DefGen-Bench: A Benchmark for Chinese Criminal Defence Opinion Generation in LegalAI

Senbo Zhang[1]*, Qiqi Wang[1]*†, Fanghao Lou[1], Guanyu Chen[1], Yihong Pan[2], Huijia Li[1]†, Qian Liu[2]†
[1]School of Statistics and Data Science, AAIS, LPMC and KLMDASR, Nankai University, Tianjin, China
[2]School of Compute Science, The University of Auckland, New Zealand
{qiqi.wang, hjli}@nankai.edu.cn ypan317@aucklanduni.ac.nz
{zhangbd,guanyu.chen,fanghao.lou}@mail.nankai.edu.cn liu.qian@auckland.ac.nz

Abstract

A defence opinion is an essential step in criminal proceedings, yet it has not been systematically formulated or evaluated as a specific LegalAI task. Grounded in legal principles and practice, we formulate this task as generating a structured defence opinion conditioned jointly on an indictment and the defendant's stated opinion, which often present conflicting claims. We formalize this setting as a dual-perspective generation problem and introduce DefGen-Bench, a benchmark comprising several Chinese criminal cases with expert-reviewed reference defence opinions. We evaluate eight large language models (LLMs) on this task and observe that existing models tend to mirror the defendant's opinion, thereby overlooking more appropriate defence strategies. To address this challenge, we propose Knowledge-Enhanced Highlighted Indictment (KHI), a legal knowledge–guided input enhancement method applicable to both open- and closed-source LLMs. Experiments demonstrate consistent improvements across all evaluated LLMs, validating the effectiveness of the proposed approach[1].

1 Introduction

LegalAI aims to improve the efficiency of legal practice and enhance the accessibility of legal services to the general public (Zhong et al., 2020a). Existing trial-based tasks have judgment summarization (T.y.s.s et al., 2024b; Liu et al., 2024a), legal judgment prediction (T.y.s.s et al., 2024a; Meng et al., 2025), and similar case retrieval (T.y.s.s and Upadhya, 2025; Liu et al., 2025b).

However, these tasks are typically descriptive or predictive (MD, 2019), requiring the users to further translate the information into legal reasoning.

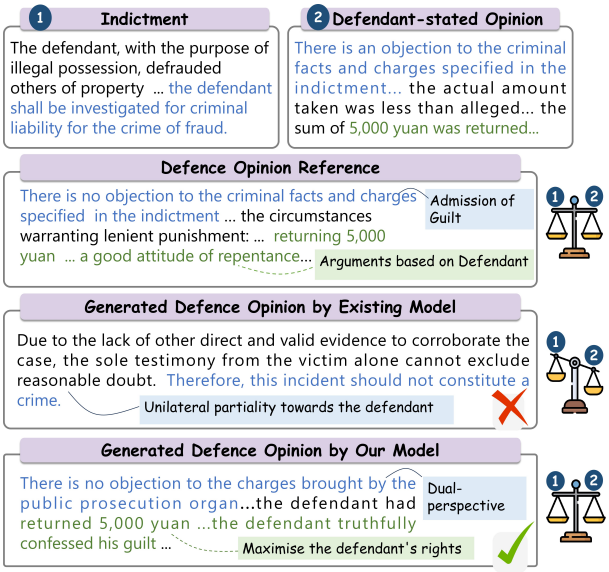


Figure 1: Example of defence opinion generation.

In criminal proceedings, defendants must determine how to respond to the indictment, which facts to highlight, and what legal arguments to construct. This intermediate reasoning process remains unsupported by existing LegalAI formulations (Wang et al., 2024). In this work, we propose a new LegalAI task, termed defence opinion generation, which aims to generate a structured defence opinion. Following real-world legal practice, for example, Section 6A of the *Criminal Procedure and Investigations Act 1996 in England and Wales*[2], we formulate the generation is conditioned on the indictment and defendant-stated position. Under this task definition, we introduce a new benchmark, DefGen-Bench, which consists of Chinese criminal cases collected from public sources.

We evaluate DefGen-Bench on eight LLMs of varying sizes, architectures, providers, and usage settings, including both open-source and closed-source models. Our results show that the generated

*Equal contribution.
†Corresponding author.
[1]The code and dataset are available on: https://github.com/Statistical-NLP-Lab/DefGen-Bench

[2]https://www.legislation.gov.uk/ukpga/1996/25/section/6A

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 35378–35392
July 2-7, 2026

defence opinions often either closely mirror the defendant-stated position or provide only generic defence arguments that overlook case-specific details. For example, Figure 1 presents a cases where the indictment is clear and the evidence is compelling, an effective defence opinion should shift towards mitigation rather than denial. Nevertheless, existing models often uncritically follow the defendant's stated position, resulting in legally implausible defence strategies that may undermine the defendant's rights. These observations indicate that existing models tend to generate outputs from a single perspective, failing to effectively integrate information from both adversarial sides. This dual-perspective requirement challenges LLMs in aligning and reasoning over conflicting inputs (Feng et al., 2024).

To address this challenge, we further propose a knowledge-enhanced highlighted indictment method (KHI), which explicitly emphasizes key indictment elements that require attention and response in defence opinion generation. The proposed method consists of two stages: legal knowledge application and knowledge-aware enhancement. To ensure applicability across both open-source and closed-source LLMs, we design two alternative enhancement strategies that operate either at the embedding level (KHI-E) or at the textual input level (KHI-T). Experimental results across all eight LLMs demonstrate that the proposed methods consistently improve defence opinion generation quality, validating their effectiveness. Moreover, we introduce two new evaluation metrics, Coverage Balance and Information–Function Consistency, to assess the balance and functional adequacy of the generated defence opinions.

Our contributions include,

- We present DefGen-Bench, a benchmark for generating defence opinions in Chinese criminal cases, framing the task as a dual-perspective generation problem based on the indictment and defendant's opinion.
- We propose Knowledge-Enhanced Highlighted Indictment (KHI), a general enhancement framework that improves generation under conflicting dual-perspective inputs by incorporating legal knowledge and highlighting key points in indictments requiring response.
- We evaluate DefGen-Bench and KHI on eight diverse LLMs. Experimental results show consistent improvements in defence statement generation performance. We also propose two new metrics to evaluate whether the generated defence is balanced for both sides.

Table 1: Prompts for Using LLMs in Defence Opinion Generation.

General Prompt	<Role Setting> You are a professional defence attorney tasked with drafting a legal defence statement for your client. <Case Information> Indictment: {*Indictment context*} Defendant-stated opinion: {*Defendant opinion context*} <*Special Information*> *if have* <Requirements> 1. Strictly adhere to the above facts; do not fabricate or exaggerate. 2. Maintain logical coherence; avoid repetition and contradictions. 3. Use professional legal language.
Special Information	
RAG	The relevant statutory provisions is {statutory provisions context}.
CoT	Please generate the result according to the following steps: 1. Case analysis and identification of disputed issues. 2. Legal element review and factual matching. 3. Defence strategy formulation and argumentation framework construction.

2 DefGen-Bench Construction

2.1 Defence Opinion Generation Definition

Following the guidance of Article 37 of the *Criminal Procedure Law of China*[3], Article 5 of the *Norms for Lawyers Handling Criminal Cases*[4], and Article 35 of the *Guiding Opinions on the Work of Sentencing Recommendations for Cases Where Defendants Plead Guilty and Accept Punishment*[5], defence opinions issued by legal professionals are independent of both the parties and the courts. A defence opinion should, in accordance with the facts of the case and in the defendant's best interests, advance one of the following positions: innocence, mitigation or reduction of punishment, or exemption from criminal responsibility. For this reason, the reference defence opinion in Figure 1 adopts a mitigation strategy rather than following the defendant's not-guilty stance. Moreover, in light of general legal principles (Ellis, 2011; Alpa, 1994)

[3] http://www.npc.gov.cn/npc/c2/c12435/201905/t20190521_276591.html

[4] https://www.szlawyers.com/file/upload/20170928/file/20170928164805_3ee1276bde7b4f7dafbd3712964e1117.pdf

[5] https://www.spp.gov.cn/spp/xwfbh/wsfbt/202112/t20211220_539038.shtml#2

and jurisdiction-specific legal knowledge (Richland, 2013; Williams, 2022), a defence opinion in criminal proceedings should reflect the defendant-stated opinion while also addressing both the factual and legal aspects relevant to the application of law to the indictment.

Accordingly, given a set of cases $x_i = \{x_{i1}, x_{i2}, \ldots, x_{il}\}$, consisting of the indictment $s_i = \{s_{i1}, s_{i2}, \ldots, s_{in}\}$ and the defendant's initial opinion $b_i = \{b_{i1}, b_{i2}, \ldots, b_{ih}\}$, the objective of the defence opinion generation task is to employ a model $F(\cdot)$ to generate a defence opinion $y_i = F(x_i)$. A high-quality defence opinion y_i should align with the information in the defendant-stated opinion b_i and provide structured, targeted responses to the indictment s_i. Table 1 shows the sample prompt for using LLMs in this task.

2.2 Data Collection

In China, there are two well-known platforms that publish judicial cases, including China Judgments Online[6] and the China Judgments Case Database[7]. However, to further protect the privacy of the parties involved and case quality, we utilize anonymized cases released through publicly approved legal competitions, specifically the Chinese Legal Intelligence Technology Evaluation (CAIL) (Ji et al., 2018; Ma et al., 2021; Yu et al., 2022b; Zhong et al., 2020b). Using this source, we initially collect approximately 7,000 Chinese criminal cases.

2.3 Data Processing

As publicly available judicial case documents in China primarily are final judgments, we derive the indictment, the defendant-stated opinion, and the defence opinion based on the judgment text and its structural layout (Moys, 2014; Ren et al., 2022). The indictment is a compulsory section in Chinese criminal judgments and can therefore be reliably extracted from the corresponding paragraph.

Since the defendant-stated opinion is not a mandatory component of judgment texts, only 2,124 collected criminal cases contain such opinions. Considering the necessity of defendant-stated opinions for our task, we retain only cases where this section is explicitly present and discard the rest. Among the 2,124 retained cases containing defendant-stated opinions, we further conduct quality filtering and exclude 396 low-quality texts. The extracted defendant-stated opinion is denoted as b_i in DefGen-Bench. Furthermore, since defence opinions are not always authored by legal professionals, we retain only high-quality defence opinions that have been reviewed by legal experts to serve as reliable reference defence opinions y_i.

After applying the above filtering steps, 1,782 cases remain, each containing the three essential components: the indictment, the defendant-stated opinion, and the defence opinion.

Table 2: Statistics of DefGen-Bench. Defence length (P90) denotes the 90th percentile of defence length.

	DefGen
Number of cases	1782
Number of crimes	8
Average length	
- Indictment length	279.90
- Defendant-stated Opinion length	73.21
- Defence length	181.66
Defence length (P90)	347.00

2.4 Data Statistics

Table 2 presents the statistics of the proposed DefGen-Bench following the above processing steps. Due to the quality reason, the collected cases' type are only cover eight frequent crimes. To better control generation length, we follow prior work (Galton, 1885) and apply the 90th percentile as the length threshold. And the assessment by legal experts verifies the high quality of the dataset's realism and consistency with legal practice.

3 Knowledge-enhanced Highlighted Indictment Method

3.1 Overall Framework

To mitigate imbalanced defence opinion generation, as illustrated in Figure 1, we propose a method that leverages legal knowledge to highlight indictment components requiring focused attention and response. We term this approach Knowledge-Enhanced Highlighted Indictment (KHI). Figure 2 shows the overall structure. KHI consists of two phases: legal knowledge application and knowledge-aware enhancement. In the first phase, relevant legal knowledge is identified to determine which aspects of the indictment should be addressed. In the second phase, this knowledge is used to extract and highlight legally salient portions of the indictment, while reducing attention to components that fall outside the scope of the

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

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

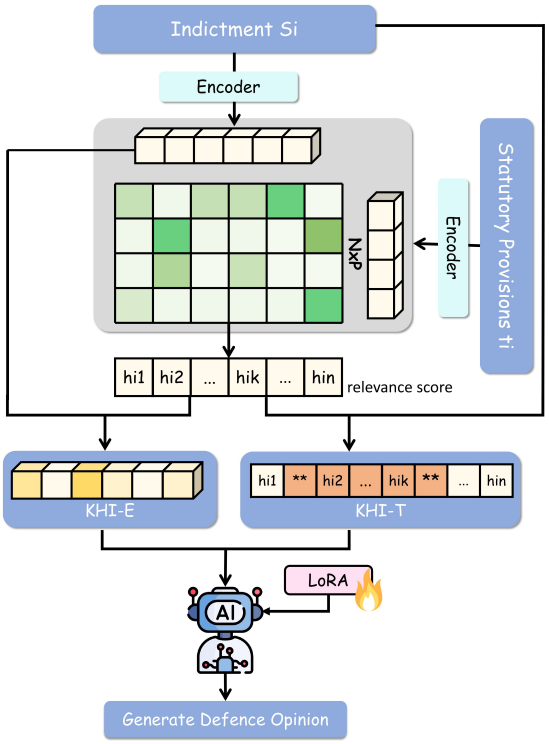


Figure 2: The overall structure of KHI.

required response. To ensure applicability to both open-source and closed-source LLMs, we further design two strategies for the highlighting phase.

3.2 Legal Knowledge Application

Inspired by previous work (Kjær, 2000; Peykani et al., 2025), we adopt statutory provisions as the source of legal knowledge, as they are both readily available and highly effective. For each case involving crime t_i, we encode the corresponding statutory provisions $\{t_{i1}, t_{i2}, \ldots, t_{im}\}$ using the embedding mechanism of the same LLM to ensure consistency in the representational space, yielding $\mathbf{E}_{t_i} \in \mathbb{R}^{d \times m}$, where d denotes the embedding dimensionality determined by the LLM.

To identify the crime type of each case, following established practices in prior research (Baayen et al., 2019; Milin et al., 2017), we employ an efficient crime identification strategy based on discriminative lexical signals. Specifically, we leverage a large-scale LLM to derive salient crime-indicative terms for each crime category and then assess their semantic coverage within the case description. The crime type is determined as the category exhibiting the strongest alignment with the case content, which enables the selection of the corresponding statutory provision set for subsequent processing. This crime identification approach achieves an accuracy of 97%, demonstrating both its effectiveness and efficiency. For more details about our prompt and crime term list, please refer to Appendix A.

3.3 Knowledge-aware Enhancement

The knowledge-aware enhancement stage aims to identify the legally important portions of the indictment $\mathcal{I}s$ that require focused responses, by using the relevant legal knowledge. Given an indictment embedding $\mathbf{E}_{c_j} \in \mathbb{R}^{n \times d}$, we measure the semantic relevance between each indictment and the corresponding statutory provisions $\mathbf{E}_{t_i}$. This relevance reflects the degree to which an indictment span aligns with legally grounded concepts that are essential for constructing an effective defence.

We define a relevance score $r_j^{(c)} \in (0, 1]$ for each indictment span as

$$r_j^{(c)} = \phi\left(\mathbf{W}_q \mathbf{E}_{c_j}, ; \mathbf{W}_k \mathbf{E}_{t_i}\right), \quad j \in \mathcal{I}_s, \qquad (1)$$

where $\phi(\cdot)$ denotes a semantic compatibility function operating in a shared representation space.

Due to the differing levels of access and usage constraints between open-source and closed-source (or extremely large) LLMs, we design two variants of the enhancement mechanism: an embedding-level approach for open-source models and a textual-level approach for closed-source models.

KHI-E version. For open-source LLMs, we rescale token representations within the indictment according to their relevance scores, enabling the model to allocate greater representational capacity to legally salient components during generation. The enhanced case embedding $\hat{\mathbf{E}}c$ is defined as

$$\hat{\mathbf{E}}_c = \begin{cases} r_j^{(c)} \mathbf{E}_{c_j}, & j \in \mathcal{I}_s \\ \mathbf{E}_{c_j}, & \text{otherwise} \end{cases}, \qquad (2)$$

This modulation preserves the original embedding structure while softly guiding the model's attention toward indictment elements that require legal rebuttal. Since the relevance scores are bounded, the transformation remains stable and does not distort the embedding space. For efficiency, the relevance weights can be precomputed and cached during both training and inference.

KHI-T version. For closed-source or extremely large LLMs, direct access to embedding layers is unavailable. In this setting, we design a textual

highlighting strategy that operationalizes relevance through input restructuring. Specifically, we apply a threshold k to the relevance scores and insert importance markers around indictment phrases whose relevance satisfies $r_j^{(c)} \geq k$. This approach encourages the model to prioritize legally relevant indictment components without requiring access to internal representations. The relevance scores for closed-source models are computed using smaller, accessible LLMs from the same model family to ensure the semantic space.

3.4 Design Analysis

We provide a theoretical justification for KHI by analyzing its effects on attention allocation, gradient propagation, and information selectivity in Transformer-based models. We show that knowledge-guided enhancement encourages the model to focus on legally salient indictment components while suppressing irrelevant noise during both inference and training. The detailed proof is shown in Appendix B.

Semantic Focus. Let $\mathbf{E}_c$ and $\hat{\mathbf{E}}_c$ denote the original and knowledge-enhanced input embeddings, respectively. For any attention head in a Transformer, the attention score for token j is given by

$$s_j = \frac{q^\top k_j}{\sqrt{d}} = \begin{cases} r_j^{(c)} \cdot \frac{q^\top \mathbf{E}_{c_j}}{\sqrt{d}}, & j \in \mathcal{I}_s \\ \frac{q^\top \mathbf{E}_{c_j}}{\sqrt{d}}, & \text{otherwise} \end{cases}, \quad (3)$$

since attention weights are computed via softmax, tokens with high legal relevance ($r_j^{(c)} \approx 1$) preserve their original attention scores, while low-relevance tokens are exponentially down-weighted. As a result, the model allocates relatively more attention to indictment components that are legally salient, effectively focusing generation on statute-relevant content.

Gradient Focus. Scaling input embeddings by relevance scores is equivalent to applying a knowledge-guided gradient mask during backpropagation. Specifically,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{E}_{c_j}} = \begin{cases} r_j^{(c)} \cdot \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{E}}_{c_j}}, & j \in \mathcal{I}_s \\ \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{E}}_{c_j}}, & \text{otherwise} \end{cases}, \quad (4)$$

Consequently, gradients associated with low-relevance tokens are suppressed, while gradients for legally critical tokens propagate normally. This encourages the model to learn primarily from statute-relevant information and reduces the influence of irrelevant noise.

Selective Enhancement. KHI enhances only the indictment while leaving the defendant's opinion unchanged. This design aligns with the task objective: defence opinions must legally respond to the indictment while remaining grounded in the defendant's opinion. Enhancing only the indictment injects legal priors where they are most needed, while avoiding the amplification of non-legal or subjective expressions in the defendant's opinion. From an information-theoretic perspective, this selective enhancement increases the effective signal-to-noise ratio at the input level, facilitating more balanced and legally grounded generation.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on DefGen-Bench. To further evaluate the balance of generated defence opinions, we ask legal experts to annotate the functional focus of each sentence in the defence opinion, including whether it primarily responds to the indictment, follows the defendant's stated opinion, or addresses both. Based on these annotations, we construct a labeled subset, denoted as DefGen-Bench-Label, which is used for in-depth evaluating the generated defence opinion quality.

Baselines. We compare our approach against three categories of text generation methods. **Traditional methods.** We select two representative encoder–decoder models commonly used for long-text generation: BART (Lewis et al., 2020a) and Longformer (Beltagy et al., 2020). **Open-source LLMs.** We evaluate six representative generative LLMs: Qwen2.5-3B (Qwen et al., 2025), Llama3.2-3B (Martra, 2025), Qwen2-7B (Team et al., 2024), Llama-3-8B-Instruct (Dubey et al., 2024), Lawyer-Llama-13B-v2 (Huang et al., 2023), and Qwen3-14B (Yang et al., 2025). **Closed-source and larger LLMs.** We further include two strong closed-source models for comparison: Qwen-PLUS (Yang et al., 2025) and DeepSeek-V3.2 (Liu et al., 2025a).

For each LLM, we select vanilla setting, supervised fine-tuning (SFT) (Dong et al., 2024) with LoRA (Hu et al., 2021), retrieval-augmented generation (RAG) (Lewis et al., 2020b), and chain-of-thought (CoT) prompting (Wei et al., 2022). The

Table 3: Comparsion on open-source LLMs. CB. and IFC. means CoverageBalance and Info-Function Consistency. Pink intensity indicates performance level (darker = better).

Model		DefGen-Bench					DefGen-Bench-Label				
		Rouge-1	Rouge-2	Rouge-L	BERTScore	GPTScore	Rouge-1	Rouge-2	BERTScore	CB.	IFC.
BART		14.21	0.75	10.26	52.24	0.00	12.46	0.23	8.28	53.26	55.47
Longformer		25.08	5.98	17.41	65.80	0.23	27.50	8.07	19.05	67.19	62.69
Qwen2.5-3B	Vanilla	10.97	1.38	5.87	60.4	0.00	11.69	1.35	61.46	71.86	52.34
	SFT	23.79	5.50	17.13	66.91	6.85	23.73	5.33	66.28	74.43	65.23
	RAG	22.45	5.41	15.02	66.17	2.05	23.07	5.69	66.80	74.77	65.95
	CoT	23.00	5.73	15.88	67.27	6.16	24.84	5.78	67.97	75.59	64.40
	KHI-E (Ours)	30.08	8.84	20.87	69.66	56.85	33.20	10.51	70.87	76.84	66.33
	KHI-T (Ours)	30.09	9.21	20.30	69.60	28.08	31.54	9.36	70.35	77.17	66.99
LLaMA3.2-3B	Vanilla	9.33	0.89	6.73	56.73	0.00	10.60	1.23	54.82	62.23	54.77
	SFT	22.61	5.58	15.47	66.86	9.59	22.65	4.66	66.39	74.29	63.21
	RAG	22.09	5.72	14.92	66.92	2.74	22.17	6.31	66.49	76.01	63.97
	CoT	22.49	6.07	15.18	67.15	4.79	23.73	5.69	67.25	75.42	64.26
	KHI-E (Ours)	28.42	7.61	19.15	69.07	51.37	29.95	7.93	69.65	77.19	65.61
	KHI-T (Ours)	28.22	8.28	18.08	69.38	30.82	30.38	8.57	69.87	79.14	69.61
Qwen2-7B	Vanilla	13.23	1.68	7.51	62.05	2.05	11.61	1.51	61.03	72.98	53.04
	SFT	24.89	6.87	17.15	68.12	10.96	25.27	7.86	68.11	76.13	63.87
	RAG	25.38	8.23	17.34	68.70	2.74	27.68	9.05	69.68	75.93	63.89
	CoT	23.36	5.98	15.86	67.49	0.68	25.96	6.79	68.06	76.40	63.86
	KHI-E (Ours)	29.85	8.55	20.88	69.08	52.74	31.92	9.67	69.63	77.47	64.56
	KHI-T (Ours)	30.57	9.58	20.30	69.61	30.82	32.39	9.69	69.77	78.67	66.56
LLaMA3-8B	Vanilla	9.49	0.86	6.42	50.98	0.00	9.49	0.67	58.21	66.65	59.56
	SFT	23.53	5.73	16.45	67.10	15.07	25.27	7.86	68.11	72.51	64.21
	RAG	24.48	6.62	17.40	67.70	3.42	25.32	7.30	67.93	74.86	65.06
	CoT	24.07	6.5	16.88	67.91	5.48	24.84	6.18	67.47	76.02	63.97
	KHI-E (Ours)	28.36	7.84	18.69	69.17	45.89	30.88	10.41	70.25	79.63	65.78
	KHI-T (Ours)	28.87	8.95	19.31	69.48	30.14	31.14	10.07	70.60	77.03	66.59
LawyerLLaMA-13B	Vanilla	14.67	2.18	8.32	62.42	0.68	16.57	2.79	63.10	76.44	55.92
	SFT	27.84	7.80	19.74	69.03	8.90	26.50	7.28	68.70	76.04	65.40
	RAG	26.45	7.88	17.46	69.29	6.16	27.08	8.57	69.05	77.87	64.94
	CoT	26.85	8.09	18.58	69.19	6.16	27.22	8.01	68.54	76.87	62.43
	KHI-E (Ours)	28.66	8.52	20.01	69.03	41.78	29.22	9.44	68.73	78.85	65.80
	KHI-T (Ours)	30.58	9.57	21.18	69.72	35.62	31.27	10.28	69.96	78.32	66.25
Qwen3-14B	Vanilla	15.03	2.51	7.85	62.36	0.68	15.18	2.84	62.17	73.56	56.97
	SFT	25.34	7.07	17.56	68.26	12.33	25.96	7.00	68.49	76.74	64.92
	RAG	23.62	6.06	17.37	67.27	2.74	24.21	5.92	67.65	73.92	64.83
	CoT	24.55	7.06	16.90	67.88	5.48	24.37	6.30	67.02	74.23	65.23
	KHI-E (Ours)	32.55	10.55	22.31	70.73	58.22	33.67	12.06	71.16	79.30	66.62
	KHI-T (Ours)	32.11	10.91	22.30	70.59	20.55	33.90	11.70	70.62	78.04	66.65

prompts used for RAG and CoT are shown in Figure 1.

Evaluation Metrics. Following prior work on natural language generation in LegalAI (Perez-Beltrachini et al., 2021), we employ three categories of automatic evaluation metrics: ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and GPTScore (Brown et al., 2020). For more details on GPTScore, please refer to Appendix D. In addition, to specifically evaluate balance and functional correctness in defence opinion generation, we propose two task-specific metrics: *CoverageBalance* and *Info-Function Consistency*. These metrics are computed on open-source LLM outputs using DefGen-Bench-Label.

CoverageBalance measures whether a generated defence opinion provides balanced coverage of the two core information sources: the indictment and the defendant's stated opinion. Let s_p and s_d denote the semantic similarity scores between the generated defence opinion and the indictment, and between the defence opinion and the defendant's stated opinion, respectively. CoverageBalance is defined as the harmonic mean of the two scores:

$$\text{CoverageBalance} = \frac{2 \cdot s_p \cdot s_d}{s_p + s_d}. \tag{5}$$

This formulation penalizes imbalanced generations that overly emphasize one source while neglecting the other.

Info-Function Consistency evaluates whether each sentence in the generated defence opinion fulfills its intended functional role as annotated by legal experts. For sentence i, let $l_i \in \{C, D, B\}$ denote its label, corresponding to responding to the indictment, following the defendant's stated opinion, or responding both. Let $s_p^{(i)}$ and $s_d^{(i)}$ denote its semantic similarity to the indictment and the defendant's opinion, respectively. The sentence-level score is defined as:

$$s^{(i)} = \begin{cases} s_p^{(i)}, & \text{if } l_i = C \\ s_d^{(i)}, & \text{if } l_i = D \\ \frac{s_p^{(i)} + s_d^{(i)}}{2}, & \text{if } l_i = B \end{cases}, \tag{6}$$

The final Info-Function Consistency score is computed as the mean over all sentences:

$$\text{IFC} = \frac{1}{N} \sum_{i=1}^{N} s^{(i)}. \tag{7}$$

We performed human evaluation validation for Coverage Balance and Info-Function Consistency in Appendix E.

Table 4: Comparison on closed-source and larger LLMs. In this comparison, we apply input contextual enhancement as the enhancement method. R-1/2/L represent ROUGE-1/2/L. BS. and GS. means BERTScore and GPTScore. Pink intensity indicates performance level (darker = better).

Model	R-1	R-2	R-L	BS.	GS.
Qwen-PLUS	24.14	6.40	15.21	67.64	4.67
RAG	23.73	6.27	13.33	67.64	39.33
CoT	23.37	5.22	13.18	66.85	1.33
KHI-T (Ours)	25.09	6.64	15.69	67.61	54.67
DeepSeek-V3.2	24.67	5.68	16.16	66.97	6.00
RAG	24.22	5.60	15.74	66.94	37.33
CoT	20.74	4.48	14.00	65.57	2.00
KHI-T (Ours)	25.41	5.73	16.70	67.17	54.67

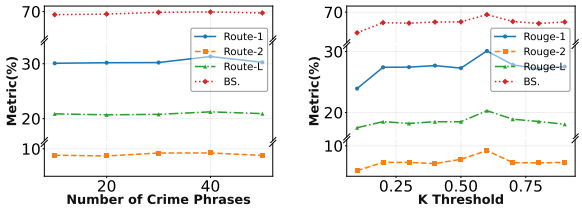
Implementation Details. We use AdamW as the optimizer with a learning rate of 3×10^{-5}. For textual enhancement in closed-source and large LLMs, we set the relevance threshold to $k = 0.6$. To identify the crime type for each case, we select the 10 most salient crime-indicative terms. All experiments are conducted on an NVIDIA A100-SXM4-80GB GPU.

4.2 Overall Evaluation

Tables 3 and 4 present the overall comparison results for all baselines.

For open-source LLMs, the vanilla setting performs poorly on this task, whereas commonly used enhancement strategies, including SFT, RAG, and CoT, yield consistent performance gains. Models in the Qwen family generally outperform LLaMA models of comparable scale. In particular, under the RAG setting, Qwen2-7B achieves a BERTScore of 68.70, surpassing LLaMA3-8B (67.70) under the same configuration. Notably, although the legal-specialized LawyerLLaMA-13B performs worse than Qwen3-14B in the vanilla setting, it achieves substantially better results after applying enhancement strategies, indicating that domain-specific legal knowledge is crucial for this task. In contrast, for closed-source and larger LLMs, RAG and CoT provide only marginal improvements over the vanilla setting.

Based on these results, we find that the proposed KHI method consistently improves the quality of generated defence opinions. For open-source LLMs, KHI brings larger gains, particularly for Qwen2-7B, LLaMA3-8B, and Qwen3-14B. Specifically, Qwen3-14B achieves Rouge-1 scores of 32.55 and 32.11 with KHI-E and KHI-T, respectively, representing about 29% improvements over



(a) Knowledge-sensitive analysis.

(b) Relevance threshold in KHI-T analysis.

Figure 3: Hyperparameter Analysis.

the corresponding baselines. For closed-source and larger LLMs, KHI-T still provides consistent performance gains, demonstrating that the proposed method can be effectively applied to both open-source and closed-source models. The different level of gain is due to the background and training process. More details discussed in Appendix C. Furthermore, with respect to the two proposed balance metrics, KHI also improves the balance of generated defence opinions.

Overall, these results confirm that KHI effectively enhances defence opinion quality while improving both balance and functional correctness.

4.3 Hyperparameter Analysis

Knowledge-sensitive Analysis. Figure 3a shows that increasing the number of crime phrases from the using 10 to 50 results in negligible changes in ROUGE-1/2/L and BERTScore. This observation indicates that the crime identification component remains consistently accurate across a wide range of term counts, and that the downstream defence opinion generation is not sensitive to this parameter. Given this robustness, we select the top 10 salient crime-indicative terms to improve computational efficiency without sacrificing performance.

Relevance Threshold in KHI-T Analysis. Figure 3b shows setting $k = 0.6$ achieves the best overall performance across ROUGE-1/2/L and BERTScore. Smaller values of k tend to introduce excessive highlighting, while larger values risk omitting legally salient indictment components. Therefore, $k = 0.6$ provides a favorable balance between emphasizing legally relevant content and preserving input coherence.

4.4 Case Study

Figure 4 shows a representative case demonstrating how KHI mitigates imbalanced defence opinion generation. The indictment highlights parts

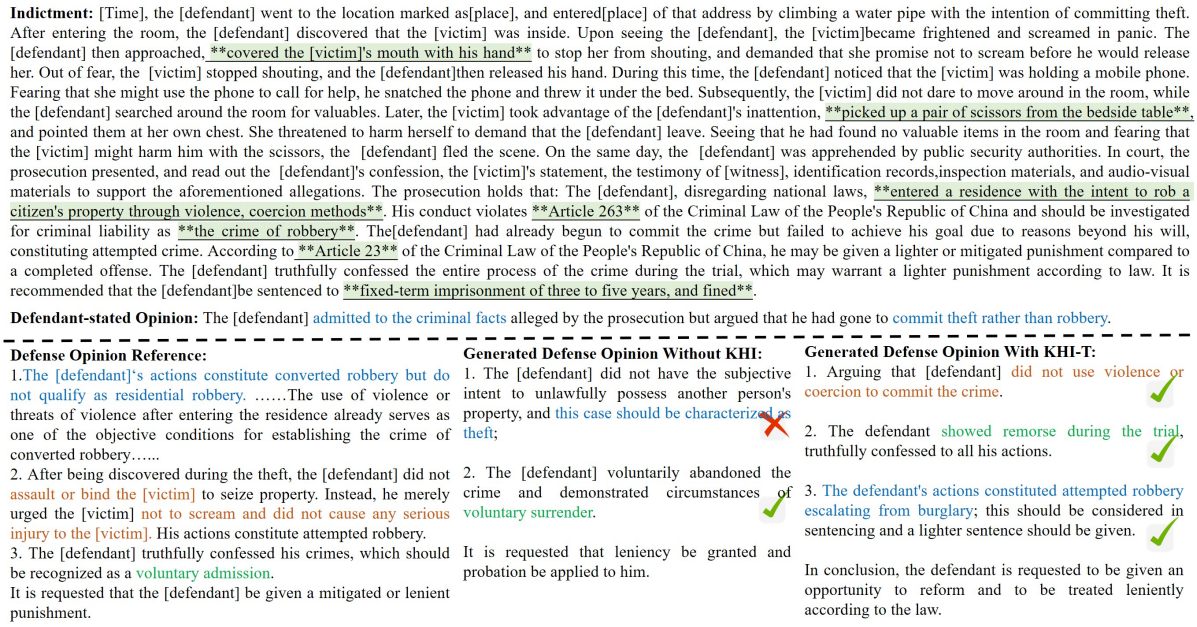


Figure 4: Case study of defence opinion generation. The text is the enhanced indictment part from KHI.

from the proposed KHI that show legally salient elements of attempted robbery, as well as offence escalation from burglary.

Although the defendant-stated opinion admits the factual conduct, it insists that the intent was limited to theft. In line with professional defence practice, the expert-reviewed reference defence does not mechanically follow this denial but instead selects a mitigation-oriented strategy (Babcock, 1983), acknowledging that the statutory elements of attempted robbery are largely satisfied while arguing for leniency based on the incomplete offence and truthful confession (Bibas, 2004). Without KHI, the model over-aligns with the defendant-stated opinion and recharacterizes the conduct as theft, failing to consider the facts mentioned in the indictment. But, within KHI-T can recognizes the attempted robbery charge and shifts its focus toward mitigation arguments, closely matching the reference defence. This case demonstrates that KHI effectively guides the model to balance the defendant's stated position with legally salient indictment components.

5 Related Work

LegalAI Tasks Overview. Existing LegalAI research has explored a broad range of application tasks, including legal judgment prediction (Medvedeva and Mcbride, 2023; Zhong et al., 2018; Lou et al., 2026), judicial fact extraction (Mistica et al., 2020), legal consultation dialogue (Yuan et al., 2025; Li et al., 2025; Sun et al., 2024), and legal text summarization (T.y.s.s et al., 2024b; Kanapala et al., 2019; Sheik and Nirmala, 2021). Although these tasks differ in application scenarios, they are generally formulated under a single-perspective setting, where models generate predictions or texts based on a single input. For instance, legal judgment prediction should, in practice, consider statements from both parties; however, current formulations typically simplify the task by predicting outcomes based on complete and legally established facts.

Legal Knowledge in LegalAI. To improve the quality and reliability of LegalAI systems, prior studies have primarily focused on enhancing LLMs with domain-specific legal knowledge. Representative approaches include incorporating external legal resources—such as statutes and case law, via retrieval-augmented or knowledge, injection mechanisms (Liu et al., 2024b), learning legal-aware or task-specific embeddings to better encode legal semantics (Belfathi et al., 2023; Alberts et al., 2020), and employing prompt engineering or structured reasoning templates to guide model behavior (Yu et al., 2022a; Trautmann, 2023). These methods have been shown to improve performance and legal consistency.

6 Conclusion

In this paper, we propose a novel and meaningful task Defence Opinion Generation, aimed at assist-

ing defendants in protecting their rights. Furthermore, we construct a benchmark DefGen-Bench to evulate current LLMs ability. The DefGen-Bench dataset contains a diverse set of real-world cases across multiple crime categories. Additionally, we propose a general input enhancement framework, KHI. Experiments demonstrate the effectiveness of our method in improving the quality of defence opinion generation. We hope our work in this paper can promte the development of defence generation.

Limitations

Although we follow established legal principles and practice to ensure that defence opinion generation closely reflects real-world proceedings, our formulation does not include evidence analysis. This limitation arises from privacy and sensitivity concerns, as evidentiary materials are often unavailable or cannot be released in publicly accessible datasets. In addition, due to jurisdictional differences, it is challenging to obtain all necessary components, namely indictments, defendant-stated opinions, and defence opinions, to construct other jurisdictions' datasets. Addressing these limitations is left for future work.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (Grant No. 63261068), the National Natural Sinence Foundation of China (Grant No. 72571150) and the Academy for Advanced Interdisciplinary Studies, Nankai University (AAIS-NKU-2025-21).

Ethical Statements

All cases in our benchmark are collected from publicly available sources. We ensure that no personal, sensitive, or private information is included. The benchmark will be released for further inspection to verify compliance with data privacy requirements.

References

Michael Aerni, Javier Rando, Edoardo Debenedetti, Nicholas Carlini, Daphne Ippolito, and Florian Tramer. 2025. Measuring non-adversarial reproduction of training data in large language models. In *International Conference on Representation Learning*, volume 2025, pages 54143–54185.

Houda Alberts, Akin Ipek, Roderick Lucas, and Phillip Wozny. 2020. Coliee 2020: Legal information retrieval and entailment with legal embeddings and boosting. In *JSAI*, pages 211–225. Springer.

Guido Alpa. 1994. General principles of law. *Ann. Surv. Int'l & Comp. L.*, 1:1.

R Harald Baayen, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019(1):4895891.

Barbara Allen Babcock. 1983. Defending the guilty. *Clev. St. L. Rev.*, 32:175.

Anas Belfathi, Nicolas Hernandez, and Laura Monceaux. 2023. Enhancing pre-trained language models with sentence position embeddings for rhetorical roles recognition in legal opinions. In *ASAIL@ICAIL*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Stephanos Bibas. 2004. Plea bargaining outside the shadow of trial. *Harvard Law Review*, pages 2463–2547.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. In *ACL 2024*, pages 177–198.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jaye Ellis. 2011. General principles and comparative law. *Eur. J. Int. Law*, 22(4):949–971.

Shangjie Feng, Buqing Cao, Ziming Xie, Zhongxiang Fu, Zhenlian Peng, and Guosheng Kang. 2024. Llmsarec: large language model with semantic alignment for web service recommendation. *IJWIS*, 21(1):37–53.

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)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Francis Galton. 1885. Anthropometric per-centiles. *Nature*, 31(793):223–225.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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

Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuanjing Huang. 2018. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *COLING 2018*, pages 3703–3714.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artif. Intell. Rev.*, 51(3):371–402.

Anne Lise Kjær. 2000. On the structure of legal knowledge: The importance of knowing legal rules for understanding legal texts. *Lang. Text Knowl. Ment. Model. Expert Commun.*, pages 127–161.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL 2020*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS 2020*, volume 33, pages 9459–9474.

Haitao Li, Yifan Chen, Hu YiRan, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. 2025. Lexrag: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. *SIGIR 2025*.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *TSBO 2004*, pages 74–81.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models.

Bulou Liu, Yiran Hu, Qingyao Ai, Yiqun Liu, Yueyue Wu, Chenliang Li, and Weixing Shen. 2025b. Generating clarifying questions for conversational legal case retrieval without external knowledge. *ACM Trans. Inf. Syst.*, pages 102:1–102:26.

Qian Liu, Hang Yu, Qiqi Wang, Qi Xu, Jinpeng Li, Zhuoqun Zou, Rui Mao, and Erik Cambria. 2025c. Legal knowledge infusion for large language models: A survey. *Information Fusion*.

Shuaiqi Liu, Jiannong Cao, Yicong Li, Ruosong Yang, and Zhiyuan Wen. 2024a. Low-resource court judgment summarization for common law systems. *Inf. Process. Manag.*, page 103796.

Yifei Liu, Yiquan Wu, Ang Li, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2024b. Unleashing the power of llms in court view generation by stimulating internal knowledge and incorporating external knowledge. In *NAACL 2024*, pages 2782–2792.

Fanghao Lou, Qiqi Wang, Guanyu Chen, Kaiqi Zhao, and Huijia Li. 2026. Zipljp: Zipped information processor for legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: A legal case retrieval dataset for chinese law system. In *SIGIR 2021*, pages 2342–2348.

Pere Martra. 2025. Fragile knowledge, robust instruction-following: The width pruning dichotomy in llama-3.2.

Pradeep MD. 2019. Legal research-descriptive analysis on doctrinal methodology. *IJMTS*, 4(2):95–103.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *NLLP 2023*, pages 73–84.

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2025. Query performance prediction using relevance judgments generated by large language models. *ACM Trans. Inf. Syst.*, pages 106:1–106:35.

Petar Milin, Laurie Beth Feldman, Michael Ramscar, Peter Hendrix, and R Harald Baayen. 2017. Discrimination in lexical decision. *PloS one*, 12(2):e0171935.

Meladel Mistica, Geordie Z. Zhang, Hui Chia, Kabir Manandhar Shrestha, Rohit Kumar Gupta, Saket Khandelwal, Jeannie Paterson, Timothy Baldwin, and Daniel Beck. 2020. Information extraction from legal documents: A study in the context of common law court judgements. In *ALTA 2020*, pages 98–103.

Jeanne-Louise Moys. 2014. Typographic layout and first impressions-testing how changes in text layout influence reader's judgments of documents. *Visible Language*, 48(1).

Laura Perez-Beltrachini, Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, and 37 others. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *GEM 2021*, pages 96–120.

Pejman Peykani, Fatemeh Ramezanlou, Cristina Tanasescu, and Sanly Ghanidel. 2025. Large language models: A structured taxonomy and review of challenges, limitations, solutions, and future directions. *Appl. Sci.*, 15(14):8103.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report.

Yong Ren, Jinfeng Han, Yingcheng Lin, Xiujiu Mei, and Ling Zhang. 2022. An ontology-based and deep learning-driven method for extracting legal facts from chinese legal texts. *Electronics*, 11(12).

Justin B Richland. 2013. Jurisdiction: grounding law in language. *Annu. Rev. Anthropol.*, 42(1):209–226.

Reshma Sheik and S Jaya Nirmala. 2021. Deep learning techniques for legal text summarization. IEEE.

Weizhe Shi, Qiqi Wang, Yihong Pan, Qian Liu, and Kaiqi Zhao. 2025. Legalchainreasoner: A legal chain-guided framework for criminal judicial opinion generation. *arXiv preprint arXiv:2509.00783*.

Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. 2024. Lawluo: A multi-agent collaborative framework for multi-round chinese legal consultation. *arXiv preprint arXiv:2407.16252*.

Qwen Team and 1 others. 2024. Qwen2 technical report.

Dietrich Trautmann. 2023. Large language model prompt chaining for long legal document classification.

Santosh T.y.s.s, Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024a. Incorporating precedents for legal judgement prediction on European court of human rights cases. In *Findings of EMNLP 2024*, pages 3743–3750.

Santosh T.y.s.s and Rohit Upadhya. 2025. LexCLiPR: Cross-lingual paragraph retrieval from legal judgments. In *ACL 2025*.

Santosh T.y.s.s, Vatsal Venkatkrishna, Saptarshi Ghosh, and Matthias Grabmair. 2024b. Beyond borders: Investigating cross-jurisdiction transfer in legal case summarization. In *NAACL 2024*, pages 4136–4150.

Xuran Wang, Xinguang Zhang, Vanessa Hoo, Zhouhang Shao, and Xuguang Zhang. 2024. Legalreasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration. *IEEE Access*, 12:166843–166854.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS 2022*, volume 35, pages 24824–24837.

Ryan C Williams. 2022. Jurisdiction as power. *JSTOR*, 89(7):1719–1792.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report.

Fang Yu, Lee Quartey, and Frank Schilder. 2022a. Legal prompting: Teaching a language model to think like a lawyer. *ArXiv*.

Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2022b. Explainable legal case matching via inverse optimal transport-based rationale extraction. In *SIGIR 2022*, pages 657–668.

Weikang Yuan, Kaisong Song, Zhuoren Jiang, Junjie Cao, Yujie Zhang, Jun Lin, Kun Kuang, Ji Zhang, and Xiaozhong Liu. 2025. Lecode: A benchmark dataset for interactive legal consultation dialogue evaluation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *EMNLP 2018*, pages 3540–3549.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. How does NLP benefit legal system: A summary of legal artificial intelligence. In *ACL 2020*, pages 5218–5230.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. Jec-qa: A legal-domain question answering dataset. In *AAAI*.

A Prompt and Crime Term Details

We employ prompt-based structured extraction with LLMs to generate high-coverage keyword tables for each charge. This design is inspired by the framework of LegalChainReasoner (Shi et al., 2025), which "uses unified and constrained

Table 5: Top 10 Core Keywords for Each Crime Category

Crime Category	Top 10 Core Keywords
Intentional Injury	intentional injury, serious injury, minor injury, crime of intentional injury, bodily harm, assault, causing injury, fighting, affray, injury condition
Traffic Accident	traffic accident, traffic offense, hit-and-run, drunk driving, crime of traffic accident, traffic violation, car crash, accident, illegal driving, violation
Robbery	robbery, crime of robbery, armed robbery, violence, coercion, property, knife-wielding, threat, snatching, forcible taking
Fraud	fraud, crime of fraud, telecom fraud, online fraud, deception, fabrication, concealment, obtaining by fraud, falsehood, swindle
Rape	rape, crime of rape, sexual assault, against victim's will, violence or coercion, indecent assault, sexual intercourse, forced act, violation, sexual violation
Theft	theft, crime of theft, stealing, illegal possession, public/private property, secret taking, pickpocketing, burglary, relatively large amount, repeated theft
Murder	intentional homicide, crime of intentional homicide, killing, deprivation of life, unlawful deprivation of life, causing death, homicide act, murder case, killing conduct, death caused
Life/Health Rights	right to life, right to bodily integrity, right to health, personality rights, personal injury, tort, civil liability, compensation, damages, tort liability

prompts to structure legal provisions into legal chains." We adopt similar prompting strategies. In contrast to generating legal chain triples, our approach produces a crime-indicative keyword set: for each charge, we generate 100 keywords ordered by importance, which support case type classification and subsequent statute similarity enhancement via embedding reweighting. Our partial precise term lists are shown in the Table 5 and the model prompt is shown below:

▶ **Prompt:** *You are a senior legal expert specializing in Chinese criminal law. Please generate 100 keywords for the crime of intentional injury under Chinese criminal law, sorted by importance in layers.*

To further ensure reproducibility and stability, we use the same prompt to generate 100 keywords for the crime of intentional injury using DeepSeek-V3.2, GPT-4, and Qwen-3.5-plus respectively. The overlap rate among the keywords generated by the three models reaches 83%, which demonstrates that there is no significant discrepancy among keywords produced by different LLMs.

B Design Analysis

Theorem 1 (Semantic Focus Theorem) *Let* $\mathbf{E}_c$ *be the original input embedding and* $\hat{\mathbf{E}}_c$ *be the knowledge-enhanced embedding. For any attention head in a Transformer layer, if the query* q *originates from the current generation position in the decoder and the key* $k_j = \hat{\mathbf{E}}_{cj}$*, then: Attention weights for claim tokens relevant to legal statutes are relatively enhanced. Attention weights for irrelevant tokens are relatively suppressed.*

Proof 1 *The attention score is:*

$$s_j = \frac{q^T k_j}{\sqrt{d}} = \begin{cases} r_j^{(c)} \cdot \frac{q^T \mathbf{E}_{cj}}{\sqrt{d}}, & j \in \mathcal{I}_s, \\ \frac{q^T \mathbf{E}_{cj}}{\sqrt{d}}, & j \notin \mathcal{I}_s, \end{cases} \tag{8}$$

Let $a_j = \frac{q^T \mathbf{E}_{cj}}{\sqrt{d}}$ *be the original attention score. The attention weight after softmax is:*

$$w_j = \frac{\exp(s_j)}{\sum_{k=1}^{n} \exp(s_k)} \tag{9}$$

For $j \in \mathcal{I}_s$*: If* $r_j^{(c)} \approx 1$ *(high relevance), then* $s_j \approx a_j$*. If* $r_j^{(c)} \ll 1$ *(low relevance), then* $s_j < a_j$*. Thus:*

$$\frac{w_j}{w_k} = \frac{\exp(s_j)}{\exp(s_k)} = \exp(s_j - s_k). \tag{10}$$

For $j \in \mathcal{I}_s$*,* $k \notin \mathcal{I}_s$*, if* $r_j^{(c)} > 0$ *and* a_j *is comparable to* a_k*, then* s_j *may be reduced relative to* a_j*. However, tokens with high relevance (*$r_j^{(c)} \approx 1$*) retain their weight, while tokens with low relevance are suppressed.* □

This indicates that the model focuses more on parts of the indictment that are relevant to legal statutes during generation.

Theorem 2 (Gradient Focus Theorem) *By scaling the embeddings of legally relevant words in the indictment between (0,1), backpropagation effectively applies gradient weighting: gradients for highly relevant words propagate normally, while gradients for low-relevance words approach zero. This forces the model to learn only from legally critical points while ignoring irrelevant noise.*

Proof 2 *Let the generation target be the defence opinion y_i, which should respond to the legally relevant points in the indictment. Define the generation loss function as cross-entropy loss $\mathcal{L}$.*

Scaling the input embeddings is equivalent to introducing a knowledge-guided weight mask:

$$\hat{\mathbf{E}}_c = \mathbf{M} \odot \mathbf{E}_c, \tag{11}$$

where $\mathbf{M}$ is a diagonal matrix with $M_{jj} = r_j^{(c)}$ for $j \in \mathcal{I}_s$, and $M_{jj} = 1$ otherwise.

This operation effectively scales the gradient during backpropagation:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{E}_{cj}} = \begin{cases} r_j^{(c)} \cdot \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{E}}_{cj}}, & j \in \mathcal{I}_s, \\ \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{E}}_{cj}}, & otherwise, \end{cases} \tag{12}$$

Therefore: For tokens with $r_j^{(c)} \approx 0$, the gradient is approximately zero, meaning the model does not learn from these "noise" tokens. For tokens with $r_j^{(c)} \approx 1$, the gradient propagates normally, allowing the model to focus on learning how to generate responses based on these key points. □

Theorem 3 (Selective Enhancement Theorem) *Only enhancing the indictment is equivalent to preemptively increasing the signal-to-noise ratio at the input stage.*

Proof 3 *Let the defined task be:*

$$y_i = F(s_i, b_i), \tag{13}$$

where s_i is the indictment and b_i is the defendant's opinion. According to the task definition, the defence opinion should primarily address the allegations in s_i and be based on the facts in b_i. However, the legal response logic should originate from the correspondence between s_i and legal statutes.

Thus: Enhancing s_i with knowledge aligns the indictment with legal statutes during encoding, facilitating later generation. Not scaling b_i avoids introducing the defendant's non-legal expressions as irrelevant legal signals, preventing noise interference.

From an information-theoretic perspective, this method is equivalent to injecting prior knowledge at the input stage, improving the signal-to-noise ratio (SNR):

$$SNR_{enhanced} = \frac{\sum_{j \in \mathcal{I}_s} (r_j^{(c)})^2 \|\mathbf{E}_{cj}\|^2}{\sum_{j \notin \mathcal{I}_s} \|\mathbf{E}_{cj}\|^2 + \sum_{j \in \mathcal{I}_s} (1 - r_j^{(c)})^2 \|\mathbf{E}_{cj}\|^2}, \tag{14}$$

After enhancement, tokens in the indictment with low relevance contribute less to the noise, thereby increasing the SNR.

Table 6: Contrast with closely related legal knowledge guided generation framework LegalChainReasoner. Pink intensity indicates performance level (darker = better).

Model	R-1	R-2	R-L	BS.
Qwen-PLUS (Vanilla)	24.14	6.40	15.21	67.64
RAG	23.73	6.27	13.33	67.64
CoT	23.37	5.22	13.18	66.85
LegalChainReasoner	22.95	5.41	15.25	66.91
KHI-T (Ours)	25.09	6.64	15.69	67.61
DeepSeek-V3.2 (Vanilla)	24.67	5.68	16.16	66.97
RAG	24.22	5.60	15.74	66.94
CoT	20.74	4.48	14.00	65.57
LegalChainReasoner	20.21	4.55	11.60	66.01
KHI-T (Ours)	25.41	5.73	16.70	67.17

C External Results and Analysis

C.1 Larger LLMs results analysis

In Table 4, we observe that the improvements in ROUGE and BERTScore brought by KHI-T over the baselines setting are limited on closed-source and larger LLMs. We believe this because such models have already encountered these cases during pretraining. To verify this hypothesis, we follow prior work (Aerni et al., 2025) and evaluate the models' ability to reproduce case facts by providing only the first 30 words of the case description and calculating the reproduction rate between the generated content and the original case facts. The results show that more than 30% of the generated content aligns with the original case facts, indicating that large proprietary models already possess substantial prior knowledge of these cases, thereby limiting the effectiveness of additional enhancements.

C.2 Contrast with existing work

Our KHI method differs from existing salience-guided attention reweighting approaches along two key dimensions. First, in terms of the signal source, the relevance scores in KHI are derived from externally retrieved statutory provisions aligned with the identified crime category, rather than from model-internal attention distributions or gradient-based salience signals. Second, in terms of the application mechanism, KHI selectively enhances only the indictment component under a dual-perspective setting, while leaving the defendant's stated opinion unchanged. This statute-grounded modulation is designed to rebalance legally structured reasoning across conflicting inputs, rather than simply amplifying salient tokens within a single text.

Table 7: Performance across crime categories. Darker color indicates better performance.

Crime	R1			R2			RL			BERT		
	CoT	KHI-E	KHI-T	CoT	KHI-E	KHI-T	CoT	KHI-E	KHI-T	CoT	KHI-E	KHI-T
Intentional Injury	22.33	27.30	27.25	5.95	7.68	8.06	14.65	18.02	16.80	67.06	68.61	69.12
Murder	24.75	29.01	29.31	6.28	6.15	8.46	17.33	20.50	19.70	67.56	69.33	69.00
Traffic Accident	19.41	25.35	26.20	5.02	6.18	7.29	11.97	16.16	16.63	65.89	67.56	68.89
Robbery	24.11	32.91	27.41	6.49	10.12	8.36	16.93	22.43	17.05	68.31	70.67	69.78
Rape	22.12	28.70	31.15	6.23	7.24	7.97	14.77	18.74	20.15	66.96	69.46	69.88
Fraud	24.21	31.08	30.13	6.89	8.61	10.23	17.50	22.86	21.59	68.01	69.90	70.47
Theft	23.12	28.63	34.12	9.44	11.68	12.84	19.34	20.27	22.87	67.18	72.06	70.97
Life/Health Rights	22.99	21.26	25.94	7.27	3.46	7.39	16.86	13.39	18.45	67.50	67.16	68.93

We also added closely related legal knowledge guided generation framework- LegalChainReasoner (Shi et al., 2025) as an additional baseline within the directly utilize provisions (RAG) and CoT thinking in the Table 6. Specifically, following the LegalChainReasoner, we constructed legal chains for the crimes involved in our DefGen-Bench. We then utilize these chains as knowledge (Liu et al., 2025c) in defense opinion generation.

The additional comparison shows that KHI achieves higher performance than LegalChainReasoner across the evaluated metrics on this task.

C.3 Detailed results in crimes

The eight categories covered in DefGen-Bench are: Murder, Traffic Accident Crime , Robbery, Fraud, Rape, Theft , Intentional Homicide , and Disputes over Life/Health/Body Rights. The Table 7 shows the performance results of KHI for 8 crimes on the LlaMa3.2-3B model, we also report the CoT in different crimes.As shown in the table, compared with the CoT, KHI achieves better performance across different crime categories.

C.4 KHI under vanilla settings

We also conducted additional experiments by applying KHI directly to the vanilla models without SFT fine-tuning. Specifically, all experiments were conducted on the same Def-Gen dataset used in the main paper. We evaluated all six backbone models reported in the original experiments (LLaMA3-8B, LawyerLLaMA-13B, Qwen3-14B, Qwen2-7B, Qwen2.5-3B, and LLaMA3.2-3B). The "Vanilla" setting corresponds to the LLMs without fine-tuning. The "KHI-E (vanilla)" and "KHI-T (vanilla)" settings apply the KHI embedding rescaling and textual highlighting mechanisms directly to these vanilla models without any supervised fine-tuning or additional training, keeping all other generation configurations identical.

Table 8: Effect of **vanilla** setting KHI on open-source LLMs. Darker color indicates better performance. Blue rows denote our methods (KHI-E and KHI-T).

Model	Setting	R-1	R-2	R-L	BS.
LLaMA3 -8B	Vanilla	9.49	0.86	6.42	50.98
	KHI-E	15.33	3.17	9.62	59.07
	KHI-T	11.26	1.59	8.68	54.73
Lawyer LLaMA -13B	Vanilla	14.67	2.18	8.32	62.42
	KHI-E	20.17	3.75	12.09	64.62
	KHI-T	17.87	2.83	11.75	61.85
Qwen3 -14B	Vanilla	15.03	2.51	7.85	62.36
	KHI-E	18.14	3.60	9.92	64.32
	KHI-T	16.95	3.14	9.19	63.85
Qwen2 -7B	Vanilla	13.23	1.68	7.51	62.05
	KHI-E	21.38	4.34	11.27	66.25
	KHI-T	18.90	3.73	10.58	64.82
Qwen2.5 -3B	Vanilla	10.97	1.38	5.87	60.40
	KHI-E	19.51	3.67	10.62	65.24
	KHI-T	16.73	2.90	9.86	63.37
LLaMA 3.2-3B	Vanilla	9.33	0.89	6.73	56.73
	KHI-E	16.19	2.75	11.75	61.18
	KHI-T	16.76	2.99	10.79	62.58

This controlled comparison isolates the effect of the KHI mechanism itself, allowing us to verify that performance gains stem from the knowledge-guided enhancement rather than from SFT adaptation. The experimental results are shown in the following Table 8.

As shown in the table , both KHI-E and KHI-T consistently improve performance across all six open-source LLMs under the vanilla setting. For example, on Qwen2-7B, ROUGE-1 increases from 13.23 to 21.38 with KHI-E and to 18.90 with KHI-T, while BERTScore improves from 62.05 to 66.25. Similar trends are observed across model sizes, indicating that the highlighting mechanism independently contributes to performance improvements rather than merely amplifying SFT effects. Of course, there remains a certain performance gap between the non-SFT setting and the SFT setting. By combining our KHI with SFT, we achieve

Table 9: Average GPTScore across multiple judge models. Mean denotes the average of three LLM judges.

Model	Method	Mean	Std. Dev.
BART	—	0.35	0.49
Longformer	—	1.12	1.42
Qwen2-7B	Vanilla	3.46	3.55
	SFT	7.12	3.72
	RAG	7.86	4.97
	CoT	4.39	2.63
	KHI-E (Ours)	48.14	4.71
	KHI-T (Ours)	27.64	3.40
LLaMA3.2-3B	Vanilla	0.69	0.98
	SFT	5.97	3.07
	RAG	7.86	4.00
	CoT	7.15	2.39
	KHI-E (Ours)	48.37	7.94
	KHI-T (Ours)	28.33	7.03

Table 10: The correlation of the Coverage Balance and Info-Function Consistency metrics with perceived legal adequacy. Darker color indicates better performance.

Model	Method	CB	IFC	Expert
BART	—	53.26	55.47	0.00
Longformer	—	67.19	62.69	0.00
Qwen2-7B	Vanilla	72.98	53.04	0.00
	SFT	76.13	63.87	6.25
	RAG	75.93	63.89	12.50
	CoT	76.40	63.86	6.25
	KHI-E (Ours)	77.47	64.56	31.25
	KHI-T (Ours)	78.67	66.56	43.75
LLaMA3.2-3B	Vanilla	62.23	54.77	0.00
	SFT	74.29	63.21	8.33
	RAG	76.01	63.97	10.42
	CoT	75.42	64.26	8.33
	KHI-E (Ours)	77.19	65.61	22.92
	KHI-T (Ours)	79.14	69.61	50.00

state-of-the-art (SOTA) performance.

D GPTScore Usage Details

To evaluate the balance and functional correctness of the generated content, we adopt an LLM-based assessment method known as GPTScore (Fu et al., 2024). This approach leverages LLMs to assess the overall quality of generated defence opinions. Following recent research, we apply the GPTScore-pairwise (Li et al., 2024) variant, which has been shown to be more effective. This method compares multiple generated denfence opinion and selects the best output. The below figure illustrates the prompt used for pairwise comparison.

► **Indictment:** *[Indictment context]...*
► **Defendant-stated opinion:** *[Defendant-stated opinion context]...*
► **Reference defence opinion:** *[Reference defence opinion content]...*
► **Generate 1:** *[Generate 1 context]...*
► **Generate 2:** *[Generate 2 context]...*
► **...**
► **Generate n:** *[Generate n context]...*
Please select the best text that, based on the given [Indictment] , [Defendant-stated opinion], and [Defence opinion]. The selected text should best balance and align with the [Defence opinion] across three dimensions: content, structure, and length.

Furthermore, to enhance the reliability and reproducibility of our method, we have extended the GPTScore evaluation by including the Qwen3.5-Plus and GPT-4 models. We employ a dedicated prompt designed for assessing the legal reasoning quality of generated text to evaluate the outputs of LLaMA3.2-3B and Qwen2-7B. The mean and standard deviation of the GPTScore across the three models are presented in Table 9.

E Human Evaluation Validation

Validating the correlation between Coverage Balance / Info-Function Consistency and perceived legal adequacy through human evaluation can further strengthen the persuasiveness of our work. Therefore, we invited legal experts to conduct a small-scale human evaluation on 50 randomly sampled cases. The results are shown in Table 10.

Our calculations show that the Spearman's rank correlation coefficient between ExpertScore and CB. (Coverage Balance) is 0.8754, while the coefficient between ExpertScore and IFC. (Info-Function Consistency) is 0.9422. This indicates that both the Coverage Balance and Info-Function Consistency metrics exhibit a significant positive correlation with perceived legal adequacy.