.

Decomposer. In judicial scenarios, the court views contain legal concepts and corresponding rationales. Given a criminal case fact, this module is designed to predict legal concepts, which are guidance to split court views into sub-views. Specifically, we first extract the representation of

Legal Concept-aware Prompt

Why the <Defendant A> constitute the ***Recidivism***?

Fact: <Defendant A>, born on April 17, 1969, ...; On December 12, 2017, he was sentenced to six months' imprisonment by Hangzhou Railway Transportation Court for committing theft, and on January 8, 2018, he was released after completing his sentence. The trial found that: at 6:00 on September 16, 2018, ..., < Defendant A> cover lookout, Li Fengnian hand Wang Moumou placed on the seat of an Apple iphoneX type (Hong Kong version) 64G cell phone stolen and sold stolen. It was determined that ....

Answer: <Defendant A>, having been sentenced to a term of imprisonment for an intentional crime, commits a second crime within five years of the completion of the execution of the sentence, is a ***recidivist*** and shall be punished more severely.

...

Fact: <Defendant A>, ..., was sentenced to six months' fixed-term imprisonment by the People's Court, for committing burglary on February 7, 2018, ..., and was released on April 14, 2018 upon completion of his sentence. The trial found that ,.., <Defendant A> stole a Huawei P20 cell phone placed on the coffee table from the victim Lu Mou while she was asleep ...

Answer:

Figure 4: Prompting template for generating rationales of specific legal concepts.

Court View Generation Prompt

Please generate the court view based on fact and sub-views?

Fact: Defendant A, ..., and was also punished with a fine of two thousand yuan, and was released on April 14, 2018 upon completion of his sentence. The trial found that at about 14:20 on May 9, 2019, <Defendant A> stole a Huawei P20 cell phone placed on the coffee table from the victim Lu, ...arrested and returned to the case. ....

Sub-Views: Defendant A stole other people's property ... constitutes the crime of theft; Defendant A voluntarily surrendered to ... lighter punishment.

Court View: This court is of the opinion that the defendant A stole other people's property ..., so he is a self-surrendered offender and shall be given a lighter punishment according to the law.

...

Fact: <Defendant A>, born on April 17, ..., he was sentenced to six months' imprisonment for committing theft, and on January 8, 2018, he was released after completing his sentence. The trial found that: at 6:00 on September 16, 2018,..., <Defendant A> ... stolen and sold stolen. It was determined that the stolen cell phone was worth RMB 6,120. .... <Defendant A> was arrested on January 14, 2019 .

Sub-Views : Defendant A secretly stole RMB 8,000 ...constituted the crime of theft; Defendant A voluntarily surrendered to the police, ...lighter penalty.

Court View:

Figure 5: Prompting template for generating court views.

the given fact description f by passing it into a pre-trained encoder (*e.g.*, Lawformer (Xiao et al., 2021)):

$$\mathbf{h}_f = \text{Decomposer}(\texttt{[CLS]}\ f\ \texttt{[SEP]}), \quad (1)$$

where [CLS] and [SEP] are the special tokens, and $\mathbf{h}_f$ is the output embedding of the token [CLS]. Then, a linear classifier and a softmax function are applied to predict the legal concepts probability $\hat{\mathbf{p}} \in \mathbb{R}^{N_c}$, where the N_c is the number of legal concept types. We train this module by optimizing the legal concept classification loss that is formulated as:

$$\mathcal{L}_c = \mathbb{E}[-\sum_{n_c=1}^{N_c} \mathbf{p}(n_c|\mathbf{h}_f)\log(\hat{\mathbf{p}}(n_c|\mathbf{h}_f))], \quad (2)$$

where $\mathbb{E}$ represents the average expectation, and $\mathbf{p}(n_c|\mathbf{h}_f)$ denotes the ground-truth probability based on the fact representation $\mathbf{h}_f$. If n_c is the ground-truth concept, $\mathbf{p}(n_c|\mathbf{h}_f)$ equals to 1 otherwise 0. After obtaining the applicable legal concepts set $\hat{\mathcal{C}} = \{\hat{c}_1, \hat{c}_2, ..., \hat{c}_n\}$, we split the court view into n sub-views, where the i-th sub-view focusing on the i-th legal concepts.

Solver. Given the predicted legal concepts set $\hat{\mathcal{C}}$, the Solver module aims to generate candidate rationales for each sub-view independently. Specifically, given the i-th legal concept, we propose a legal concept enriched prompt to employ LLM to generate m rationales $r_i = \{r_i^1, r_i^2, ..., r_i^m\}$ independently. The prompting template is shown in Fig. 4. The prompt begins with a question to elicit the LLM to generate a rationale based on the given demonstrations containing fact-answer pairs. For prompt construction, it has been proved that similar cases contain more knowledge benefiting Legal AI task (Wei et al., 2024). Moreover, we believe that similar cases may share the same applicable legal concepts. Inspired by this, we first use the BM25 algorithm[2] to measure the semantic similarity between the fact of cases. Then, instead of directly selecting the top-k similar cases, we collect the top-k similar cases sharing the same legal concept set with the input case to enrich the prompt.

Verifier. Rationales r_i could be seen as the summary of the partial case fact that is related to the legal concept c_i. To ensure the consistency of case fact and generated rationale, we develop the verifier to filter the rationale uninformative or inconsistent. Specifically, we pass the case fact and rationale into a pre-trained encoder (*e.g.*, Lawformer (Xiao et al., 2021)) to obtain the representation for the j-th rationales of i-th concept:

$$\mathbf{v}_i^j = Verifier(\texttt{[CLS]}\ f\ \texttt{[SEP]}\ r_i^j\ \texttt{[SEP]}), \quad (3)$$

where [CLS] and [SEP] are the special tokens. The $\mathbf{v}_i^j$ is the output embedding of the token [CLS]. Then we apply a linear classifier followed by a softmax function to predict whether the rationale can be deduced from the case fact. We train the module

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

by optimizing classification loss. Specifically, we treat the fact f and rationale r_i derived from the same case as the positive pair denoted as (f, r_i), so $y = 1$. For the negative pair construction, given the f, we first retrieve the most similar case, but irrelevant to the legal concept c_i. We take the retrieved case fact and rationale r_i as a negative pair denoted as (f^-, r_i), $y = 0$ in this case. Moreover, we remove the defendant's name from the rationale r_i, which deters the model from learning shortcuts. The classification loss is formulated as follows:

$$\mathcal{L}_v = -\mathbb{E}[y \log(\hat{\mathbf{p}}(y|\mathbf{v}_i^j)) + (1-y)\log(\hat{\mathbf{p}}(y|\mathbf{v}_i^j))], \quad (4)$$

where y represents the ground-truth binary label for positive and negative pairs. $\hat{\mathbf{p}}(y|\mathbf{v}_i^j)$ denotes the predicted probability based on the representation $\mathbf{v}_i^j$. After collecting the predicted positive rationale with the highest probability for each sub-view, we obtain the generated rationales $\{\hat{r}'_1, \hat{r}'_2, ..., \hat{r}'_n\}$, where $\hat{r}'_i$ denotes the rationale for i-th sub-view.

Geneartor. After generating the generated rationales for all sub-views, the court view is obtained by incorporating the sub-views and case fact in the LLM. The prompting template is shown in Fig.5. The template begins with an instruction asking the model to generate court views based on the case fact and sub-views. We use similar cases retrieved in the solver module to enrich the prompt.

4 Experiment

4.1 LLMs

We first introduce the used LLMs in our work. **ChatGPT** (Ouyang et al., 2022), which is trained to follow instructions to adapt to human preferences. The version of `gpt-3.5-turbo` is used, and it is available from OpenAI API [3]. **GLM130B** (Du et al., 2022), an open bilingual (English & Chinese) bidirectional dense model with 130 billion parameters. The version of `glm-3-turbo` is used, and it is available from Zhipu API [4]. **GLM6B** [5] is an open bilingual language model based on GLM (Du et al., 2022) framework, with 6 billion parameters. **BaiChuan** (BaiChuan-Inc, 2023) is an open-source pre-trained large language model with 7 billion parameters [6].

[3]https://openai.com/
[4]https://open.bigmodel.cn/
[5]https://huggingface.co/THUDM/chatglm3-6b
[6]https://huggingface.co/baichuan-inc/Baichuan-7B

Item	LCVG	C3VG
# Train set	60,744	50,312
# Development set	20,257	-
# Test set	20,290	12,627
# Type of legal concept	101	62
# Avg. Legal Concept per case	2.1	-
# Avg. Length of fact	781.4	456.9
# Avg. Length of rationale	52.8	-
# Avg. Length of court view	245.9	276.8

Table 1: Statistics of LCVG and C3VG, where "#" denotes the number of data in the set.

4.2 Dataset and Evaluation Metrics

Dataset. As far as we know, the ground truth of the existing datasets ignore the legal concepts and corresponding rationales. Therefore, we construct a new dataset LCVG from CJO [7], the widely used government website in the Legal AI task. Following (Yue et al., 2021), we extract the defendant's background information, fact descriptions, legal concepts, and rationale of legal concepts using regular expressions. In this study, we only focus on the cases with one defendant, leaving the complex cases with multiple defendants in future work. As shown in Table 1, our dataset is split into the train set, development set, and test set, following a ratio of 3:1:1.

We evaluate our method on both our dataset LCVG and commonly used dataset C3VG (Yue et al., 2021), which are detailed in the datasets in Table 1. Due to the large size of the dataset (>12k for the test set of LCVG and C3VG), following Shui et al. (2023), we randomly a balanced small set from the test set of both datasets, which contains 522 cases and 536 cases, respectively. We denoted the sampled small set as LCVG*, and C3VG*, respectively. Moreover, we detail the construction of dataset LCVG and instruction datasets for fine-tuning LLMs in Appendix A.

Evaluation Metrics. To evaluate the quality of the generated court views, we employ ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BertScore (Zhang et al., 2020) as metrics. Following Yue et al. (2024), we report F1 scores of ROUGE-1, ROUGE-2, and ROUGE-L, and we keep the result of BLEU-1, BLEU-2, and BLEU-N (*i.e.*, an average score of BLEU-1∼4). While traditional metrics like ROUGE and BLEU are standard in evaluating text generation, they may not fully

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

Models	Rouge (%)			Bleu (%)			BertScore (%)			GPT4-Score
	Rouge-1	Rouge-2	Rouge-L	Bleu-1	Bleu-2	Bleu-N	Prec.	Recall	F1	
* LCVG:Traditional Methods with Fine-tuning Stage										
AttS2S	55.78	33.25	45.32	48.86	38.6	33.87	76.56	74.88	75.56	2.91
PGN	54.23	31.33	43.88	48.45	39.2	32.46	76.44	74.9	75.2	2.96
Transformer	55.45	30.2	42.45	49.05	34.43	30.78	76.09	73.76	75.87	2.87
Label-AttS2S	56.5	35.54	43.89	49.76	38.8	30.77	74.05	73.22	73.56	2.87
C3VG	57.2	37.45	**45.5**	50.12	40.52	**37.64**	**80.67**	77.02	78.78	3.01
LCVG:Prompt Methods with LLMs										
GLM6B	35.58	16.62	24.85	30.57	12.96	7.77	70.01	71.73	70.75	3.04
+LeGen-PT	41.54(+5.96)	22.93(+6.31)	28.33(+3.48)	34.88(+4.31)	13.47(+0.51)	10.97(+3.2)	71.76(+1.75)	75.91(+4.18)	73.66(+2.91)	3.17
GLM130B	40.67	21.53	28.27	30.46	16.0	10.07	71.36	76.44	73.69	3.42
+LeGen-PT	53.61(+12.94)	36.99(+15.46)	40.18(+11.91)	57.13(+26.67)	**41.84(+25.84)**	37.52(+27.45)	80.13(+8.77)	78.82(+2.38)	79.29(+5.6)	**3.97**
ChatGPT	44.89	17.79	29.01	52.95	25.87	15.62	75.83	73.36	74.5	3.79
+LeGen-PT	**57.82(+12.93)**	**38.48(+20.69)**	43.02(+14.01)	**57.56(+4.61)**	39.22(+13.35)	36.4(+20.78)	79.63(+3.8)	**80.67(+7.31)**	**80.56(+6.06)**	**4.13**
* C3VG:Traditional Methods with Fine-tuning Stage										
AttS2S	57.41	38.23	58.9	52.27	36.67	34.05	72.3	73.36	72.64	2.56
PGN	57.23	39.56	58.5	52.33	37.3	34.46	73.65	75.3	74.14	2.46
Transformer	61.05	40.67	58.45	51.86	39.4	35.78	75.43	74.05	74.97	2.81
Label-AttS2S	58.67	40.54	54.43	48.4	37.2	31.77	73.61	71.22	72.56	2.68
C3VG	62.12	**42.7**	**60.5**	60.78	**42.98**	**40.64**	76.45	73.45	74.78	2.82
* C3VG:Prompt Methods with LLMs										
GLM6B	36.23	17.56	24.43	29.76	14.45	10.67	71.23	72.33	71.89	3.13
+LeGen-PT	53.28(+17.05)	34.48(+16.92)	39.4(+14.97)	45.79(+16.03)	27.88(+13.43)	23.97(+13.3)	71.19(-0.04)	75.25(+2.92)	73.62(+1.73)	3.24
GLM130B	45.03	25.74	25.05	32.3	15.72	11.35	70.23	73.12	71.45	3.45
+LeGen-PT	59.5(+14.47)	42.01(+16.27)	49.3(+24.25)	60.71(+28.41)	40.8(+25.08)	35.3(+23.95)	82.4(+12.17)	**81.4(+8.28)**	**80.8(+9.35)**	**4.21**
ChatGPT	43.21	22.35	27.23	47.95	27.65	18.99	73.12	71.66	72.78	3.73
+LeGen-PT	**62.55(+19.34)**	42.35(+20.0)	56.6(+29.47)	**61.07(+13.12)**	41.88(+14.23)	37.83(+18.84)	**82.89(+9.77)**	79.34(+7.68)	80.33(+7.55)	**4.17**

Table 2: Overall performance on dataset LCVG* and C3VG*, where the "LeGen-PT" represents implementing LeGen by prompting LLM. The second-best score is underlined and the best is marked with **bold**.

capture the nuances of court view generation. Recent studies show that large language models can serve as evaluation methods, following Li et al. (2024b); Gao et al. (2024), we conduct score-based evaluations based on prompting GPT-4 (OpenAI, 2023b), which are detailed in Appendix B.

4.3 Experiment Setting

Baseline Models. We compare LeGen with the following two groups of baseline models: **(1) Traditional Methods** that are implemented based on GRU (Cho et al., 2014) or Transformer (Vaswani et al., 2017): **AttS2S** (Bahdanau et al., 2015), which is an attention-based sequence-to-sequence model. **PGN** (See et al., 2017), an attention-based model augmented with a hybrid pointer-generator network. **Transformer** (Vaswani et al., 2017), a widely used encoder-decoder framework. **Label-AttS2S** (Ye et al., 2018), a legal knowledge enhanced court views generation model. **C3VG** (Yue et al., 2021), which generates court views based on different parts of fact. **(2) Modern Methods** that are implemented based on pre-trained (large) language models: **C3VG-BART** (Yue et al., 2021), which implement C3VG with BART (Lewis et al., 2020) as the backbone. **LexiLaw** (Li, 2023), a LLM fine-tuned using court views generation dataset. **EGG** (Yue et al., 2024), which generates court views grounded by events with cooperative LLM (*i.e.*, BaiChuan (BaiChuan-Inc, 2023)).

Implement Details. We implement the baseline models based on the released source codes. For our framework LeGen, we implement the decomposer and the verifier by fine-tuning Lawformer (Xiao et al., 2021). We set the batch size, learning rate, dropout rate, warmup steps, and max length of fact as 16, 1×10^{-4}, 0.1, 500, and 1000, respectively. We implement the solver and the generator by prompting LLMs, and we set the top-k as 3. Moreover, we also implement the solver and the generator module by fine-tuning GLM6B on the instruction dataset using LoRA (Hu et al., 2022), a parameter-efficient fine-tuning method. Specifically, we set the rank, α, and dropout rate of LoRA as 16, 64, and 0.1, respectively. We set the max sentence length, max input sentence length, batch size, and learning rate as 1560, 1024, 4, and 1×10^{-5}, respectively. All the experiments are conducted on 2 Tesla A100 40G GPUs with AdamW (Loshchilov and Hutter, 2019) optimizer for 10 epochs.

4.4 Main Results

Table 2 shows the overall performance. **(1)** It is observed that our LeGen achieves the best performance in terms of the GPT4-Score, showing the effectiveness of the legal concept-guided method for court view generation. **(2)** As the prompt-based method, compared with other prompt-based methods that utilize the same LLM, our LeGen-PT achieves 14.01%, 21.9%, and 6.06% increase in terms of Rouge-L, Bleu-N, and F1 score of

Models	Rouge (%)			Bleu (%)		
	Rouge-1	Rouge-2	Rouge-L	Bleu-1	Bleu-2	Bleu-N
LCVG						
LexiLaw	49.39	25.07	36.00	48.81	27.96	26.59
C3VG-BART	73.11	62.50	59.70	55.70	49.23	46.05
EGG (BaiChuan)	73.20	66.21	60.30	56.35	52.60	50.9
LeGen-FT (GLM3)	**77.76**	67.80	**63.44**	**60.78**	**54.98**	51.88
LeGen-FT (BaiChuan)	75.93	**68.04**	63.23	59.93	54.50	**52.67**
C3VG						
LexiLaw	50.77	24.80	33.56	47.56	27.30	23.50
C3VG-BART	75.59	64.22	65.11	56.16	53.58	52.71
EGG (BaiChuan)	**76.86**	**65.15**	**65.90**	56.92	54.43	**53.59**
LeGen-FT (GLM3)	74.32	64.88	65.60	**57.03**	53.78	51.43
LeGen-FT (BaiChuan)	73.01	64.72	63.60	55.86	**54.89**	52.78

Table 3: Overall performance on LCVG and C3VG, where the "LeGen-FT" represents our LeGen implemented with a fine-tuned LLM (*i.e.*, GLM6B (Du et al., 2022)). The second-best score is underlined and the best is marked with **bold**.

BertScore. This illustrates the substantial impact of incorporating legal concepts into the model. **(3)** For prompt-based methods, most of them achieve worse performances compared with traditional methods, especially the most competitive method, C3VG. This is reasonable, as directly prompting an LLM does not require extensive training, whereas traditional methods benefit from large amounts of annotated data. Our method, however, achieves comparable results to C3VG and even outperformed it in Rouge-1, BLEU-1, and BertScores, which underscores the superiority of our approach.

Table 3 reports the performance of LeGen-FT and baselines based on pre-trained (large) language models. It is observed that our methods surpasses the prior state-of-the-art method EGG (Yue et al., 2024) by improving the Rouge-1, Rouge-2, Rouge-L, Bleu-1, and Bleu-2 by 4.56%, 1.59%, 3.16%, 4.43%, and 2.38%, respectively. However, our framework shows inferior performance on the C3VG dataset, which may lie in the absence of the defendant's background information (such as previous convictions and the date of release from prison) in the cases of C3VG. The defendant's background information plays an important role in legal concepts reasoning (such as `Recidivist`) according to the Criminal Law of China [8]. Therefore, we filter out cases involving the `Recidivist` and conduct experiments on the remaining cases. The results show that our framework achieves the best performance, please refer to Appendix C for more details.

[8]Any criminal who has been sentenced to fixed-term imprisonment or above and who commits another crime punishable by fixed-term imprisonment or above within five years after the execution of the sentence or pardon shall be deemed a **Recidivist** and shall be punished more severely, except for crimes committed by negligence and crimes committed by persons under the age of 18.

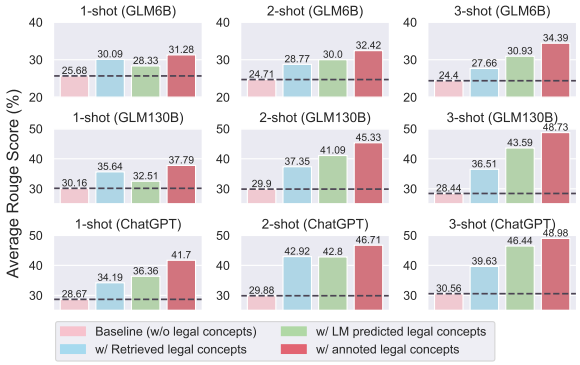

Figure 6: Model performance by different legal concepts.

4.5 In-Depth Analysis

To evaluate the efficacy of LeGen, we will answer the following research questions: **RQ-1**: Can legal concepts be accurately identified and effectively enhance the quality of generated court views? **RQ-2**: Can legal concepts effectively guide the generation of sub-views (rationals) for the criminal fact? **RQ-3**: How does LeGen perform in human evaluation? **RQ-4**: Why does LeGen generate the more precise court views with guidance of legal concepts?

Efficiency of Legal Concepts (RQ-1). In the *Decomposer* module, we obtain the legal concepts with two strategies, *i.e.*, information retrieval and fine-tuned language models; and our experiments on the LCVG dataset demonstrate recall rates of 83.32% and 87.68% for these strategies, respectively. The high recall achieved by the fine-tuned language model suggests that legal concepts can be accurately detected. In Fig. 6, we compared the average rouge scores using different legal concepts[9]. It is evident that the fine-tuned LM achieves superior results compared to retrieved legal concepts, underscoring the importance of high-quality legal concepts in enhancing model performance. However, there is still room for improvement compared to manually annotated legal concepts, indicating that further accuracy enhancements in legal concept detection could lead to better generation.

Efficiency of the Legal Concept-guided Rational Generation (RQ-2). In our method, given the legal concept, the *Solver* module generate corresponding rationales and the *Verifier* module select rationle with the highest align score. We report the performance of the solver module on the in-

[9]This experiment is conducted using three LLMs (*i.e.*ChatGPT,GLM130B,GLM6B) under few-shot settings, which are detailed in Appendix D

Fact Description		*Wang, born on 20 July 1969 in xx County, xx Province, …. Due to a sudden strong wind, the burning paper money was blown into the thatched grass near the grave, causing a forest fire in Yandouling Mountain Field, resulting in damage to the trees on the hill of responsibility of a group of villagers in Duijiang Village. …., surrender to the Xinning County Forest Public Security Bureau, truthfully confessed his crimes, apologized to the victim Huang Mou, and compensated him for his economic losses of 4,000 yuan, and obtained the victim Huang Mou's forgiveness, …*
Court View (Ground Truth)		*This court believes that the defendant Wang violated forest regulations …, causing a forest fire, resulting in a burned area of 2.78 hectares, all of which were forested land. His behavior constituted the crime of arson…, accusation is established. After the incident, the defendant Wang took the initiative to surrender and truthfully confessed his criminal facts. He surrendered himself and can be given a lighter punishment. He compensated the victim for his economic losses and obtained the victim's forgiveness, so he can be given a lighter punishment at his discretion, …, it is recommended that he be subject to community correction. This court decided to give the defendant Wang a lighter punishment and apply suspended sentence.*
LeGen-FT (Ours)	**Sub-Views (Decomposer)**	❶Why the Wang constitutes the crime of arson? ❷Why the Wang constitutes the Surrender? ❸Why the Wang constitutes the suspended sentence ?
	Rationales (Solver & Verifier)	❶ Wang accidentally caused a forest fire during the sacrificial ceremony, and his behavior constituted the crime of arson. ❷ Wang took the initiative to surrender to the public security organ after the incident and truthfully confessed his crime, which is a case of self-surrender and he can be given a lighter punishment. ❸ Wang surrendered himself and admitted his guilt after the incident. He actively compensated the victim for his economic losses and obtained forgiveness, showing a remorseful attitude, and was given a suspended sentence.
	Generated Court View (Generator)	This court believes that the defendant Wang accidentally caused a forest fire during the sacrificial ceremony, and his behavior constituted the crime of arson. The public prosecutor's charge …,guilty of arson is upheld. The defendant Wang took the initiative to surrender to the public security organs after the incident and truthfully confessed his crime, which is a circumstance of surrendering himself and can be given a lighter punishment. He actively compensated the victim for economic losses and obtained forgiveness,…, be given a lighter punishment at his discretion. Taking into account the defendant's surrender, confession attitude, and compensation behavior, it is in accordance with the law to apply community correction to him. Therefore, a suspended sentence was given to the defendant Wang in accordance with the law.
ChatGPT		This court believes that the source and form of the above evidence are legal, ... The evidence corroborates each other and the facts can be determined. Wang caused a forest fire due to negligence during the sacrifice, causing a large area of forest damage. His behavior constituted the crime of arson. However, in view of his active compensation for the victim's losses and obtaining forgiveness, and his remorseful behavior, this court decided to adopt the suggestion of the Judicial Bureau and apply community correction to Wang.
LexiLaw		This court believes that the defendant Wang caused a forest fire by arson, which was a relatively minor circumstance, and his behavior constituted the crime of arson. The facts of the … and the accusation is established. Based on the defendant Wang 's surrender and other criminal circumstances and his remorseful performance, combined with the community correction investigation and evaluation opinions, this court decided to impose a lighter punishment on the defendant Wang and apply suspended sentence .
EGG		This court believes that the defendant Wang caused a forest fire by burning paper money during the sacrifice, and his behavior constituted the crime of arson. Wang voluntarily surrendered himself to the police, truthfully confessed his crime, apologized to the victim Huang Mou and compensated him for his economic losses of 4,000 yuan, and obtained the victim Huang Mou's forgiveness, so he can be given a lighter punishment.

Figure 7: Case study. Examples of generated sub-views, rationales, and final court views by our method, compared with court reviews generated by other baseline methods. Text with a dark background represents legal concepts, and text with a light background represents the corresponding rationales.

Score	Fluency	Legal Professionality
1	Nonsense.	No Professionality. The generated texts are unrelated to all applicable legal concepts.
2	Very unfluent.	Almost Professionality. The generated texts are unrelated to almost all applicable legal concepts.
3	Partial fluent.	Half of them are Professional. The generated texts are related to about half of the applicable legal concepts.
4	Highly fluent.	Highly Professional. The generated texts are related to most of the applicable legal concepts.
5	Very fluent.	Exactly Professional. The generated texts are related to all applicable legal concepts.

Table 4: The criteria for human evaluation.

	LexiLaw	EGG	LeGen-PT	LeGen-FT
Fluency	4.36	4.47	**4.71**	4.58
Professionality	3.92	4.03	4.23	**4.55**

Table 5: The results of human evaluation.

struction dataset. Specifically, the solver based on fine-tuned GLM6B achieves 67.73%, 50.33%, and 85.79% in terms of the Rouge-L, Bleu-N, and F1 score of BertScore, respectively. The solver based on the ChatGPT achieves 49.42%, 29.23%, and 78.86%, respectively. The results prove the effectiveness of our LeGen-FT in terms of generating rationale for legal concepts. Moreover, we conduct an ablation study and find that, when removing the *Verifier*, there is 6.79%, 3.0%, 1.78% performance drop in terms of Rouge-L, Bleu-N, and F1 score of BertScore, respectively, indicating the effectiveness of *Verifier* in selecting better rationals.

Human Evaluation (RQ-3). To gain further insights, we conduct a human evaluation on the court view generated by our LeGen, EGG (Yue et al., 2024), and LexiLaw (Li, 2023). Specifically, we sample 200 cases and employ three legal experts to evaluate the generated court view. The evaluation adopts *Fluency* and *Legal Professionality* as metrics. Specific scoring criteria are provided in Table 4, with each metric scored on a scale of 1 (lowest) to 5 (highest). Table 5 shows that all methods achieve promising results in terms of *Fluency* due to the strong text-generation capability of LLMs. Additionally, our method outperforms other methods in terms of *Legal Professionality*, which proves the efficiency of the legal concept for the court view generation task. For a specific case, our method concentrates more on the explanations of the applicable legal concepts. This can lead to generating more professional and precise court views.

Case Study (RQ-4). Fig.7 shows an example of sub-views, corresponding rationales, and court views generated by our LeGen, which shows the high coherence with the ground truth court view in terms of the legal concepts and rationales. Please refer to Appendix E for more examples.

Moreover, we present the court views that is generated by strong baselines. We observe that the ChatGPT successfully generates the rationales for the `crime of arson` and `suspended sentence` but ignores the key `Voluntary surrender` and corresponding rationale. For LexiLaw, we find that although it mentions all the involved legal con-

cepts, the court view is absent of the rationale of the `Voluntary surrender`. For EGG, the generated court view mentions all the legal concepts but the rationales of the `Voluntary surrender` and `suspended sentence` are intertwined, which is less clear than our method.

5 Conclusion

In this study, we introduce a legal concept-guided court view generation framework (LeGen). Specifically, given the fact, The LeGen first divides the court view into several sub-views based on the predicted legal concepts. Then the solver and verifier are employed to generate and select rationales respectively. Finally, the court view is generated by incorporating the fact and rationales. Comprehensive experiments demonstrate the effectiveness of our proposed method.

Limitations

In this work, we propose a new legal concept-guided criminal court view generation framework LeGen. The limitation of the work is that LeGen splits the court view into sub-views based on the predicted legal concepts. Each sub-view is generated independently ignoring the relation of these sub-views. Although the strategy shows the advantage of the legal concept guidance, the potential to take full advantage of these legal concepts is still under-explored.

Ethics Statement

Each case in our dataset LCVG is obtained from the Chinese government website, with sensitive information appropriately anonymized to protect privacy. It is important to note that our work aims to automatically generate court view based on the case fact, which can alleviate the workload of legal professionals. Like many other LegalAI system, our framework may generate some uncontrollable content. Therefore, we emphasize that our method should only serve as an auxiliary tool in the legal field. The ultimate decision-making should always depend on legal professionals.

Acknowledgments

This work is supported by the Key Technology Partnerships Seed Funding Scheme. We thank the anonymous reviewers for their insightful comments.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR*.

BaiChuan-Inc. 2023. A large-scale 7b pretraining language model developed by baichuan-inc.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1724–1734.

Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. 2024. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *EECV*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*, abs/2306.16092.

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, EMNLP 2023, Singapore, December 6-10, 2023*.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and llms for legal case judgement summarization? In *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023) co-located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023)*, CEUR Workshop Proceedings.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391.

Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2022a. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. pages 15460–15475.

Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. 2022b. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning*, pages 6373–6391.

Frandberg. 1998. Legal concepts. Publisher: Taylor and Francis.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Congqing He, Tien-Ping Tan, Sheng Xue, and Yanyu Tan. 2023a. Explaining legal judgments: A multitask learning framework for enhancing factual consistency in rationale generation. *J. King Saud Univ. Comput. Inf. Sci.*, 35(10):101868.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. 2023b. Hanfei-1.0. https://github.com/siat-nlp/HanFei.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR*.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *CoRR*, abs/2305.15062.

Weijing Huang, Xianfeng Liao, Zhiqiang Xie, Jiang Qian, Bojin Zhuang, Shaojun Wang, and Jing Xiao. 2020. Generating reasonable legal text through the combination of language modeling and question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3687–3693.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*.

Haitao Li. 2023. Lexilaw: A large-scale 6b pretraining language model in legal domain.

Jinpeng Li, Hang Yu, Xiangfeng Luo, and Qian Liu. 2024a. Cosign: Contextual facts guided generation for knowledge graph completion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguisticsn (NAACL)*.

Quanzhi Li and Qiong Zhang. 2021. Court opinion generation from case fact description with legal basis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024b. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. Xiezhi chinese law large language model. https://github.com/LiuHC0428/LAW_GPT.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024. Prott3: Protein-to-text generation for text-based protein understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 5949–5966.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR*.

OpenAI. 2023a. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2023b. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Jaromír Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (GPT-4). *CoRR*, abs/2306.09525.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*.

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.

Yucheng Suo, Zhedong Zheng, Xiaohan Wang, Bang Zhang, and Yi Yang. 2024. Jointly harnessing prior structures and temporal consistency for sign language video generation. *ACM Trans. Multim. Comput. Commun. Appl.*, 20(6):185:1–185:18.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*.

Xiao Wei, Qi Xu, Hang Yu, Qian Liu, and Cambria Erik. 2024. Through the mud: A multi-defendant charge prediction benchmark with linked crime elements. In *Proceedings of the 2024 Conference of the Association for Computational Linguistics: ACL 2024*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning*.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 763–780.

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, EMNLP*, 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. *arXiv preprint arXiv:2105.03887*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL*, pages 13326–13365.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, pages 1854–1864.

Linan Yue, Qi Liu, Han Wu, Yanqing An, Li Wang, Senchao Yuan, and Dayong Wu. 2021. Circumstances enhanced criminal court view generation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1855–1859.

Linan Yue, Qi Liu, Lili Zhao, Li Wang, Weibo Gao, and Yanqing An. 2024. Event grounded criminal court view generation with cooperative (large) language models. *CoRR*, abs/2404.07001.

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. *CoRR*, abs/2309.11325.

Lidong Zeng, Zhedong Zheng, Yinwei Wei, and Tat-Seng Chua. 2024. Instilling multi-round thinking to text-guided image generation. *CoRR*, abs/2401.08472.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 9701–9708.

{ "Instruction": " You are a lawyer, please explain why the defendant constitutes {Legal Concept} ? ",
"Input": " {Fact}",
"Output": " {Rationale}"}

(a) Template for fine-tuning the solver module

{ "Instruction": " You are a judge, please summarize the court view.",
"Input": " {Fact}
{Sub-Views}",
"Output": "{Court View} " }

(b) Template for fine-tuning the generator module

Figure 8: Templates for fine-tuning the solver and generator modules

A Dataset Preparation

To support the Legal concept-enhanced criminal court view generation study, we construct a dataset with annotated legal concepts. Specifically, we first assign legal concepts to relevant sentences in the court view using regular expressions. Then we employed 20 Ph.D. students in law to perform the annotations and correct errors. The guidelines provided to the human annotators are as follows: (1) Review each annotated sentence and its legal concepts to verify accuracy. (2) Based on Step 1, provide corrected legal concepts if any annotations are inaccurate. (3) Check for any missed legal concepts in this court view. (4) Based on Step 3, provide the sentences and the corresponding legal concepts if any were missed. Finally, to assess inter-annotator agreement, we divided the 20 annotators into three groups. Each group was assigned the same set of 1,000 cases, randomly sampled from the LCVG dataset. Therefore, each of these 1,000 cases received 3 annotations from different annotators. We found that 99.6% of the cases had completely consistent annotations, indicating that this annotation task is very clear to the legal experts, with high annotation quality and strong inter-annotator agreement.

As shown in Figure 9, the distribution of legal concepts in LCVG has a long-tail distribution, which is in line with the real-world Chinese legal systems. We implement the legal concepts predictor model using the bm25 retriever (Wei et al., 2024) and small language model (Xiao et al., 2021). The results shown in Table 6 indicate that there is room for improvement. For the implementation of the solver and generator module of LeGen, we construct the instruction dataset based on the LCVG dataset, where the templates are shown in Fig.8.

Model	Acc.(Exact Match)	Precision	Recall	F1 Score
bm25	0.4934	0.754	0.8332	0.7916
Lawformer	0.714	0.914	0.8768	0.8950

Table 6: The results of legal concepts prediction on dataset LCVG.

Figure 9: The long-tail distribution of Legal concept of dataset LCVG.

B Evaluation with GPT4

Although traditional metrics such as ROUGE and BLEU are commonly used to evaluate text generation, they may fall short in capturing the subtle complexities of generating court opinions. Recent research suggests that large language models can be leveraged as effective evaluation tools. Following Li et al. (2024b); Gao et al. (2024), we conduct score-based evaluations based on prompting GPT-4 (OpenAI, 2023b), where the prompt template is shown in Figure 10.

C Cases in C3VG

As shown in the Fig.11, some cases in the C3VG overlook the defendant's previous convictions or the date of release from prison. These elements play a important role for legal concept reasoning according to the Criminal Law of China, as shown in Table 7. To provide a fair comparison, we filter out the cases labeled with *Article* 65 or *Article* 66. Since these cases involved the legal concept `Recidivist` according to the Criminal of China. After that, we evaluate our framework on the remaining 11675 cases. The overall performance is shown in Table 8. It is observed that our framework achieves the best results, showing the effectiveness of our LeGen with the help of legal concepts.

Article	**Legal Definition**
Article 65	Any criminal who has been sentenced to fixed-term imprisonment or above and who commits another crime punishable by fixed-term imprisonment or above **within five years after the execution of the sentence or pardon** shall be deemed a **Recidivist** and shall be punished more severely, except for crimes committed by negligence and crimes committed by persons under the age of 18.
Article 66	Criminals who commit crimes endangering national security, terrorist activities, or organized crime of a mafia nature, who commit any of the above crimes again at any time after the execution of their sentence or after being pardoned, shall be treated as repeat offenders.

Table 7: The definition of legal concept `Recidivist`.

Given the ground truth references court view: [*Ground Truth Court View*]
Given the model-generated court view: [*Generated Court View*]
Please score the quality of the generated text from 1 (worst) to 5 (best)

Figure 10: The template of prompting GPT-4 for score-based evaluation

Models	Rouge(%)			Bleu(%)		
	Rouge-1	Rouge-2	Rouge-L	Bleu-1	Bleu-2	Bleu-N
LexiLaw	49.88	25.3	32.78	47.16	29.74	25.73
C3VG-BART	74.66	64.47	64.83	58.32	55.34	53.95
EGG	77.34	65.73	66.82	57.81	56.29	52.17
LeGen-FT	**78.73**	**68.73**	**69.78**	**63.03**	**57.83**	**55.28**

Table 8: Overall performance on the filtered dataset C3VG. The best results are marked with **bold**.

Case without Defendant's Background Information in C3VG

"id": "270823 "
" fact " : " The trial found that the defendant Li Xi allowed Fu, Yu and Hua to take methamphetamine five times at his temporary residence ,... 1. On December 19, 2013, the defendant Li Xi provided money and asked Fu to help buy drugs...; 5. On a day in late January 2014, the defendant Li Xi provided drug-taking tools ...; while the defendant Li Xi was being investigated by the public security organs for taking drugs, he took the initiative to confess to the public security organs the criminal facts of allowing others to take drugs that had not yet been grasped. ..., the process of arrest and other evidence confirmed that it was sufficient to determine "
" court_view " : " This court believes that the defendant Li Xi allowed others to take drugs, ..., If he commits a crime punishable by fixed-term imprisonment or above within five years after the execution of the sentence, he is a **recidivist** and should be punished more severely. According to the ..."

Figure 11: An example of the case without the defendant's background information (*i.e.*, previous convictions and the date of release from prison), which is an important legal element for the legal concept (*i.e.*, *Recidivist*) reasoning

D Ablation Study

We conduct the ablation study under a few shot settings to illustrate the effectiveness of each module. The results are shown in the Table 9.

E Case Analysis

Given criminal facts, Table 10 provides more examples of generated sub-views, rationales, and final court views generated by our LeGen. we can observe that our method divide the court view into multiple sub-views and generates the rationale for each sub-view. Each rationale shows high coherence with the ground truth which leads to the more precise generated court view.

Setting	Models	Rouge			Bleu			BertScore		
		Rouge-1	Rouge-2	Rouge-L	Bleu-1	Bleu-2	Bleu-N	Prec.	Rcall	F1
1-Shot	ChatGPT	0.3894	0.1805	0.2903	0.4858	0.2264	0.1327	0.7437	0.7339	0.7379
	ChatGPT(RC/S/V)	0.3925	0.2872	0.346	0.4977	0.2331	0.1471	0.7517	0.7285	0.7391
	ChatGPT(PC/S)	0.4579	0.2703	0.3154	0.4719	0.2003	0.1199	0.735	0.7097	0.7214
	ChatGPT(PC/S/V)	0.4695	0.3028	0.3186	0.5061	0.2323	0.145	0.7462	0.7213	0.7328
	ChatGPT(OC/S/V)	0.5495	0.3228	0.3786	0.5147	0.3056	0.3658	0.757	0.7353	0.7451
2-Shot	ChatGPT	0.3904	0.2125	0.2935	0.5097	0.2486	0.1504	0.7505	0.7344	0.7416
	ChatGPT(RC/S/V)	0.558	0.3373	0.3923	0.4983	0.3137	0.208	0.7676	0.7641	0.765
	ChatGPT(PC/S)	0.527	0.3189	0.3719	0.5051	0.336	0.2697	0.7593	0.7441	0.7508
	ChatGPT(PC/S/V)	0.5439	0.334	0.4061	0.5511	0.3634	0.3205	0.788	0.772	0.777
	ChatGPT(OC/S/V)	0.5905	0.3723	0.4385	0.5723	0.3953	0.353	0.804	0.785	0.7987
3-Shot	ChatGPT	0.4489	0.1779	0.2901	0.5295	0.2587	0.1562	0.7583	0.7336	0.745
	ChatGPT(RC/S/V)	0.5466	0.2848	0.3575	0.4881	0.2633	0.1786	0.7647	0.7643	0.7636
	ChatGPT(PC/S)	0.5607	0.3544	0.3839	0.5556	0.3388	0.3076	0.778	0.7616	0.764
	ChatGPT(PC/S/V)	0.5782	0.3848	0.4302	0.5756	0.3922	0.364	0.7963	0.8067	0.8056
	ChatGPT(OC/S/V)	0.6001	0.416	0.4533	0.6121	0.4101	0.389	0.8183	0.8003	0.8154
1-Shot	GLM130B	0.4067	0.2153	0.2827	0.3046	0.16	0.1007	0.7136	0.7644	0.7369
	GLM130B(PC/S)	0.4154	0.2215	0.3082	0.3402	0.2009	0.1125	0.7172	0.76	0.7369
	GLM130B(PC/S/V)	0.421	0.2336	0.3208	0.3489	0.2207	0.1212	0.7218	0.7615	0.74
	GLM130B(RC/S/V)	0.45	0.2678	0.3514	0.3835	0.2364	0.1503	0.7372	0.7806	0.7573
	GLM130B(OC/S/V)	0.4799	0.2989	0.355	0.4809	0.3131	0.2481	0.7384	0.7816	0.7583
2-Shot	GLM130B	0.402	0.2175	0.2775	0.2864	0.1514	0.0966	0.7126	0.7552	0.732
	GLM130B(PC/S)	0.4371	0.3217	0.3524	0.4582	0.3016	0.2406	0.7307	0.7754	0.7511
	GLM130B(PC/S/V)	0.5206	0.3368	0.3753	0.4754	0.3297	0.2494	0.7435	0.7794	0.7545
	GLM130B(RC/S/V)	0.4671	0.3004	0.3531	0.4637	0.3032	0.2578	0.735	0.7916	0.7666
	GLM130B(OC/S/V)	0.5233	0.4005	0.436	0.5586	0.341	0.2732	0.7536	0.8042	0.7769
3-Shot	GLM130B	0.3937	0.2064	0.2532	0.2441	0.1268	0.0807	0.7025	0.7569	0.7278
	GLM130B(PC/S)	0.5005	0.3251	0.3646	0.545	0.3522	0.3019	0.789	0.7656	0.7705
	GLM130B(PC/S/V)	0.5361	0.3699	0.4018	0.5713	0.4184	0.3752	0.8013	0.7882	0.7929
	GLM130B(RC/S/V)	0.4664	0.2844	0.3444	0.4506	0.3145	0.1546	0.7439	0.7985	0.7692
	GLM130B(OC/S/V)	0.5945	0.425	0.4423	0.6056	0.4274	0.402	0.8233	0.8052	0.813
1-Shot	GLM6B	0.3558	0.1662	0.2485	0.3057	0.1296	0.0777	0.7001	0.7173	0.7075
	GLM6B(PC/S)	0.3701	0.1922	0.243	0.2188	0.1076	0.0621	0.687	0.7233	0.7116
	GLM6B(RC/S/V)	0.4082	0.2265	0.2679	0.2219	0.1211	0.0825	0.6949	0.7471	0.7188
	GLM6B(PC/S/V)	0.3856	0.2054	0.2588	0.2244	0.113	0.0729	0.693	0.7401	0.7145
	GLM6B(OC/S/V)	0.423	0.2273	0.2882	0.2724	0.151	0.1036	0.7115	0.7655	0.7363
2-Shot	GLM6B	0.3431	0.1561	0.2421	0.2999	0.1194	0.0706	0.6959	0.7092	0.7012
	GLM6B(PC/S)	0.3865	0.1978	0.2498	0.2032	0.1013	0.0553	0.6988	0.7474	0.7156
	GLM6B(RC/S/V)	0.3918	0.2036	0.2677	0.2156	0.1178	0.0706	0.7103	0.7596	0.7306
	GLM6B(PC/S/V)	0.4023	0.2189	0.2787	0.2233	0.1125	0.0865	0.7083	0.7527	0.7288
	GLM6B(OC/S/V)	0.445	0.2354	0.2923	0.3166	0.1216	0.1519	0.7359	0.7632	0.7376
3-Shot	GLM6B	0.34	0.1515	0.2405	0.2976	0.1134	0.0699	0.702	0.7086	0.7023
	GLM6B(PC/S)	0.396	0.2121	0.2828	0.2276	0.1134	0.0938	0.7075	0.7437	0.7139
	GLM6B(RC/S/V)	0.3718	0.1978	0.2601	0.2385	0.1215	0.08	0.7022	0.731	0.7148
	GLM6B(PC/S/V)	0.4154	0.2293	0.2833	0.2488	0.1347	0.1097	0.7176	0.7591	0.7366
	GLM6B(OC/S/V)	0.4565	0.253	0.3222	0.3021	0.1747	0.157	0.7587	0.7801	0.762

Table 9: The results of the ablation study where *PC*, *RC*, an *OC* denote obtaining legal concepts by using pre-trained language model, retriever, and human annotated legal concepts respectively. *S* and *V* denote the sovler and verifier respectively.

Fact-1	Tan Moujun, male, born on ...it was found that ..., Tan Jun and Yan Debing and others he summoned carried out high-altitude welding and construction without the qualification of steel structure construction and high-altitude operation operation certificate,..., It was also found that Tan Jun was notified by the public security organ to appear in court by phone, and after appearing in court, he truthfully confessed his criminal facts. It was further found that Tan Jun voluntarily signed a confession and confession of guilt...
Court View (Ground Truth)	This court believes that the defendant Tan Moujun and others violated relevant safety management regulations during construction work, and an accident occurred in which one person died. Their behavior constituted a major responsibility accident crime and should be punished. The People's Procuratorate of Shunqing District, Nanchong City, accused the defendant Tan Moujun of committing a major responsibility accident crime and the facts and charges were established. The defendant Tan Moujun was notified by the public security organ by phone to appear in court. After appearing in court, he truthfully confessed his crime and surrendered himself, which can be mitigated or reduced. The defendant voluntarily pleaded guilty and accepted punishment and can be given a lenient punishment. In view of the fact that the family of the deceased in this case has received compensation and expressed understanding of the defendant's behavior, this court will give full consideration when sentencing. The defense opinion raised by the defendant's defense counsel is supported by this court. In summary, according to the defendant Tan Moujun's criminal facts, nature, circumstances, and degree of harm to society, he can be given a suspended sentence according to law.
Sub-Views	Why did the defendant constitute the crime of major responsibility accident ? Why did the defendant constitute the Surrender? Why did the defendant constitute the suspended sentence?
Rationales	"When Tan Moujun was working at height, he violated safety management regulations, resulting in the death of construction worker Yan Debing and causing serious casualties. His behavior constituted a crime of major responsibility accident", "In view of the fact that Tan Moujun voluntarily surrendered after the incident and truthfully confessed his crime, it is a case of self-surrender", "Taking into account that Tan Moujun voluntarily surrendered after the incident and truthfully confessed the facts of the crime and that he had made civil compensation and obtained forgiveness from the defendant's close relatives after the incident, in order to reduce his criminal responsibility and encourage him to correct his mistakes, it is in accordance with the law to apply probation to him."
Court View (Generated)	This court believes that the defendant Tan Moujun violated safety management regulations when performing high-altitude operations, resulting in the death of construction worker Yan Debing and causing major casualties. His behavior has constituted a major responsibility accident crime. The charges brought by the public prosecution agency are established, and this court supports them. In view of the fact that the defendant Tan Moujun voluntarily surrendered after the incident and truthfully confessed his crimes, it is a case of self-surrender. Considering that the defendant Tan Moujun voluntarily surrendered after the incident and truthfully confessed the facts of the crime and that he had made civil compensation to the victim's close relatives and obtained forgiveness after the incident, in order to reduce his criminal responsibility and encourage him to correct his mistakes, it is in accordance with the law to apply probation to him. In summary, this court has made a suspended sentence judgment on the defendant Tan Moujun in accordance with the law.

Fact-2	Li Mouyun, citizen ID number XX, ..., it was found that one day in early December 2017, Li Mouyun bought two homemade rifles from a stranger for 450 yuan in the mountains of Goujiaqu Village, Yima Town, Qingcheng County. In January 2018, Li Mouyun used one rifle to hunt near his home...., In addition, at 12:00 on May 9, 2018, Li Mouyun took the initiative to go to the Yima Police Station of the Qingcheng County Public Security Bureau for investigation....
Court View (Ground Truth)	This court believes that the defendant Li Mouyun violated the state's regulations on the management of firearms and illegally possessed two non-military firearms that could be fired normally and had the power to cause injury. His behavior has violated the criminal law and constituted the crime of illegal possession of firearms, and the circumstances are serious. The public prosecutor's criminal charges against the defendant Li Mouyun are clear in facts, and the evidence is solid and sufficient. The charges are established and should be supported. The defendant Li Mouyun took the initiative to surrender and truthfully confessed the facts of the crime, which is a self-surrender and can be mitigated. According to the defendant's remorse, there is no risk of re-offending if a suspended sentence is applied to him, and a suspended sentence can be applied according to law.
Sub-Views	Why does the defendant constitute the crime of illegal possession of firearms? Why did the defendant constitute the Surrender? Why did the defendant constitute the suspended sentence?
Rationales	"Li Mouyun was previously sentenced to one year of probation for illegal possession of firearms, but during the probation period he still illegally possessed two homemade rifles, which shows that he had no respect for the law and a serious criminal mentality, posing a serious threat to social security", "Li Mouyun took the initiative to surrender to the public security organs after the incident and truthfully confessed his criminal facts of illegal possession of firearms, which is a circumstance of self-surrender", "Based on Li Mouyun's criminal facts, circumstances, harmful consequences, and remorse, combined with his lack of criminal record, this court believes that probation can be applied to him"
Court View (Generated)	This court believes that the defendant Li Mouyun illegally possessed two homemade rifles, and his behavior constituted the crime of illegal possession of firearms. The charges of the public prosecutor are established. The defendant Li Mouyun was previously sentenced to one year of probation for the crime of illegal possession of firearms, but he still illegally possessed firearms during the probation period, which shows that he has no respect for the law and a serious criminal concept, posing a serious threat to social security. The defendant Li Mouyun took the initiative to surrender to the public security organs after the incident and truthfully confessed his criminal facts of illegal possession of firearms, which is a circumstance of self-surrender. Based on the defendant Li Mouyun's criminal facts, circumstances, harmful consequences and remorse, as well as the fact that he has no record of illegal crimes, this court believes that the suspended sentence can be applied to him. The sentencing recommendation of the public prosecutor is appropriate.

Table 10: Examples of criminal court view generation process of our LeGen method.