

LegalChainReasoner: Grounding Criminal Judicial Opinion Generation via Structured Legal Chains

Weizhe Shi^{1*}, Qiqi Wang^{2*†}, Yihong Pan¹, Qian Liu^{1†}, Kaiqi Zhao^{3,4†}

¹School of Computer Science, University of Auckland, Auckland, New Zealand

²School of Statistics and Data Science, AAIS, LPMC and KLMDASR, Nankai University, Tianjin, China

³School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

⁴Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen, China

{wshi590, ypan317}@aucklanduni.ac.nz qiqi.wang@nankai.edu.cn

liu.qian@auckland.ac.nz zhaokaiqi@hit.edu.cn

Abstract

A criminal judicial opinion represents the judge’s disposition of a case, including the decision rationale and sentencing. Automatically generating such opinions can assist in analyzing sentencing consistency and provide judges with references to past similar cases. However, current research typically approaches this task by dividing it into two isolated subtasks: legal reasoning and sentencing prediction. This separation often leads to inconsistency between the reasoning and predictions, failing to meet real-world judicial requirements. Furthermore, prior studies rely on manually creating knowledge to enhance applicability, yet such methods remain limited in practical deployment. To address these limitations and better align with legal practice, we propose a new LegalAI task: Criminal Judicial Opinion Generation, which simultaneously produces both legal reasoning and sentencing decisions. To achieve this, we introduce the LegalChainReasoner framework that applies structured legal chains to guide the model through comprehensive case assessments. By integrating factual premises, composite legal conditions, and sentencing conclusions, our approach ensures flexible knowledge injection and end-to-end opinion generation. Experiments on real-world, open-source Chinese legal case datasets demonstrate that our method outperforms baseline models. ¹

1 Introduction

In criminal judicial systems, particularly within Civil Law jurisdictions, legal judgments inherently integrate two inseparable elements: structured legal reasoning and consequent sentencing decisions (Cui et al., 2023). This integration reflects the fundamental nature of judicial decision-making, where sentencing conclusions must be derived from

*Equal contribution.

†Corresponding author.

¹The code and data are available at <https://github.com/Statistical-NLP-Lab/LegalChainReasoner>.

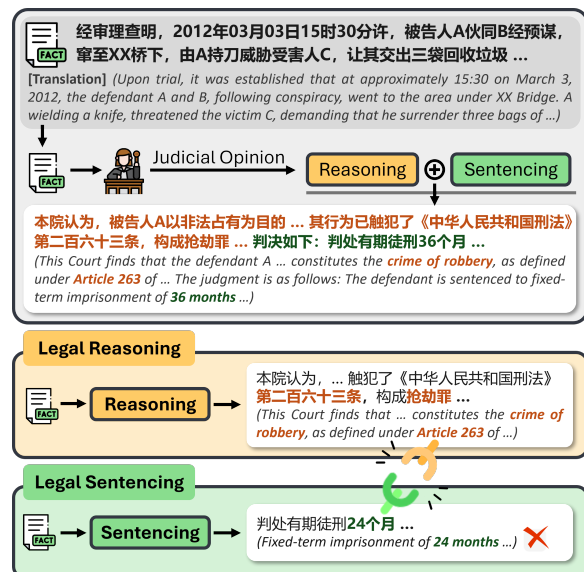


Figure 1: Traditional Approaches vs. Real-world Practice. Traditional methods model legal reasoning and sentencing separately, whereas practice treats them as a unified process where sentencing follows from (and is justified by) reasoning over case facts.

and justified by the systematic application of statutory provisions to established facts.

Despite this intrinsic connection, current LegalAI research artificially divides this coherent process into isolated tasks. The legal reasoning task (Ye et al., 2018; Wu et al., 2020; Li et al., 2024a; Yue et al., 2024; Mishra et al., 2025) focuses exclusively on constructing plausible arguments without addressing sentencing implications, while the sentencing prediction task (Yue et al., 2021a; Liu et al., 2022; Wu et al., 2023; Sun et al., 2024; Li et al., 2025) generates penalty terms without providing necessary legal justifications. This separation creates two critical problems: (1) reasoning models may produce analyses inconsistent with the possible sentencing outcome, and (2) prediction models generate sentences lacking transparent legal foundations. Such fragmentation fundamen-

tally contradicts the judicial principle that sentencing must be grounded in and justified by proper legal reasoning.

Beyond this task separation, prior work faces two further limitations. First, legal knowledge is often manually designed and simplistic (Gan et al., 2021), relying on predefined rules or direct statutory citations rather than abstracting essential reasoning patterns. This weakens deeper legal grounding and generalization across subtly different but legally similar cases, a challenge that has recently been framed more broadly as legal knowledge infusion for LLMs (Liu et al., 2026). Second, explainability remains limited, especially in sentencing prediction. Although some studies introduce manually defined generation processes for reasoning generation (Xu et al., 2024) or explore structure-aware legal text generation such as legal judgment summarization (Wang et al., 2023), they still do not jointly model judicial reasoning and sentencing. Nevertheless, most existing research on sentence prediction primarily focuses on extracting semantic features from factual information, overlooking the underlying reasoning mechanisms and the interpretability of the predicted outcomes. This limits the transparency and trustworthiness of these models, especially in the criminal domain.

To address these fundamental limitations, we propose a novel **Criminal Judicial Opinion Generation (CJOG)** task, wherein legal reasoning and sentencing prediction are integrated into one paradigm like human judges, as illustrated in Figure 1. This task offers two significant advantages: (1) sentencing predictions become explicitly grounded in legal reasoning, enhancing both interpretability and jurisprudential validity; and (2) the approach mirrors authentic judicial processes, where verdicts are derived through structured legal analysis rather than isolated numerical predictions.

To solve this task, we introduce **LegalChainReasoner**, a framework designed to ground generation in structured legal knowledge. It comprises two principal components: the structured Legal Chain and the Chain-Aware Encoding method. The structured Legal Chains are developed by transforming statutory provisions into *premise-situation-conclusion* triplets, based on the Legal Norm theory (Keuth, 1975; Raz, 1980). These chains formalize implicit legal inference patterns into explicit structured representations, enabling the model to learn step-by-step judicial decision-making processes. Furthermore, the Chain-Aware encoding

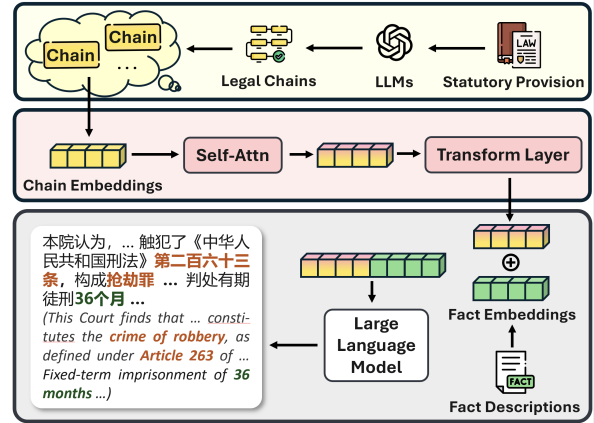


Figure 2: LegalChainReasoner Architecture. (top) Construct legal chains from statutory provisions; (middle) encode them via Chain-Aware Encoding; (bottom) fuse the resulting legal representations with case-fact embeddings in the LLM to generate judicial opinions.

utilizes multi-linear transformation layers to capture complex relationships between legal elements while applying crime-specific reasoning transformations, ensuring effective jurisprudential alignment with established judicial reasoning patterns.

The main contributions of this work are: (1) We define the CJOG task that produces legal reasoning and sentencing prediction in a single process, ensuring that predictions are supported by judicial analyses; (2) We propose LegalChainReasoner, a novel framework that formalizes statutory provisions into structured Legal Chains. By employing a Chain-Aware Encoding method, our approach effectively facilitates interpretable reasoning grounded in legal logic. (3) We evaluate our method on real-world datasets. Experimental results demonstrate that LegalChainReasoner achieves improvements in both the quality of generated judicial opinions and the accuracy of sentencing prediction.

2 Methodology

Task Definition. Given a set of factual descriptions of criminal cases $\mathcal{F} = \{F_1, F_2, \dots, F_i, \dots\}$, the CJOG task aims to generate a judicial opinion O_i for each fact F_i , by jointly considering legal reasoning and sentencing decision. Formally, we define the task as a function $f : F_i \xrightarrow{C_i} O_i$, where C_i denotes the criminal charge. In this work, we omit the charge prediction step, as it has already achieved remarkably high accuracy (>95%) in multiple established studies (Hu et al., 2018; Le et al., 2020; Zhao et al., 2022). Our focus is therefore on the CJOG task, given the charges. Additional

experimental evidence for this choice is presented in the Appendix A.

Overall Framework. The fundamental principle of our method lies in mirroring authentic judicial processes, producing comprehensive judicial opinions that jointly consider both reasoning and sentencing in accordance with established legal principles. Figure 2 presents the overall framework of the proposed method, LegalChainReasoner. The framework consists of the Legal Chain construction and the Chain-Aware Encoding modules. The Legal Chain construction addresses the need for formalized legal knowledge representation. By transforming legal provisions into structured *premise-situation-conclusion* triplets, complex legal conditions and their relationships are explicitly formalized, enabling the model to learn step-by-step judicial decision-making processes. The Chain-Aware Encoding module leverages these structured chains to guide the opinion generation process. This specialized mechanism is integrated with case fact embeddings before being passed to the language model. This architecture ensures that the CJOG process is explicitly grounded in proper legal analysis.

2.1 Legal Chain

In criminal law practice, judges typically derive case outcomes through a structured reasoning process. This process traditionally follows a syllogistic pattern (Dai et al., 2025), where *major premises* (statutory provisions) and *minor premises* (factual descriptions) are combined to reach judicial decisions, encompassing both legal reasoning and sentencing. However, judges do not apply statutory provisions in their raw form, as these provisions often contain complex structures, including multiple conditional clauses and nested relationships. To navigate this complexity, the legal norm theory (Keuth, 1975; Raz, 1980) offers a principle that decomposes each statutory provision into one or more triplets, each comprising a *premise*, a *situation*, and a *conclusion*. These triplets capture the constitutive elements of a crime and the conditional factors affecting sentencing. Judges typically select and apply the most relevant triplet during the reasoning process to reach a legally sound and context-appropriate judgment. Even in complex cases involving multiple crimes, the legal norm theory still applies: judges analyze each crime in accordance with the corresponding statutory provi-

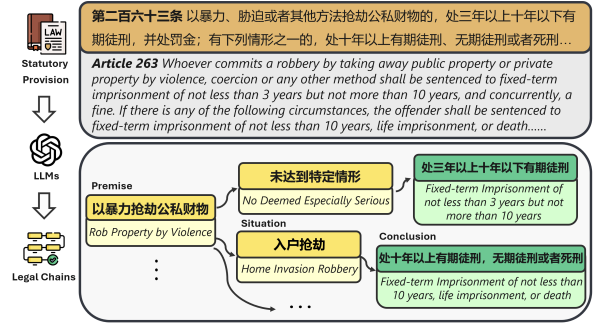


Figure 3: Legal Chain Construction. The statute (top) is decomposed into *premise-situation-conclusion* triplets (bottom) following Legal Norm Theory.

sions.

To align with real-world judicial practice, we propose using the triplet structure, *premise-situation-conclusion*, as a means of integrating legal knowledge into language models for generating judicial opinions. We refer to this triplet as a **Legal Chain**, which aims to formalize the implicit reasoning patterns embedded within statutory provisions into explicit, structured representations, enabling models to learn the step-by-step judicial decision-making process. Figure 3 illustrates an example of the Legal Chains for robbery.

Formally, for each criminal charge C , all relevant statutory provisions defined within the legal domain are denoted as $\mathcal{SP} = \{SP_1, SP_2, \dots, SP_j, \dots\}$. They are systematically decomposed into a set of Legal Chains $\Phi_C = \{\phi_1, \phi_2, \dots\}$, where each Legal Chain $\phi \in \Phi_C$ is represented as a triplet $\langle p, s, c \rangle$. Here p (*premise*) represents a condition derived from legal facts, typically characterizing a legal element or prerequisite that must be satisfied; s (*situation*) captures the consequences and severity assessments, reflecting nuanced legal requirements where multiple factors must be considered simultaneously; and c (*conclusion*) specifies a sentencing recommendation directly grounded in statutory provisions.

Current works utilize various forms of legal knowledge, such as legal graphs (Zhong et al., 2020; Yang et al., 2022), prediction processes (Xu et al., 2024), and explanations of statutory provisions (Yue et al., 2021a). However, most of these approaches rely on manually designed structures or expert annotations (Chalkidis et al., 2019), which significantly increase the cost of incorporating legal knowledge and may result in unstable model performance when the knowledge is incorrect or of low quality. To address this, we propose an auto-

matic extraction process for Legal Chains. This process transforms statutory provisions into structured triplets through a prompt-based extraction methodology using large language models. The extraction prompts are carefully designed with constraints to ensure that the resulting chains maintain both legal accuracy and structural consistency. The complete prompt template, extraction constraints, and failure analysis of the extraction process are provided in Appendix B. The generated Legal Chains are subsequently validated by legal experts to ensure their fidelity to the original statutory provisions. Then, for a factual description F_i and its corresponding criminal charge C_i , a set of Legal Chains $\Phi_{C_i} = \{\phi_1, \phi_2, \dots, \phi_n\}$ is constructed, where each ϕ_j represents a distinct decision-making pathway.

2.2 Chain-Aware Encoding

To better make models understand Legal Chains Φ_{C_i} , we propose a Chain-Aware Encoding module to leverage these chains for the CJOJ task.

For a given criminal charge C_i , each chain $\phi_j = \langle p_j, s_j, c_j \rangle$ in the set Φ_{C_i} is embedded by tokenizing each component and averaging its token embeddings, as mean pooling consistently outperformed alternative strategies (see Appendix C for detailed comparisons).

$$\mathbf{e}_{x_j} = \frac{1}{|x_j|} \sum_{t \in x_j} \mathbf{E}_{\text{base}}(t), \quad \text{for } x \in \{p, s, c\}, \quad (1)$$

where \mathbf{E}_{base} represents the token embedding function of language models. These component embeddings are then concatenated to form a chain representation $\mathbf{H}_{\phi_j} = [\mathbf{e}_{p_j}; \mathbf{e}_{s_j}; \mathbf{e}_{c_j}] \in \mathbb{R}^{3 \times d}$, where d is the model’s hidden dimension.

To capture interactions between the elements of each Legal Chain, we apply a self-attention mechanism:

$$\mathbf{A}_{\phi_j} = \text{MultiHeadAttn}(\mathbf{H}_{\phi_j}), \quad (2)$$

$$\text{MultiHeadAttn}(\mathbf{H}_{\phi_j}^h) = \text{Attn}(\mathbf{H}_{\phi_j}^h, \mathbf{H}_{\phi_j}^h, \mathbf{H}_{\phi_j}^h), \quad (3)$$

where $\mathbf{A}_{\phi_j} \in \mathbb{R}^{3 \times d}$ is the attended output representation, and h means heads of multi-head attention. A residual connection preserves both the original and attention-weighted information: $\mathbf{A}'_{\phi_j} = \mathbf{H}_{\phi_j} + \mathbf{A}_{\phi_j}$. This attended representation is then pooled to obtain a single vector representation $\mathbf{r}_{\phi_j} \in \mathbb{R}^d$ for each chain.

After obtaining the processed chain representation \mathbf{r}_{ϕ_j} , as different types of criminal charges require distinct reasoning patterns based on their

statutory provisions and jurisprudential traditions, inspired by the Mixture-of-Experts (MoE) (Mansournia and Ebrahimpour, 2014; Cai et al., 2025), we design a crime-specific transformation layer to capture specific information. In addition, we introduce a general transformation layer to extract shared, generalizable features. We refer to this module as the crime transformation block, which processes the chain representations. The general transformation first applies to all criminal charges:

$$\mathbf{T}_G(\mathbf{r}_{\phi_j}) = \mathbf{W}_{G_2} \cdot \text{ReLU}(\mathbf{W}_{G_1} \cdot \mathbf{r}_{\phi_j} + \mathbf{b}_{G_1}) + \mathbf{b}_{G_2}, \quad (4)$$

where $\mathbf{W}_{G_1}, \mathbf{W}_{G_2} \in \mathbb{R}^{d \times d}$ are weight matrices and $\mathbf{b}_{G_1}, \mathbf{b}_{G_2} \in \mathbb{R}^d$ are bias vectors. Then, for each criminal charge C_i , a dedicated transformation is applied to the generally transformed representation:

$$\mathbf{T}_{C_i}(\mathbf{T}_G(\mathbf{r}_{\phi_j})) = \mathbf{W}_{C_i} \cdot \mathbf{T}_G(\mathbf{r}_{\phi_j}) + \mathbf{b}_{C_i}, \quad (5)$$

where $\mathbf{W}_{C_i} \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_{C_i} \in \mathbb{R}^d$ are charge-specific parameters that encode reasoning patterns unique to each crime type.

These transformations are then combined through a gating mechanism to determine the appropriate balance between general and crime-specific reasoning:

$$\mathbf{G}(\mathbf{T}_G(\mathbf{r}_{\phi_j})) = \sigma(\mathbf{W}_g \cdot \mathbf{T}_G(\mathbf{r}_{\phi_j}) + \mathbf{b}_g), \quad (6)$$

$$\mathbf{t}_{\phi_j} = \mathbf{G}(\mathbf{T}_G(\mathbf{r}_{\phi_j})) \odot \mathbf{T}_{C_i}(\mathbf{T}_G(\mathbf{r}_{\phi_j})) + (1 - \mathbf{G}(\mathbf{T}_G(\mathbf{r}_{\phi_j}))) \odot \mathbf{T}_G(\mathbf{r}_{\phi_j}), \quad (7)$$

where σ is the sigmoid function that produces content-dependent interpolation weights.

After applying transformations and the gating mechanism, each chain representation is processed through a final fusion layer that combines the original and transformed representations:

$$\mathbf{f}_{\phi_j} = \mathbf{W}_{\text{fusion}} \cdot [\mathbf{r}_{\phi_j}; \mathbf{t}_{\phi_j}] + \mathbf{b}_{\text{fusion}}, \quad (8)$$

where $\mathbf{W}_{\text{fusion}} \in \mathbb{R}^{d \times 2d}$ and $\mathbf{b}_{\text{fusion}} \in \mathbb{R}^d$ are learnable parameters.

The final chain representation for all chains is then constructed:

$$\mathbf{E}_{\text{Chain}} = [\mathbf{f}_{\phi_1}; \mathbf{f}_{\phi_2}; \dots; \mathbf{f}_{\phi_n}] \in \mathbb{R}^{n \times d}, \quad (9)$$

where n is the number of chains in Φ_{C_i} .

2.3 Judicial Opinion Generation

The factual description F_i is embedded using the same embedding layer of the language model, and then integrated with the chain representations:

$$\mathbf{E}_{\text{Combined}} = [\mathbf{E}_{\text{Chain}}; \mathbf{H}_{F_i}] \in \mathbb{R}^{(n+l_{F_i}) \times d}, \quad (10)$$

where $\mathbf{H}_{F_i} \in \mathbb{R}^{l_{F_i} \times d}$ represents the embedded tokens of F_i and l_{F_i} is the sequence length. This integrated representation is processed by the language model to generate the judicial opinion \hat{O} :

$$\hat{O} = \text{LLM}(\mathbf{E}_{\text{Combined}}). \quad (11)$$

The entire framework is optimized to jointly consider both sentencing prediction accuracy and reasoning quality. The loss function is defined as:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \alpha \mathcal{L}_{\text{Reasoning}} + \beta \mathcal{L}_{\text{Sentencing}}, \\ \mathcal{L}_{\text{Reasoning}} &= -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C R_{i,c} \log(\hat{R}_{i,c}), \\ \mathcal{L}_{\text{Sentencing}} &= -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C S_{i,c} \log(\hat{S}_{i,c}), \end{aligned} \quad (12)$$

where α and β are weighting parameters, $\mathcal{L}_{\text{Reasoning}}$ represents the cross-entropy loss for judicial reasoning generation quality, and $\mathcal{L}_{\text{Sentencing}}$ denotes the cross-entropy loss for sentencing prediction accuracy. Additionally, R_i and \hat{R}_i represent the ground truth and predicted reasoning text respectively, while S_i and \hat{S}_i represent the actual and predicted sentencing outcomes. Both loss components use standard cross-entropy formulation to ensure joint optimization of reasoning coherence and sentencing precision.

3 Experiments

3.1 Experimental Setup

Datasets. In the legalAI domain, we use a publicly accessible benchmark for our CJOG setting constructed from Legal AI Challenge 2021 (LAIC-2021), which contains Chinese criminal cases with factual descriptions, court reasoning, and sentencing outcomes. To ensure stable training and fair comparison across crime types, we retain only crime types with more than 500 cases, resulting in 12 crime categories. Detailed statistics are provided in Appendix D. This filtering is an experimental design choice for stable comparison rather than a limitation of the framework. Since legal chains are automatically extracted from statutory provisions, the construction cost scales primarily with the number of relevant statutory provisions rather than the number of case instances. To further assess the effectiveness of our proposed approach, we collect a more complex and up-to-date set of cases of the same crime types from the People’s Court Case

Database², denoted as PCCD. The new dataset includes 100 representative cases (balanced across the 12 crime types) and is used for zero-shot validation, meaning it serves solely to evaluate the models trained on the LAIC-2021.

Baselines. Based on recent studies, we select two types of baselines for a fair comparison. For the CJOG Task, we evaluate several natural language generation models, including AttS2S (Bahdanau et al., 2015), BART (Lewis et al., 2020), Llama-3.2-3B (Team, 2024a), Qwen-2.5-7B (Team, 2024b), DeepSeek-R1-Distill-Qwen-7B (Team, 2025), Lawyer-Llama-13B-V2 (Huang et al., 2023) and DeepSeek-R1-Distill-Qwen-32B (Team, 2025), on the complete CJOG task. For sentencing prediction, we selected methods such as BERT (Devlin et al., 2019), fine-tuned BERT (BERT-FT), and ELECTRA (Clark et al., 2020); manually designed knowledge-based methods like NeurJudge (Yue et al., 2021a) and ML-LJP (Liu et al., 2023); and case-based methods such as R-Former (Dong and Niu, 2021).

Evaluation metrics. Since the CJOG task primarily involves generating text, it can be evaluated using standard and widely adopted text generation metrics, such as ROUGE-1/2/L (Lin, 2004), BLEU-1/2/N (Papineni et al., 2002), BERTScore (Zhang et al., 2020a), and GPTScore-pairwise (Li et al., 2024b). Additionally, as sentencing prediction involves numerical outputs, we employ a pair of commonly used regression evaluation metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Further details are provided in Appendix E.

Implementation. We instantiate LegalChainReasoner on top of Llama-3.2-3B, Qwen-2.5-7B, DeepSeek-R1-Distill-Qwen-7B, Lawyer-Llama-13B-V2 and DeepSeek-R1-Distill-Qwen-32B. We fine-tune these LLM backbones with Low-Rank Adaptation (LoRA) (Hu et al., 2022), using rank $r=64$, scaling factor $\alpha=128$, dropout 0.2, and target modules q_proj, k_proj, v_proj, and o_proj. The base models are loaded in bfloat16 precision. In the Chain-Aware encoder, we use 8 attention heads with 0.1 dropout. As the $\mathcal{L}_{\text{Reasoning}}$ already include the sentence part, to enhance the sentencing performance, we found $\alpha = \beta = 1$ is the best.

²The People’s Court Case Database (<https://rmfyalk.court.gov.cn>) is operated by the Supreme People’s Court of China, with cases selected as references for adjudicating similar cases by lower courts.

Table 1: Results on LAIC-2021. Comparison across prediction and generation metrics (GPTScore represents the GPTScore-pairwise, and the GPTScore-pairwise scores for AttS2S and BART reflect their average performance across all comparisons). LegalChainReasoner performs best across backbones.

Model	MAE	RMSE	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-N	BERTScore	GPTScore
AttS2S	20.86	30.96	42.93	25.24	39.04	28.25	20.34	19.12	72.72	0.00
BART	28.99	41.68	51.73	35.95	46.72	39.99	32.93	31.35	79.68	17.70
Llama-3.2-3B	17.29	28.22	45.38	27.59	40.39	33.05	25.31	23.81	78.14	4.10
+ Statutory provisions	18.88	30.99	50.56	32.25	45.05	35.52	28.28	26.75	79.63	14.90
+ LegalChainReasoner	15.41	25.33	52.82	34.91	48.00	37.50	30.42	28.85	80.42	52.05
Qwen-2.5-7B	20.57	31.39	49.48	30.78	43.69	38.32	29.80	28.15	79.17	16.00
+ Statutory provisions	18.11	30.38	51.20	30.90	44.19	40.73	31.31	29.70	78.87	24.00
+ LegalChainReasoner	15.61	26.72	51.85	31.57	44.94	41.63	32.10	30.46	79.18	42.00
DeepSeek-R1-Distill-Qwen-7B	22.47	32.06	47.64	29.44	42.61	35.91	27.90	26.16	78.87	6.61
+ Statutory provisions	18.01	28.92	50.34	31.68	44.71	37.52	29.52	27.87	79.26	20.75
+ LegalChainReasoner	15.74	26.17	53.39	34.59	47.71	42.35	33.77	31.94	80.48	50.96
Lawyer-Llama-13B-V2	16.98	27.14	52.53	34.30	47.93	40.05	32.07	30.06	80.53	13.97
+ Statutory provisions	14.73	25.11	54.96	36.07	49.95	42.30	34.03	32.03	81.19	25.27
+ LegalChainReasoner	13.54	24.05	55.99	37.28	50.65	44.31	35.89	33.89	81.35	50.88
DeepSeek-R1-Distill-Qwen-32B	16.75	26.79	54.51	36.72	47.54	43.53	35.32	33.41	81.23	24.00
+ Statutory provisions	14.20	24.87	56.48	37.88	48.19	45.95	37.28	35.38	81.43	26.00
+ LegalChainReasoner	13.85	23.60	56.19	37.61	48.03	46.60	37.68	35.76	81.33	40.00

Table 2: Zero-shot Results on PCCD (generalization to complex cases). Comparison across prediction and generation metrics (GPTScore represents the GPTScore-pairwise, and the GPTScore-pairwise scores for AttS2S and BART reflect their average performance across all comparisons). LegalChainReasoner is most robust.

Model	MAE	RMSE	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-N	BERTScore	GPTScore
AttS2S	39.13	52.96	27.20	12.54	22.85	19.28	12.17	11.88	67.02	1.00
BART	62.53	77.33	29.54	15.84	23.51	24.12	16.45	15.86	70.10	19.40
Llama-3.2-3B	38.85	52.13	28.45	12.89	21.93	21.16	13.28	12.69	68.50	1.00
+ Statutory provisions	26.53	37.24	32.76	16.13	25.88	24.04	16.22	15.50	70.17	31.00
+ LegalChainReasoner	24.96	36.59	32.96	16.56	26.23	24.25	16.39	15.70	70.18	32.00
Qwen-2.5-7B	49.32	65.49	28.92	12.16	20.44	24.93	14.92	14.39	68.85	20.00
+ Statutory provisions	36.92	53.62	33.71	14.67	24.09	28.74	17.84	17.18	70.11	26.00
+ LegalChainReasoner	32.39	47.34	33.24	14.93	24.10	28.64	17.99	17.26	70.21	35.00
DeepSeek-R1-Distill-Qwen-7B	40.71	56.41	28.13	11.91	20.60	21.75	13.06	12.47	68.11	8.00
+ Statutory provisions	37.08	50.12	31.26	13.71	22.44	25.97	16.01	15.34	69.37	17.00
+ LegalChainReasoner	33.43	46.52	33.95	16.45	25.68	29.09	18.99	18.19	70.24	45.00
Lawyer-Llama-13B-V2	30.57	43.84	31.34	14.97	24.40	26.25	16.86	15.79	69.35	17.00
+ Statutory provisions	27.51	39.50	33.33	16.43	26.01	26.67	17.61	16.57	69.95	27.00
+ LegalChainReasoner	24.96	38.03	33.90	16.84	26.03	27.71	19.39	17.40	70.07	43.00
DeepSeek-R1-Distill-Qwen-32B	34.44	49.84	31.02	14.74	23.67	26.95	17.27	16.49	69.92	24.00
+ Statutory provisions	29.94	43.77	34.45	16.16	23.76	28.62	18.71	17.85	70.93	34.00
+ LegalChainReasoner	29.65	45.36	35.86	17.58	26.69	30.76	20.34	19.40	70.97	38.00

We apply Adam optimizer with learning rates at 10^{-5} level. All the experiments were conducted on NVIDIA A100 GPUs with 80GB memory.

3.2 Comparison among Criminal Judicial Opinion Generation

Tables 1 and 2 compare the proposed method with baselines on LAIC-2021 and PCCD for CJOG.

With LLM fusion, LegalChainReasoner achieves the best overall results in both text generation and sentence prediction across the datasets. A performance gap is observed between LAIC-2021 and PCCD across all methods, characterized by higher prediction errors and lower generation quality scores. There are some reasons: (1) PCCD

contains more complex cases, as it is published to guide judges; (2) it includes a higher proportion of appellate and final ruling cases involving multiple legal considerations; and (3) these cases need more complex legal reasoning.

When incorporating legal knowledge into LLMs, our Legal Chain fusion outperforms statute-only conditioning and vanilla LLMs. While fine-tuning legal-domain LLMs can help, it is not always efficient: integrating our Legal Chain approach with general-purpose LLMs surpasses the fully fine-tuned Lawyer Llama. Furthermore, when scaling to a larger 32B backbone, LegalChainReasoner continues to deliver improvements, indicating that structured chain injection remains complementary

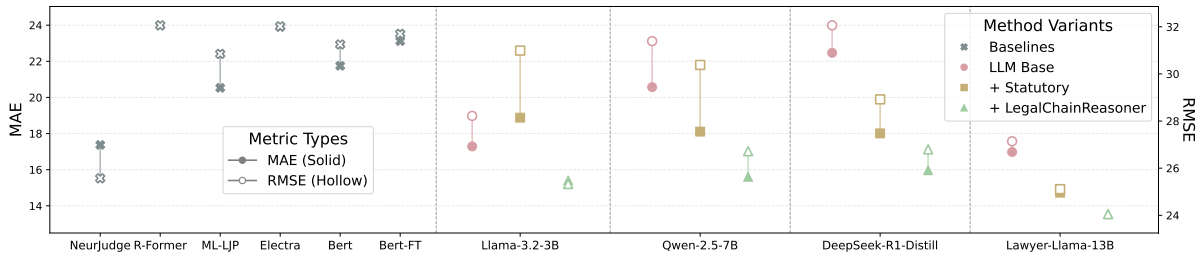


Figure 4: Comparison on Sentencing Prediction. Visual comparison between our unified approach and specialized sentencing prediction models. The left section illustrates the performance trend of existing baselines, while the right sections demonstrate the consistent improvements of the LegalChainReasoner (red triangles) across different LLM backbones compared to base models and statutory augmentation. Lower MAE (solid) and RMSE (hollow) indicate better performance.

to stronger base models rather than being eliminated by model scale alone.

These results indicate robustness across diverse legal reasoning scenarios and show that *premise-situation-conclusion* triplets effectively formalize judicial reasoning. We further assess semantic consistency and statutory compliance via Legal Question Assessment (Legal-QA) with LLMs as detailed in Appendix F, LegalChainReasoner consistently leads in statutory compliance, sentencing appropriateness, consistency, and completeness, significantly outperforming baselines in pairwise comparisons.

3.3 Comparison among Legal Sentencing Prediction

Figure 4 compares our CJOG-based approach with baselines that focus solely on sentencing prediction. Overall, CJOG-based methods achieve lower MAE/RMSE, suggesting that treating sentencing prediction in isolation is less effective. With LLM integration, LegalChainReasoner yields the best MAE and RMSE, highlighting the value of structured legal knowledge for accurate sentencing prediction while also providing high-quality legal reasoning.

Among specialized predictors, NeurJudge performs best, followed by ML-LJP, whereas the remaining models lag behind. This ranking differs from original reports, likely because we formulate sentencing prediction as regression rather than classification: discretizing numeric outcomes into bins can obscure finer-grained differences and complicate cross-setting comparisons.

It is noteworthy that the simple incorporation of statutory provisions without structured reasoning chains produced inconsistent results. In some cases, such as with Llama-3.2-3B, direct addition

Table 3: Sentencing Prediction on a CAIL2018 Subset. Results on an evaluation subset covering the same crime categories.

Model	MAE	RMSE
AttS2S	20.43	26.29
BART	26.28	35.90
Llama-3.2-3B	17.63	26.31
+ Statutory Provisions	18.00	27.89
+ LegalChainReasoner	14.69	24.04

of statutory provisions actually degraded sentencing accuracy. This observation underscores the insufficiency of merely exposing models to legal text without proper structuring of the reasoning process.

Generalization on the CAIL2018 Dataset. To further validate the generalization of LegalChainReasoner, we conduct additional experiments on CAIL2018 (Xiao et al., 2018), a widely used Chinese legal benchmark. Since CAIL2018 provides fact descriptions, accusations, and prison terms, but does not contain reference judicial opinions, it is not suitable for full CJOG evaluation, but is appropriate for evaluating sentencing prediction. For compatibility with our framework, we use the fact description as input, map the accusation label to the corresponding crime categories, and treat the term of imprisonment as the numerical target. We construct an evaluation subset by randomly sampling 10 cases for each of the 12 crime categories, resulting in 120 cases in total. Models are trained on LAIC-2021 and directly evaluated on this CAIL2018 subset without additional fine-tuning.

As shown in Table 3, LegalChainReasoner achieves the lowest MAE and RMSE among the included baselines in this CAIL2018 setting, indi-

Table 4: Ablation on Knowledge Integration Strategies. LegalChainReasoner outperforms RAG-based concatenation and the variant without crime-specific transformations (MoE); GPTScore denotes GPTScore-pairwise.

Model	MAE	RMSE	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-N	BERTScore	GPTScore
Llama-3.2-3B	17.29	28.22	45.38	27.59	40.39	33.05	25.31	23.81	78.14	8.79
+ RAG (Statutory provisions)	18.88	30.99	50.56	32.25	45.05	35.52	28.28	26.75	79.63	17.49
+ RAG (LegalChain)	16.96	29.01	48.14	29.66	42.44	34.49	26.74	25.27	78.53	9.46
+ LegalChain w/o MoE	15.67	26.87	50.70	32.30	45.11	37.61	29.67	28.08	79.54	18.49
+ LegalChainReasoner	15.41	25.33	52.82	34.91	48.00	37.50	30.42	28.85	80.42	64.26

Table 5: Rule-Based Case Study Evaluation. Assessing generated opinions across defendant identification, situation description, and sentencing accuracy. Our method achieves the highest combined consistency score.

Model	Defendant Accuracy (%)	Situation Accuracy (%)	Sentencing Accuracy (%)	Combined Performance Score (%)
AttS2S	8.45	42.26	76.15	2.72
BART	99.50	65.27	12.22	7.93
Lawyer-Llama-13B-V2	99.08	56.82	71.30	40.12
+ Statutory provisions	99.41	63.18	74.39	46.66
+ LegalChainReasoner	99.41	67.20	78.49	52.39

cating that it generalizes beyond LAIC-2021 and PCCD to another benchmark.

3.4 Ablation Study

We ablate knowledge integration strategies in our framework using Llama-3.2-3B as the base model, results are in Table 4. We compare: (1) LLM + RAG (Statutory Provisions): concatenate raw statutory text as prompts; (2) LLM + RAG (LegalChain): concatenate text-based Legal Chains as prompts (no specialized encoding); (3) LLM + LegalChainReasoner w/o MoE: Chain-Aware Encoding without the crime-specific transformation layers (MoE); (4) LLM + LegalChainReasoner: full model with Legal Chains and the complete Chain-Aware Encoding module.

LegalChainReasoner consistently surpasses the RAG-LegalChain baseline on all metrics, indicating that the proposed chain-aware encoder provides gains beyond simple text injection. The best performance is obtained only with all components: using statutory provisions alone leads to legal-knowledge overload (better generation but worse sentencing prediction), while plain-text Legal Chains yield more balanced results via the *premise-situation-conclusion* structure. With Chain-Aware Encoding (and MoE), the model captures deeper interdependencies among legal elements, achieving the highest scores in both generation and prediction.

3.5 Case Study

Rule-based Screening. We use a rule-based screening protocol to evaluate generated opinions along defendant identification, situation description, and sentencing-range appropriateness (details in Appendix E.5). As shown in Table 5, most models are accurate on defendant identification, while situation and sentencing exhibit larger gaps. LegalChainReasoner achieves the strongest situation and sentencing accuracy, yielding the best combined score.

Case Demonstration. Figure 5 and Appendix G provide examples across three crime types; Figure 5 shows only summaries, with full outputs in Table 10. AttS2S misclassifies the case as fraud (not robbery), identifies the defendant as “Li XX,” and misses key elements (violence/threats), although its sentence (36 months) is close to the gold (42 months) by chance. BART identifies the defendant and charge correctly but predicts 30 months, below the statutory minimum for robbery (Chinese Criminal Law Art. 263: 3–10 years), indicating weak statutory compliance without explicit legal reasoning. In contrast, LegalChainReasoner is coherent across dimensions: it identifies the defendant and charge, captures violence/threats, and predicts 48 months, slightly above the gold but aligned with the reasoning and case severity.

Failure Patterns. Beyond the overall gains, the remaining errors of LegalChainReasoner are relatively systematic. In the representative cases in Appendix G, our model usually identifies the correct charge-specific article in the Criminal Law *Special Part*, but may omit routinely cited *General Part* provisions used for sentencing formalities, especially when fines, recidivism, confiscation, or mitigation rules are involved. This is consistent with our current chain construction, which primarily covers charge-specific *Special Part* provisions rather than the full *General Part*. By contrast, several baselines show less stable behaviors, including irrelevant article substitution and over-citation of

Case Facts: <i>The court ascertained that on January 2, 2012, at approximately 22:30, defendant Fu XX, together with Cui Zhongyu and Ren Zhenguo (both already sentenced), drove a black Honda sedan to the Shell gas station in Tangerlizhen, Bazhou City. Wearing masks and armed with knives and pickaxe handles, they threatened [staff] and robbed over 300 yuan in cash from the station. At approximately 1:00 on January 3, the aforementioned individuals drove to the Shell gas station on East Ring Road in Bazhou City and used the same methods to rob over 1,000 yuan in cash. The stolen money has been spent.</i>			
Model	Defendant	Reasoning	Sentencing
AttS2S	[Incorrect] Incorrectly identify.	[Misclassified] Fraud instead of robbery.	[Inconsistent] Sentencing inconsistent with reasoning.
BART	[Correct] Correctly identify.	[Correct] Used violence and threats to rob.	[Inconsistent] Sentencing inconsistent with reasoning and actual sentencing range.
Lawyer-Llama-13B-V2	[Correct] Correctly identify.	[Correct] Used violence and threats to rob.	[Consistent] Sentencing consistent with reasoning.
+ Statutory Provisions	[Correct] Correctly identify.	[Correct] Used violence and threats to rob.	[Inconsistent] Sentencing inconsistent with reasoning and actual sentencing range.
+ LegalChain Reasoner	[Correct] Correctly identify.	[Correct] Used violence and threats to rob.	[Consistent] Sentencing consistent with reasoning and actual sentencing range (most close).

Figure 5: Case Study Overview. LegalChainReasoner best preserves fact–reasoning–sentencing consistency, while baselines often contradict. Colors: green/yellow/red = correct/partly/inaccurate.

unrelated provisions. This suggests that the remaining errors of LegalChainReasoner are mainly due to the current coverage of the chain set rather than uncontrolled hallucination.

4 Related Work

Legal Reasoning Generation. Several studies have formalized legal reasoning as a text-to-text natural language generation task, proposing various approaches to enhance performance. Early work applies sequence-to-sequence models to generate legal reasoning (Ye et al., 2018). Subsequent research leveraged pre-trained language models such as BART (Lewis et al., 2020), including versions fine-tuned on legal datasets (Yue et al., 2021b), as well as large language models (Xu et al., 2024; Yue et al., 2024). Beyond language model applications, recent studies have incorporated different forms of legal knowledge, either implicitly or through explicit processing. This includes multi-step reasoning (Wu et al., 2022; Xu et al., 2024) and legal provisions understanding (Li et al., 2024a; Yue et al., 2024). However, integrating preprocessed legal knowledge presents several challenges. First, reliance on human experts for annotation and validation incurs high costs, limiting scalability. Moreover, preprocessing steps may introduce biases or oversimplifications, reducing the robustness of gen-

erated reasoning, especially using LLMs.

Legal Sentencing Prediction. Legal sentencing prediction in criminal domains targets predicting prison terms. Several types of models are used in this task, including word-to-embedding methods like the Word2Vec-based approach combined with legal corpora (Xia et al., 2019), RNN-variants (Xu et al., 2020), language models (Zhang et al., 2023; Sun et al., 2024) and LLMs (Deng et al., 2024). The knowledge used in this area can also be divided into explicit and processed categories. The processed knowledge includes explanatory statutory provision (Yue et al., 2021a) and statute embeddings (Meng et al., 2025), similar cases retrieved through Case Triple Modeling (Zhang et al., 2023).

5 Conclusion

This paper introduces CJOG to align legal reasoning and sentencing with actual judicial workflows. We propose LegalChainReasoner, which models statutes as structured Legal Chains and uses Chain-Aware Encoding to capture legal logic and charge-specific reasoning. Experiments on public Chinese criminal-case datasets show that LegalChainReasoner improves both reasoning quality and sentencing accuracy, whether evaluated on the full CJOG task or a single task. This work marks a step forward in building models that generate judicial opinions with stronger legal coherence and real-world applicability.

Limitations

While LegalChainReasoner demonstrates significant improvements in CJOG, there are two limitations in evaluation methodology and temporal considerations.

While our automated evaluation using LLMs provides scalability, it may not capture all nuances of legal reasoning that human judges would consider. The combined accuracy metric, though comprehensive, treats all three components (defendant identification, situation description, and sentencing range) as equally important, which may not reflect actual judicial priorities.

The dataset used in this study reflects legal practices and interpretations from a specific time period. Legal frameworks evolve over time through legislative amendments and changing judicial interpretations. The model’s ability to adapt to such temporal shifts in legal understanding without re-training remains unexplored.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. 63261068) and the Shenzhen Science and Technology Program (No. SYSPG20241211173609009).

Ethical Considerations

Privacy Statement. All cases used in this paper are publicly accessible and have been published through China Judgment Online or the People’s Court Case Database. All personally identifiable information has been anonymized, as demonstrated in the case studies.

Usage Statement. This work is intended solely for research purposes. We have no intention of applying the current version in the real court to judge or influence people. While our proposed method outperforms baseline models, it still contains errors and requires further improvement.

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 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. [A survey on mixture of experts in large language models](#). *IEEE Trans. Knowl. Data Eng.*, 37(7):3896–3915.
- Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. [Neural legal judgment prediction in english](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4317–4323. ACL.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: pre-training text encoders as discriminators rather than generators](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. [A survey on legal judgment prediction: Datasets, metrics, models and challenges](#). *IEEE Access*, 11:102050–102071.
- Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2025. [Laiw: A chinese legal large language models benchmark](#). In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10738–10766. ACL.
- Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024. [Enabling discriminative reasoning in llms for legal judgment prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 784–796. ACL.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL.
- Qian Dong and Shuzi Niu. 2021. [Legal judgment prediction via relational learning](#). In *SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 983–992. ACM.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. [Gptscore: Evaluate as you desire](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6556–6576. ACL.
- Leilei Gan, Kun Kuang, Yi Yang, and Fei Wu. 2021. [Judgment prediction via injecting legal knowledge into neural networks](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12866–12874. AAAI Press.
- 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 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. [Few-shot charge prediction with discriminative legal attributes](#). In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 487–498. ACL.
- 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.

- Herbert Keuth. 1975. [On some logical characteristics of legal norms](#). *Jurimetrics Journal*, 15(3):160–170.
- Yuquan Le, Congqing He, Meng Chen, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. [Learning to predict charges for legal judgment via self-attentive capsule network](#). In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, *Frontiers in Artificial Intelligence and Applications*, pages 1802–1809. IOS Press.
- 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 2020, Online, July 5-10, 2020*, pages 7871–7880. ACL.
- Ang Li, Yiquan Wu, Ming Cai, Adam Jatowt, Xiang Zhou, Weiming Lu, Changlong Sun, Fei Wu, and Kun Kuang. 2025. [Legal judgment prediction based on knowledge-enhanced multi-task and multi-label text classification](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6957–6970. ACL.
- Ang Li, Yiquan Wu, Yifei Liu, Kun Kuang, Fei Wu, and Ming Cai. 2024a. [Enhancing court view generation with knowledge injection and guidance](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 5896–5906. ELRA and ICCL.
- Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024b. [Leveraging large language models for NLG evaluation: Advances and challenges](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 16028–16045. ACL.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81. ACL.
- Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. 2022. [Augmenting legal judgment prediction with contrastive case relations](#). In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2658–2667. ICCL.
- Qian Liu, Hang Yu, Qiqi Wang, Qi Xu, Jinpeng Li, Zhuoqun Zou, Rui Mao, and Erik Cambria. 2026. [Legal knowledge infusion for large language models: A survey](#). *Inf. Fusion*, 125:103426.
- Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. [ML-LJP: multi-law aware legal judgment prediction](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1023–1034. ACM.
- Saeed Masoudnia and Reza Ebrahimpour. 2014. [Mixture of experts: a literature survey](#). *Artif. Intell. Rev.*, 42(2):275–293.
- Chunyun Meng, Yuki Todo, Cheng Tang, Li Luan, and Zheng Tang. 2025. [DPFSI: A legal judgment prediction method based on deontic logic prompt and fusion of law article statistical information](#). *Expert Syst. Appl.*, 272:126722.
- Venkatesh Mishra, Bimsara Pathiraja, Mihir Parmar, Sat Chidananda, Jayanth Srinivasa, Gaowen Liu, Ali Payani, and Chitta Baral. 2025. [Investigating the shortcomings of llms in step-by-step legal reasoning](#). In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, *Findings of ACL*, pages 7795–7826. ACL.
- 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.
- Joseph Raz. 1980. *The Concept of a Legal System: An Introduction to the Theory of a Legal System*. Oxford University Press.
- Jingyun Sun, Shaobin Huang, and Chi Wei. 2024. [Chinese legal judgment prediction via knowledgeable prompt learning](#). *Expert Syst. Appl.*, 238(Part E):122177.
- DeepSeek-AI Team. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nat.*, 645(8081):633–638.
- Llama Team. 2024a. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Qwen Team. 2024b. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Qiqi Wang, Ruofan Wang, Kaiqi Zhao, Robert Amor, Benjamin Liu, Xianda Zheng, Zeyu Zhang, and Zijian Huang. 2023. [Towards legal judgment summarization: A structure-enhanced approach](#). In *ECAI 2023 - 26th European Conference on Artificial Intelligence, September 30 - October 4, 2023, Kraków, Poland - Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, *Frontiers in Artificial Intelligence and Applications*, pages 2491–2498. IOS Press.

- 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 2020, Online, November 16-20, 2020*, pages 763–780. ACL.
- 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 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4787–4799. ACL.
- Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. **Precedent-enhanced legal judgment prediction with LLM and domain-model collaboration**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12060–12075. ACL.
- Chunyu Xia, Tieke He, Wenlong Li, Zemin Qin, and Zhipeng Zou. 2019. **Similarity analysis of law documents based on word2vec**. In *19th IEEE International Conference on Software Quality, Reliability and Security Companion, QRS Companion 2019, Sofia, Bulgaria, July 22-26, 2019*, pages 354–357. IEEE.
- 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**. *CoRR*, abs/1807.02478.
- Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. **Distinguish confusing law articles for legal judgment prediction**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3086–3095. ACL.
- Qi Xu, Xiao Wei, Hang Yu, Qian Liu, and Hao Fei. 2024. **Divide and conquer: Legal concept-guided criminal court view generation**. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 3395–3410. ACL.
- Jun Yang, Weizhi Ma, Min Zhang, Xin Zhou, Yiqun Liu, and Shaoping Ma. 2022. **Legalgnn: Legal information enhanced graph neural network for recommendation**. *ACM Trans. Inf. Syst.*, 40(2):33:1–33:29.
- 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, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1854–1864. ACL.
- Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021a. **Neurjudge: A circumstance-aware neural framework for legal judgment prediction**. In *SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 973–982. ACM.
- Linan Yue, Qi Liu, Han Wu, Yanqing An, Li Wang, Senchao Yuan, and Dayong Wu. 2021b. **Circumstances enhanced criminal court view generation**. In *SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1855–1859. ACM.
- 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**. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2221–2230. ACM.
- Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. **Contrastive learning for legal judgment prediction**. *ACM Trans. Inf. Syst.*, 41(4):113:1–113:25.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. **Bertscore: Evaluating text generation with BERT**. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. **Bertscore: Evaluating text generation with BERT**. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Jie Zhao, Ziyu Guan, Cai Xu, Wei Zhao, and Enze Chen. 2022. **Charge prediction by constitutive elements matching of crimes**. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4517–4523. ijcai.org.
- Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. **JECQA: A legal-domain question answering dataset**. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9701–9708. AAAI Press.

A Pipeline Validation Beyond Given Charges

As discussed earlier, we omit charge prediction in our main setup since prior work has already achieved greater than 95% accuracy (Hu et al., 2018; Le et al., 2020; Zhao et al., 2022). For completeness and to assess real-world applicability, we additionally validate a two-stage pipeline where charge prediction precedes CJOG. Our implementation uses the DeepSeek API for charge prediction, followed by different configurations for opinion generation: (1) base LLM only, (2) base LLM with statutory provisions, and (3) base LLM with our LegalChainReasoner.

The charge prediction stage achieved over 99% accuracy across the same crime categories, confirming its reliability and ensuring that the subsequent opinion generation can be meaningfully assessed. Validation was conducted on 10% of samples randomly selected from both training and testing datasets.

The results, summarized in Table 6, show that LegalChainReasoner consistently outperforms both the base model and the statutory provisions baseline. These results demonstrate that our framework maintains its effectiveness in this process, supporting the practical viability of the pipeline approach for real-world deployment.

B Legal Chain Extraction Pipeline

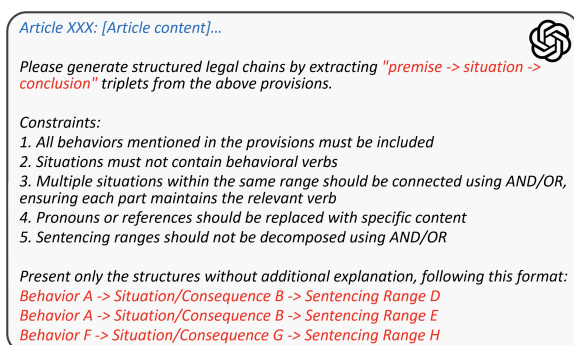


Figure 6: Prompt Template for Legal Chain Extraction. The instruction used to guide the LLM in decomposing statutory provisions into structured triplets.

Extraction Pipeline. For each crime type, we first collect the relevant statutory provisions from the Criminal Law and then prompt an LLM to decompose each provision into one or more *premise-situation-conclusion* triplets under the constraints in Figure 6. During prompt development, these

constraints were iteratively refined to target the most common early-stage extraction errors, including omitted behaviors, incorrect segmentation of logical scope (especially AND/OR), violations of the requirement that the situation node should not contain behavioral verbs, and incorrect parsing of sentencing ranges. Because the number of relevant statutory provisions per crime type is limited, the resulting chain set can be reused across many cases.

Extraction Quality. We asked legal experts to evaluate the generated Legal Chains. Across all 12 crime categories in Chinese criminal law, the generated Legal Chains achieved an accuracy of 98%, confirming the effectiveness of the proposed extraction prompt. Our empirical study further shows that large language models, such as ChatGPT, have consistently high accuracy in both the extraction and structuring of reasoning chains, including the correct treatment of subjective terms such as “severe circumstances,” “particularly serious cases,” and “relatively large amounts.” Moreover, this construction method allows legal knowledge to be represented more systematically and efficiently than in previous studies.

Failure Modes and Propagation. Although the extraction quality is high, failures may still occur, especially for provisions with open-ended enumerations such as “etc.,” “as well as”, or “other”. If such a branch is omitted, the resulting chain set suffers from *coverage loss*, reducing the availability of relevant statutory logic for certain fact patterns and potentially weakening reasoning completeness and sentencing prediction. If such a branch is incorrectly attached, the resulting chain introduces *semantic bias*, injecting inaccurate legal abstractions that may distort both legal grounding and the sentencing signal. In our experiments, all extracted chains are human-verified and corrected before use, which effectively cuts off the downstream impact of extraction noise and yields a high-quality, reusable Legal Chains for each crime type.

While robustness across different legal languages and jurisdictions remains a broader challenge for the LegalAI field, our empirical results demonstrate that the proposed approach is effective for the Chinese criminal law domain. This validation provides confidence that the method can reliably handle the types of ambiguous legal expressions that frequently occur in statutory provisions.

Table 6: Two-stage Pipeline without Given Charges (Charge Prediction→Opinion Generation). Comparison across prediction and generation metrics; GPTScore denotes GPTScore-pairwise. LegalChainReasoner performs best, supporting real-world use.

Model	MAE	RMSE	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-N	BERTScore	GPTScore
Llama-3.2-3B	14.42	23.39	46.83	28.94	41.93	35.76	27.61	25.97	79.18	3.37
+ Statutory provisions	14.52	23.80	51.42	33.51	46.54	37.51	30.16	28.60	80.45	34.45
+ LegalChainReasoner	13.31	21.75	51.93	34.18	46.95	39.64	31.96	30.32	80.59	62.18

Table 7: Impact of Pooling Strategies. Ablation results comparing different pooling methods for chain representations, showing that Mean Pooling yields the most stable performance (GPTScore represents the GPTScore-pairwise).

Model	MAE	RMSE	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-N	BERTScore	GPTScore
Mean Pooling	15.41	25.33	52.82	34.91	48.00	37.50	30.42	28.85	80.42	36.00
Sum Pooling	16.79	27.50	48.94	30.70	43.78	35.56	27.93	26.33	79.26	8.00
Min Pooling	15.63	25.87	52.64	34.67	47.67	37.42	30.29	28.72	80.37	28.00
Max Pooling	15.63	26.05	52.79	34.88	47.91	37.68	30.55	28.97	80.41	28.00

Table 8: Statistics of LAIC-2021 and PCCD Datasets. Distribution of case counts and text lengths across the training and testing sets.

	LAIC-Train	LAIC-Test	PCCD
Total Cases	33,841	1,200	100
Text Length (Characters)			
Fact Description			
Average	548.6	509.6	394.0
Median	321.0	306.5	321.5
Court Opinion			
Average	343.6	366.1	394.0
Median	280.0	299.0	332.0

C Pooling Strategies Analysis

To determine the most effective method for aggregating token embeddings into a unified node representation within the Legal Chain, we conducted an analysis comparing four common pooling strategies: Mean, Sum, Min, and Max pooling. We utilized Llama-3.2-3B with LegalChainReasoner as the base model and evaluated performance on the test dataset.

As shown in Table 7, Mean Pooling achieves the best overall performance across both regression metrics and generation quality metrics, it provides a more robust representation of the overall semantic content of the legal chain components, which is crucial for the subsequent logic transformation layers. Consequently, we adopted Mean Pooling as the standard configuration for all main experiments.

D Datasets

We utilize open-source, real-world datasets: the Legal AI Challenge 2021 (LAIC-2021) and the People’s Court Case Database (PCCD). Both datasets contain cases published by courts in China. The

key difference is that the PCCD dataset was curated by the Supreme People’s Court of China in 2023 and is intended to guide judges in handling similar cases. Therefore, the cases of PCCD are much more complex, and the number of cases is much smaller than LAIC-2021. Statistics for the datasets are shown in Table 8.

E Evaluation Metrics

We select several types of evaluation metrics to assess the quality of the generated judicial opinions.

E.1 Numerical Evaluation

For sentencing prediction, we formulate prison-term estimation as a regression task rather than a classification task, because discretizing sentence lengths may lose fine-grained information and impose arbitrary category boundaries. We therefore evaluate performance using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

E.2 Lexical Evaluation

To assess the quality of legal reasoning generation from a lexical perspective, we utilize standard n-gram overlap-based metrics. Specifically, we use ROUGE-1, ROUGE-2, and ROUGE-L to measure unigram, bigram, and longest common subsequence overlaps, respectively. Additionally, we report BLEU-1, BLEU-2, and BLEU-N scores to evaluate n-gram precision at different levels. These metrics capture surface-level similarity between the generated and reference texts.

E.3 Embedding Evaluation

To assess semantic similarity beyond surface-level lexical overlap, we utilize BERTScore (Zhang

Table 9: Legal Question Assessment (Legal-QA). Pairwise evaluation by LLMs focusing on deep legal validity dimensions: statutory compliance, sentencing appropriateness, consistency, and completeness.

Model	Statutory Compliance	Sentencing Appropriateness	Case-Sentence Consistency	Content Completeness	Average Score
AttS2S	0.00	0.00	0.00	0.00	0.00
BART	0.14	0.08	0.10	0.08	0.10
Lawyer-Llama-13B-V2	0.14	0.18	0.18	0.22	0.18
+ Statutory provisions	0.14	0.16	0.16	0.10	0.14
+ LegalChainReasoner	0.58	0.58	0.56	0.60	0.58

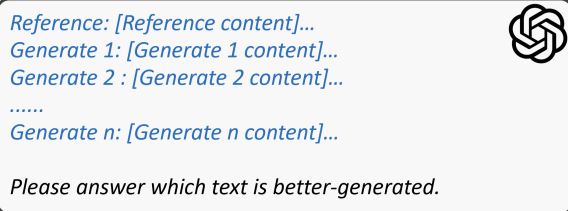


Figure 7: Evaluation Prompt for GPTScore. The instruction used for the pairwise comparison of generated judicial opinions.

et al., 2020b), which leverages contextualized embeddings from pre-trained transformer models to compare generated and reference judicial opinions. This metric captures deeper semantic alignment and is particularly effective in identifying meaning-preserving variations that traditional n-gram-based metrics may overlook.

E.4 LLM-based Evaluation

To evaluate the logical consistency and legal soundness of the generated content, we adopt an LLM-based assessment method known as GPTScore (Fu et al., 2024). Following recent research, we employ the GPTScore-pairwise (Li et al., 2024b) variant, which has been shown to be more effective. Figure 7 illustrates the prompt used for pairwise comparison, based on prior work.

Tables 1 and 2 report GPTScore-pairwise results. For AttS2S and BART, we report only averaged values in the main tables: on LAIC, 0.0/0.0/0.0 and 9.87/21.68/28.97, respectively; on PCCD, 2.0/1.0/1.0 and 11/29/35, respectively.

E.5 Rule-based Evaluation

We design a rule-based method to evaluate the information consistency between the reference and generated judicial opinions. This method focuses on three key aspects: defendant, situation, and sentencing. First, we compare the defendant’s name (e.g., "Li XX") to ensure it is consistently represented in both the reference and generated opinions. Second, we assess whether both opinions describe the same situation, based on structured

legal chains extracted from the statutory provisions. Finally, given the situation mentioned in the generated opinion, we verify whether the predicted sentence aligns with the expected sentencing level for that situation. The final score is computed as the product of the three component accuracies.

F Legal Question Assessment

While standard NLG metrics (e.g., ROUGE, BLEU) and regression metrics (e.g., MAE, RMSE) are commonly used to evaluate generated judicial opinions, they primarily capture surface-level similarity or numerical accuracy. Such metrics are insufficient for assessing the correctness of legal reasoning, particularly in terms of compliance with statutory provisions or judicial precedents. Although our rule-based screening method provides a degree of legal validation, deeper semantic and legal validity checks are still required.

To address this, we propose a Legal Question Assessment (Legal-QA) metric, designed to evaluate the correctness of legal reasoning in generated opinions. Legal-QA measures four key dimensions of reasoning validity: (1) Statutory Compliance: Accuracy of legal provisions referenced and applied; (2) Sentencing Appropriateness: Correctness of sentencing decisions relative to case circumstances; (3) Case-Sentence Consistency: Alignment between case facts and sentencing outcomes; (4) Content Completeness: Comprehensiveness of legal analysis and reasoning.

We conducted this evaluation on a randomly selected subset of 50 cases from the test dataset. For each case, we provided the reference judicial opinion, our model output, and all baseline outputs. LLMs were then used to identify the highest-quality output for each evaluation dimension. As shown in Table 9, LegalChainReasoner achieves the highest selection ratios and average scores on all four dimensions, demonstrating superior performance in generating legally sound opinions.

G Case Study

Table 10: Case Demonstration Example 1. (Background colors denote the evaluation of generation quality: Green denotes correct results; Red indicates error results; and Yellow represents results that are partially correct but suboptimal or inconsistent)

Case Information	
Type	Robbery
Facts	<p>经审理查明，2012年1月2日22时30分许，被告人A伙同B、C（二人已判刑）驾驶一辆黑色轿车至X市Y镇1号加油站，蒙面持刀、工具，并以言语相威胁，抢走该站现金300余元。1月3日凌晨1时许，上述几人又驾车至X市Z路2号加油站，以同样手段抢走现金1000余元。赃款已挥霍。</p> <p>The court ascertained that on January 2, 2012, at approximately 22:30, Defendant A, together with B and C (both already sentenced), drove a black sedan to Gas Station 1 in Town Y, City X. Wearing masks and armed with knives and tools, they threatened the staff and robbed over 300 yuan in cash from the station. At approximately 1:00 on January 3, the aforementioned individuals drove to Gas Station 2 on Road Z in City X and used the same methods to rob over 1,000 yuan in cash. The stolen money has been spent.</p>
Opinion	<p>本院认为，被告人A以非法占有为目的，以暴力威胁手段抢劫他人财物，其行为已构成抢劫罪。公诉机关指控的罪名成立，本院依法予以支持。被告人A持械抢劫，酌情从重处罚。被告人A当庭自愿认罪，酌情从轻处罚。依照《中华人民共和国刑法》第二百六十三条，第二十五条第一款，第五十二条，第五十三条之规定，判决如下：判处有期徒刑42个月。</p> <p>This court holds that Defendant A, with the purpose of illegal possession, used violence and threats to rob others' property. His actions constitute the crime of robbery. The prosecution's charge is established and is supported by this court. Defendant A committed armed robbery, which warrants a heavier punishment. Defendant A voluntarily admitted guilt in court, which warrants a lenient punishment. In accordance with Articles 263, 25(1), 52, and 53 of the Criminal Law of the People's Republic of China, the judgment is as follows: 42 months of fixed-term imprisonment.</p>
Evaluation Metrics	
Defendant	Defendant A
Reasoning	Robbery (with knives and tools); Multiple instances (two gas stations)
Sentencing	42 months of imprisonment
Generated Opinions	
AttS2S	<p>Opinion: This court holds that Defendant D, with the purpose of illegal possession, used methods of fabricating facts and concealing the truth to defraud others' property, with the amount being huge. These actions constitute the crime of fraud. The prosecution's charge is established. The defendant truthfully confessed his crimes, which warrants a lenient punishment. Based on the defendant's criminal facts, circumstances, and remorseful attitude, a lenient punishment is appropriate. In accordance with Articles 266, 67(1), and 62 of the Criminal Law of the People's Republic of China, the judgment is as follows: 36 months of fixed-term imprisonment.</p> <p>Evaluation: × Incorrectly identifies defendant as "Defendant D"; × Completely mischaracterizes the crime as fraud instead of robbery and fails to mention used violence and threats as premise; × Sentencing inconsistent with reasoning</p>
BART	<p>Opinion: This court holds that Defendant A, with the purpose of illegal possession, together with others used violence and threats to forcibly rob others' property. These actions constitute the crime of robbery. The prosecution's charge is established. The Defendant A played a minor role in the joint crime and is an accomplice who should receive reduced punishment according to law. The Defendant A truthfully confessed his crimes after being arrested, which warrants a lenient punishment according to law. In accordance with Articles 263, 25(1), 27, 67(3), and 64 of the Criminal Law of the People's Republic of China, the judgment is as follows: 30 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly identifies crime as robbery and mention used violence and threats as premise; × Sentencing inconsistent with reasoning and actual sentencing range</p>
Lawyer-Llama-13B-V2	<p>Opinion: This court holds that Defendant A, together with others, with the purpose of illegal possession, used violence and threats to rob public and private property. His actions constitute the crime of robbery. The prosecution's charge is supported by clear facts and sufficient evidence. The Defendant A voluntarily admitted guilt in court, which warrants consideration for leniency. In accordance with Articles 263(1) and 55(1) of the Criminal Law of the People's Republic of China, the judgment is as follows: 60 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly identifies crime as robbery and mention used violence and threats as premise; ✓ Sentencing consistent with reasoning though higher than actual sentence</p>
Lawyer-Llama-13B-V2 + Statutory Provisions	<p>Opinion: This court holds that Defendant A, together with others, with the purpose of illegal possession, used violence and threats to rob public and private property. His actions constitute the crime of robbery. The prosecution's charge is supported by clear facts and sufficient evidence. The Defendant A voluntarily admitted guilt in court, which warrants consideration for leniency. In accordance with Article 263(4) of the Criminal Law of the People's Republic of China, the judgment is as follows: 120 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly identifies crime as robbery and mention used violence and threats as premise; × Sentencing inconsistent with reasoning and actual sentencing range</p>
Lawyer-Llama-13B-V2 + Legal-Chain-Reasoner	<p>Opinion: This court holds that Defendant A, together with others, with the purpose of illegal possession, used violence and threats to rob public and private property. His actions constitute the crime of robbery. The prosecution's charge is supported by clear facts and sufficient evidence. The Defendant A voluntarily admitted guilt in court, which warrants consideration for leniency. In accordance with Article 263 of the Criminal Law of the People's Republic of China, the judgment is as follows: 48 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly identifies crime as robbery and mention used violence and threats as premise; ✓ Sentencing consistent with reasoning though higher than actual sentence but most closed</p>

Table 11: Case Demonstration Example 2. (Background colors denote the evaluation of generation quality: Green denotes correct results; Red indicates error results; and Yellow represents results that are partially correct but suboptimal or inconsistent)

Case Information	
Type	Fraud
Facts	<p>经审理查明，被告人A于2013年1月至4月，以能为被害人B办理工厂上班为名，两次骗取被害人B人民币16万元。被告人A于2014年11月份，以给C的儿子办理工作为由，骗取C人民币24万元。被告人A于2017年2月份，以帮助D办事为由先后三次骗取D人民币22万元。综上，被告人A共诈骗三起，诈骗金额计人民币62万元。</p> <p>The court ascertained that from January to April 2013, Defendant A defrauded Victim B by claiming he could arrange employment at Factory X for Victim B’s son, receiving 160,000 CNY on two occasions. In November 2014, the defendant defrauded Victim C of 240,000 CNY by claiming he could arrange employment for Victim C’s son. In February 2017, the defendant defrauded Victim D of 220,000 CNY on three occasions by claiming he could help with certain matters. In total, Defendant A committed three instances of fraud amounting to 620,000 CNY.</p>
Opinion	<p>本院认为，公诉机关指控被告人A以非法占有为目的，虚构事实，隐瞒真相，多次骗取他人财物，数额特别巨大的事实，有经过庭审查实的证据予以证实，A的行为构成诈骗罪，公诉机关指控的犯罪事实和罪名成立。关于A辩称称，其没有承诺为C儿子办工作，收到C24万元系借款，其行为不构成诈骗罪的意见，经审理认为，被害人C、证人E、证人F证言均证实A谎称能为C儿子办工作，C共给付A24万元，书证收据（欠条），载明为出款人孩子办工作；短信截图证实A谎称为C儿子办理工作，让C拿钱的事实，上述证据能够形成完整证据链条，足以证实A谎称能为C儿子办工作，诈骗C24万元的事实，故A的辩解不成立；关于A辩称称，其在D处分三次拿22万元后让G拿走了，钱没花着，不能认定其构成诈骗罪的辩护意见，经审理认为，被害人D的陈述、证人G的证言均证实将22万元交付给A，且G对A称其将22万元借走一事予以否认亦无证据证实，故A的辩解不成立。被告人A诈骗被害人B案系刑满释放后五年内重新犯罪，是累犯，应当从重处罚；被告人A如实供述诈骗被害人B的罪行，可以从轻处罚。依照《中华人民共和国刑法》第二百六十六条、第六十五条第一款、第六十七条第三款、第五十二条、第五十三条、第六十四条之规定，判决如下：判处有期徒刑138个月。</p> <p>This court holds that the prosecution’s charges against Defendant A—that with the purpose of illegal possession, he fabricated facts, concealed the truth, and repeatedly defrauded others of property in an especially large amount—are supported by evidence verified during the trial. Defendant A’s actions constitute the crime of fraud, and the prosecution’s charges regarding the criminal facts and offense are established. Regarding Defendant A’s defense that he did not promise to arrange employment for Victim C’s son and that the 240,000 CNY received from Victim C was a loan rather than fraud, the court finds that testimonies from Victim C, Witness E, and Witness F all confirm that Defendant A falsely claimed he could arrange employment for Victim C’s son, and Victim C paid Defendant A 240,000 CNY in total. Documentary evidence including receipts (IOUs) indicate the money was for arranging employment for the payer’s child. Text message screenshots confirm Defendant A’s false claims about arranging employment and requesting money from Victim C. This evidence forms a complete chain sufficient to prove that Defendant A defrauded Victim C of 240,000 CNY by falsely claiming he could arrange employment for Victim C’s son. Therefore, Defendant A’s defense is not established. Regarding Defendant A’s defense that after receiving 220,000 CNY from Victim D in three installments, he let Witness G take it away and did not spend the money himself, thereby not constituting fraud, the court finds that testimonies from Victim D and Witness G confirm that 220,000 CNY was given to Defendant A, and Witness G denies Defendant A’s claim that he borrowed the 220,000 CNY. There is no evidence to support Defendant A’s claim, so his defense is not established. The defendant committed the fraud against Victim B after being released from prison for less than five years, making him a recidivist, which warrants a heavier punishment. The defendant truthfully confessed to defrauding Victim B, which warrants a lenient punishment. In accordance with Articles 266, 65(1), 67(3), 52, 53, and 64 of the Criminal Law of the People’s Republic of China, the judgment is as follows: 138 months of fixed-term imprisonment.</p>
Evaluation Metrics	
Defendant	Defendant A
Reasoning	Multiple defrauded others of property (three separate victims); Especially large amount (620,000 CNY)
Sentencing	138 months of imprisonment
Generated Opinions	

AttS2S	<p>Opinion: The Court holds that Defendant Z, with the purpose of illegal possession, used methods of fabricating facts and concealing the truth to defraud others' property, with the amount being large. These actions constitute the crime of fraud. The prosecution's charge is established. The defendant truthfully confessed his crimes after committing them, which warrants lenient punishment according to law. Based on the defendant's criminal facts, circumstances, and remorseful attitude, a lenient punishment is appropriate. In accordance with Articles 266, 67(1), and 62 of the Criminal Law of the People's Republic of China, the judgment is as follows: 36 months of fixed-term imprisonment.</p> <p>Evaluation: × Incorrectly identifies defendant; × Correctly classifies premise as defraud others' property but mischaracterizes amount as "large" rather than "especially large"; × Sentencing inconsistent with reasoning and actual sentencing range</p>
BART	<p>Opinion: The Court holds that Defendant A, with the purpose of illegal possession, used methods of fabricating facts and concealing the truth to defraud others' property, with the amount being especially large. These actions constitute the crime of fraud. The prosecution's charge against the defendant for the crime of fraud is supported by clear facts and sufficient evidence. The charge is established and supported by this court. The defendant truthfully confessed his crimes after being arrested, which warrants lenient punishment according to law. In accordance with Articles 266, 67(3), 52, 53, and 64 of the Criminal Law of the People's Republic of China, the judgment is as follows: 120 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly classifies premise as defraud others' property and correctly characterizes amount as "especially large"; ✓ Sentencing consistent with reasoning though lower than actual sentence</p>
Lawyer-Llama-13B-V2	<p>Opinion: The Court holds that Defendant A, with the purpose of illegal possession, fabricated facts and concealed the truth, to repeatedly defraud others' property, with the amount being especially large. These actions constitute the crime of fraud. The prosecution's charge is supported by evidence and is established. The defendant truthfully confessed his crimes after being arrested, which constitutes a confession and warrants lenient punishment according to law. The defendant has actively returned the defrauded property and obtained forgiveness from the victims, which may warrant lenient punishment. In accordance with Articles 266 and 67(3) of the Criminal Law of the People's Republic of China, the judgment is as follows: 96 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly classifies premise as defraud others' property and correctly characterizes amount as "especially large"; × Sentencing inconsistent with reasoning and actual sentencing range</p>
Lawyer-Llama-13B-V2 + Statutory Provisions	<p>Opinion: The Court holds that Defendant A, with the purpose of illegal possession, fabricated facts and concealed the truth to defraud others' property, with the amount being especially large. These actions constitute the crime of fraud. The prosecution's charge is supported by clear facts and sufficient evidence, and is established. The defendant voluntarily admitted guilt in court, which warrants lenient punishment. After the case occurred, the defendant's family actively compensated the victims for economic losses and obtained forgiveness, which warrants lenient punishment. The defense counsel's relevant defense opinions are consistent with the established facts and are adopted. In accordance with Articles 266, 52, and 64 of the Criminal Law of the People's Republic of China, the judgment is as follows: 120 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly classifies premise as defraud others' property and correctly characterizes amount as "especially large"; ✓ Sentencing consistent with reasoning though lower than actual sentence</p>
Lawyer-Llama-13B-V2 + Legal-Chain-Reasoner	<p>Opinion: The Court holds that Defendant A, with the purpose of illegal possession, fabricated facts and concealed the truth to defraud others' property, with the amount being especially large. These actions constitute the crime of fraud. The prosecution's charge is supported by clear facts, sufficient evidence, and is established. Given that the defendant voluntarily admitted guilt in court, a relatively lenient punishment may be considered. In accordance with Articles 266, 52, and 53 of the Criminal Law of the People's Republic of China, the judgment is as follows: 120 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly classifies premise as defraud others' property and correctly characterizes amount as "especially large"; ✓ Sentencing consistent with reasoning though lower than actual sentence</p>

Table 12: Case Demonstration Example 3. (Background colors denote the evaluation of generation quality: Green denotes correct results; Red indicates error results; and Yellow represents results that are partially correct but suboptimal or inconsistent)

Case Information	
Type	Misappropriation of Funds
Facts	<p>经审理查明：2017年3月5日，被告人A与B某签定股权债权转让协议书，在2017年3月10日至3月13日，B某分三次通过银行转账转入被告人A账户150万元的合同应付款，被告人A将150万元挪用还银行贷款。2017年4月21日至5月9日，B某将X公司Y分公司的售房款分6次转入被告人A账户106万元，被告人A将其中85万元售房款用于还自己的贷款和个人使用。案发前已归还。另查明，被害人X公司及X公司Y分公司均表示对被告人A谅解。</p> <p>Facts as determined by the court: On March 5, 2017, Defendant A signed an equity and debt transfer agreement with Person B. Between March 10-13, 2017, Person B transferred contractual payments totaling 1.5 million yuan to Defendant A's account in three separate bank transfers, which Defendant A misappropriated to repay bank loans. From April 21 to May 9, 2017, Person B transferred housing sales proceeds totaling 1.06 million yuan from Branch Y of Company X to Defendant A's account in six installments. Defendant A used 850,000 yuan of these funds to repay personal loans and for personal expenses. The funds were returned before the case was filed. Further investigation revealed that the victims, Company X and its Branch Y, have both expressed understanding toward Defendant A</p>
Opinion	<p>本院认为，被告人A利用职务之便挪用资金，数额较大，犯罪事实清楚，证据确实、充分，其行为构成挪用资金罪。公诉机关指控的罪名及事实成立，本院予以确认。被告人A到案后如实供述自己的罪行，可从轻处罚。被告人A获得被害人谅解，可酌情从轻处罚。公诉机关的量刑建议符合法律规定，本院予以确认。辩护人关于被告人如实供述、获得被害人谅解、认罪认罚可从轻处罚的辩护意见予以采纳。辩护人关于起诉书指控第一起被告人挪用150万元不应认定及第二起犯罪数额应为45万元的辩护意见，经查与事实不符，不予采纳。根据被告人犯罪的事实、犯罪的性质、情节和对于社会的危害程度，依照《中华人民共和国刑法》第二百七十二條、第六十四條、第六十七條第三款之規定，判決如下：判处有期徒刑7个月。</p> <p>This court holds that Defendant A took advantage of his position to misappropriate funds in a relatively large amount. The facts of the crime are clear and the evidence is reliable and sufficient. His behavior constitutes the crime of misappropriation of funds. The prosecution's charge and alleged facts are established and confirmed by this court. The defendant truthfully confessed his crimes after being arrested, which warrants a lenient punishment. The defendant obtained the forgiveness of the victims, which also warrants consideration for leniency. The sentencing recommendation of the prosecution complies with the law and is confirmed by the court. The defense's opinion that the defendant truthfully confessed, obtained the victims' understanding, and admitted guilt and accepted punishment warranting leniency is adopted. The defense's opinion regarding the prosecution's allegations about the first misappropriation of 1.5 million yuan and that the amount in the second crime should be 450,000 yuan is not consistent with the facts and is not adopted. Based on the facts, nature, circumstances, and social harm of the defendant's crime, in accordance with Articles 272, 64, and 67(3) of the Criminal Law of the People's Republic of China, the judgment is as follows: 7 months of fixed-term imprisonment.</p>
Evaluation Metrics	
Defendant	Defendant A
Reasoning	Misappropriated funds; Relatively large amounts
Sentencing	7 months of imprisonment
Generated Opinions	
AttS2S	<p>Opinion: This court holds that Defendant Z, with the purpose of illegal possession, used methods of fabricating facts and concealing the truth to defraud others' property, with the amount being large. These actions constitute the crime of fraud. The prosecution's charge is established. The defendant's criminal facts and circumstances warrant lenient punishment according to law. Based on the defendant's criminal facts, circumstances, and the degree of social harm, in accordance with Articles 266, 67(1), and 60 of the Criminal Law of the People's Republic of China, the judgment is as follows: 36 months of fixed-term imprisonment.</p> <p>Evaluation: × Incorrectly identifies defendant; × Mischaracterizes crime as fraud rather than misappropriation of funds and incorrectly characterizes amount as "large" rather than "relatively large"; × Sentencing inconsistent with reasoning and actual sentencing range</p>

BART	<p>Opinion: This court holds that Defendant A, as a company employee, took advantage of his position to misappropriate company funds for personal use, with the amount being large, and not returned for over three months. His actions constitute the crime of misappropriation of funds. The prosecution's charge is established and supported by this court. The defendant truthfully confessed his crimes after being arrested, which warrants lenient punishment according to law. The defendant has obtained the understanding of the victims, which warrants consideration for leniency. In accordance with Articles 272(1), 67(3), and 64 of the Criminal Law of the People's Republic of China, the judgment is as follows: 36 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; × Correctly identifies premise as misappropriate company funds but incorrectly characterizes amount as "large" rather than "relatively large"; × Sentencing inconsistent with reasoning and actual sentencing range</p>
Lawyer-Llama-13B-V2	<p>Opinion: This court holds that Defendant A took advantage of his position to misappropriate public funds for profit-making activities, with the amount being large. His actions constitute the crime of misappropriation of funds. The prosecution's charge is established with clear facts and sufficient evidence. The defendant voluntarily surrendered himself after the crime and truthfully confessed, constituting voluntary surrender, which warrants lenient punishment. The defendant voluntarily admitted guilt and obtained the victims' understanding, which also warrants consideration for leniency. In accordance with Articles 272(1), 67(1), 72(1), 72(3), 73(2), and 73(3) of the Criminal Law of the People's Republic of China, the judgment is as follows: 36 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; × Incorrectly characterizes as "misappropriation of public funds" rather than "misappropriation of company funds" and incorrectly describes amount as "large" rather than "relatively large"; × Sentencing inconsistent with reasoning and actual sentencing range</p>
Lawyer-Llama-13B-V2 + Statutory Provisions	<p>Opinion: This court holds that Defendant A took advantage of his position to misappropriate company funds for personal use, with the amount being relatively large and not returned for over three months. His actions constitute the crime of misappropriation of funds. The prosecution's charge is established with clear facts and sufficient evidence. Before the case was filed, the defendant returned most of the misappropriated funds to the victim company and obtained their understanding, which warrants consideration for leniency. The defendant voluntarily admitted guilt during the trial, which also warrants leniency. The defense counsel's opinion that the defendant is a first-time offender who has actively compensated for the victims' losses and obtained their understanding, requesting a suspended sentence, is verified and adopted. In accordance with Articles 272(1), 67(3), 72(1), 73(2), and 73(3) of the Criminal Law of the People's Republic of China, the judgment is as follows: 12 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly identifies premise as misappropriation of funds and correctly characterizes amount as "relatively large"; ✓ Sentencing consistent with reasoning though higher than actual sentence</p>
Lawyer-Llama-13B-V2 + Legal-Chain-Reasoner	<p>Opinion: This court holds that Defendant A took advantage of his position to misappropriate company funds for personal use, with the amount being relatively large and not returned for over three months. His actions constitute the crime of misappropriation of funds. The prosecution's charge and alleged facts are established and legally supported. Given that the defendant actively returned all misappropriated funds before the case was filed, voluntarily admitted guilt in court, and obtained the understanding of the victims, a lenient punishment is warranted. In accordance with Articles 272(1) and 64 of the Criminal Law of the People's Republic of China, the judgment is as follows: 12 months of fixed-term imprisonment.</p> <p>Evaluation: ✓ Correctly identifies defendant; ✓ Correctly identifies premise as misappropriation of funds and correctly characterizes amount as "relatively large"; ✓ Sentencing consistent with reasoning though higher than actual sentence</p>