A CAUSAL LEGAL REASONING METHOD FOR JUDICIAL SUBJECTIVE QUESTIONS VIA KEY LEGAL FACT IDENTIFICATION

Jinze Sang*
China University of Political Science and Law
Beijing, China
2401421143@cupl.edu.cn

Jiawen Zhang†
Zhejiang University
Hangzhou, China
zhangjiawenzju@zju.edu.cn

ABSTRACT

LLMs are increasingly used in legal tasks, yet they still rely primarily on data-driven learning, associative pattern extraction, and probabilistic generation. While effective in open-domain question answering, this mechanism tends to treat background narratives, superficially relevant details, and legally decisive facts in a similar manner in judicial subjective questions, leading to misplaced reasoning focus, weak rule grounding, and unstable conclusions. Causal research suggests that association is not causation: compared with spurious associations induced by confounding or selection bias, causal relations are generally more interpretable, more robust, and more useful for decision-making. Motivated by this perspective, we propose a causal legal reasoning method for judicial subjective questions centered on key legal fact identification. Instead of generating answers directly from raw case descriptions, our framework decomposes legal reasoning into four components: legal fact extraction, key legal fact identification, rule grounding, and legal judgment generation. A fact is treated as a key legal fact if changing it, while keeping other core conditions approximately fixed, would alter the legal assessment or final conclusion. This intermediate layer enables more targeted legal retrieval, norm application, and answer generation. To support this framework, we construct a task-oriented intermediate representation for judicial subjective questions, including legal facts, key legal facts, rule references, and gold answers. Experiments on the CAIL2025 judicial subjective-question dataset show that the proposed framework achieves strong end-task performance across multiple backbone models. Ablation results further show that both key legal fact identification and retrieval grounding contribute substantially to judicial scoring. These findings suggest that explicit fact-centered reasoning provides a feasible way to improve legal answer generation for complex judicial subjective questions.

1 INTRODUCTION

In recent years, large language models (LLMs) have achieved remarkable progress in question answering, text generation, and complex reasoning, and have rapidly pushed legal AI from retrieval-oriented systems toward generative systems. In the legal domain, substantial efforts have been devoted to legal question answering, legal judgment prediction, legal retrieval, statute matching, and legal reasoning (Aletras et al., 2016; Xiao et al., 2018; Chalkidis et al., 2022; Guha et al., 2023). In particular, with the emergence of legal benchmarks and legal-domain LLMs, legal NLP has gradually shifted from asking whether legal texts can be processed to asking whether reliable legal reasoning can be performed (Xiao et al., 2018; Chalkidis et al., 2022; Guha et al., 2023).

Among legal reasoning tasks, judicial subjective questions are especially challenging. Unlike ordinary legal QA, they do not merely ask whether a model can retrieve a relevant rule or produce a short answer. Instead, the model must read complex case materials, identify the facts that are legally deci-

*Corresponding author.
†Equal contribution.

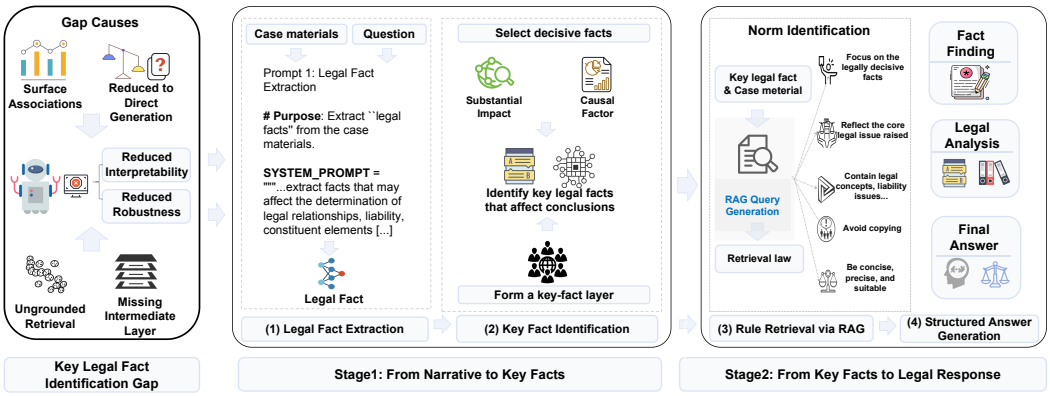

Figure 1: A Causal Legal Reasoning Method for Judicial Subjective Questions via Key Legal Fact Identification

sive, organize a path from facts to rules, and generate a conclusion-oriented response that can satisfy rubric-based judicial scoring. Prior studies suggest that both legal judgment prediction and legal QA are highly sensitive to case complexity, task structure, and explainability requirements (Xiao et al., 2018; Chalkidis et al., 2022). This suggests that the central difficulty of judicial subjective questions lies not merely in insufficient legal knowledge, but in whether the model can distinguish legally decisive facts from background details within long narratives, multiple actors, intertwined actions, and layered legal relations.

This difficulty is closely related to the underlying mechanism of current LLMs. At their core, LLMs are still built upon large-scale data-driven training, statistical association learning, and probabilistic autoregressive generation: they learn token-level dependencies and broader linguistic regularities from massive corpora, and then produce outputs by predicting the most probable continuation under a given context (Vaswani et al., 2017; Brown et al., 2020; Wu et al., 2024). Such a mechanism is highly effective for tasks that benefit from broad exposure to textual patterns, but it does not naturally distinguish which factual details are legally outcome-determinative and which are merely narratively salient. In judicial subjective questions, if all case facts are treated as roughly equal in the reasoning context, the model may mistake superficially relevant details for decisive grounds, which in turn leads to misplaced focus, drifting argumentation, and weak rule application.

A causal perspective helps clarify this problem. Classical causal theory suggests that observed associations may arise from different sources, including genuine causal relations, spurious associations induced by confounding bias, and spurious associations induced by selection bias (Pearl, 2010; Pearl & Mackenzie, 2018). Among these, causal relations are generally more interpretable, more robust across changing environments, and more useful for decision-making, whereas spurious associations may break down once the data distribution shifts (Schölkopf et al., 2021; Wu et al., 2024). This distinction is particularly relevant to legal reasoning. Legal reasoning is not simply a matter of reproducing the most probable answer pattern. Rather, it requires the model to explain why certain facts matter to legal assessment, attribution, and norm application, while other facts do not.

Building on this insight, we adopt a simple causal principle for judicial subjective questions: a fact is important if changing it, while other core conditions remain approximately fixed, would change the legal assessment or final conclusion. We do not construct a full structural causal model for legal text. Instead, we use this control-variable perspective to distinguish facts that materially affect rule application or legal outcome from facts that serve mainly as background or narrative detail. This yields a more explicit intermediate structure for legal reasoning and provides a principled basis for fact selection before rule retrieval and answer generation.

Accordingly, we propose a causal legal reasoning method for judicial subjective questions via key legal fact identification. Rather than directly generating an answer from the raw case description, our framework organizes the reasoning process into four components: legal fact extraction, key legal fact identification, rule grounding, and legal judgment generation. More specifically, Stage 1 transforms the raw case narrative into a set of key legal facts, and Stage 2 uses these facts to retrieve

relevant legal rules and generate the final structured response. In this way, the framework introduces an explicit path from facts to rules to conclusions, instead of implicitly treating all case details as equally relevant.

This design is motivated by both legal reasoning and model behavior. From the perspective of legal reasoning, judicial subjective questions are evaluated not only by whether the answer is fluent, but by whether it identifies the right issues, applies the right rules, and covers the required scoring points. From the perspective of model behavior, an explicit intermediate layer can reduce the influence of weakly relevant details and provide a clearer basis for downstream retrieval and answer generation. Compared with direct generation or purely knowledge-augmented methods, our framework places the emphasis on selecting legally decisive facts before invoking legal rules and producing the final answer.

To operationalize this idea, we build a task-oriented intermediate representation on top of the original judicial subjective-question dataset, including legal facts, key legal facts, rule references, and gold answers. This representation lets us model legal reasoning as a structured process rather than a single-step generation task. We then evaluate the proposed framework on the CAIL2025 judicial subjective-question dataset. The results show strong end-task performance across multiple backbone models. Ablation results further show that removing key legal fact identification or RAG leads to clear drops in judicial scoring. Together, these findings suggest that explicit fact-centered reasoning better supports answer generation in complex judicial subjective questions.

The main contributions of this paper are threefold. First, we revisit judicial subjective-question reasoning from a causal perspective and argue that the main bottleneck lies not simply in legal knowledge acquisition, but in identifying the key legal facts that materially shape legal conclusions. Second, we propose a two-stage legal reasoning framework centered on key legal fact identification, in which rule grounding and answer generation are built on selected decisive facts rather than the full factual narrative. Third, through task-oriented intermediate representation and multi-dimensional experiments, including in-framework comparison and ablation analysis, we show the value of explicit key-fact-centered reasoning for judicial subjective answer generation.

2 Related Work

2.1 Legal Prediction and Question Answering

Early legal AI research mainly focused on Legal Judgment Prediction (LJP) and Legal Question Answering (LQA). These tasks were usually framed as text classification, label prediction, or retrieval-and-matching problems. Case facts were encoded to predict charges, applicable statutes, or penalty terms. For example, Aletras et al. (2016) showed that judicial decisions can be predicted with NLP methods. Zhong et al. (2018) proposed TopJudge, which jointly modeled charges, relevant statutes, and prison terms in a multitask framework and substantially advanced Chinese legal judgment prediction. Compared with LJP, LQA requires stronger reasoning. It asks models not only to retrieve legal knowledge, but also to interpret legal concepts, organize arguments, and generate explainable answers under factual scenarios. JEC-QA significantly advanced legal QA research in the context of China's national judicial examination (Zhong et al., 2020). It showed that legal QA requires not only statute memorization, but also multi-hop reasoning, conceptual distinction, and complex reading comprehension (Zhong et al., 2020). Overall, prior work on LJP and LQA established core tasks, datasets, and evaluation paradigms for legal intelligence. However, most approaches still map case text directly to labels or answers and pay limited attention to which facts actually drive legal conclusions.

2.2 LLM Legal Reasoning

With the rise of large language models, legal NLP has shifted from traditional discriminative methods to generative modeling. LexGLUE systematically benchmarked legal language understanding tasks and showed that legal-domain-specific models consistently outperform generic pretrained models (Chalkidis et al., 2022). LegalBench extended this line of work by providing a broader framework for evaluating legal reasoning abilities of LLMs across diverse tasks (Guha et al., 2023). LawBench further assessed legal LLMs from multiple dimensions, including legal knowledge, analytical reasoning, and legal application, and revealed their limitations in complex legal reasoning

settings (Fei et al., 2024). At the model level, the legal domain has also seen domain-adapted large language models. A representative example is Lawyer LLaMA. It incorporates legal knowledge during continual training and further improves professional legal capabilities through supervised fine-tuning (Huang et al., 2023). These studies show that domain adaptation can improve legal fluency and domain familiarity. Yet stronger domain knowledge does not by itself ensure that a model relies on the legally decisive facts of a case. This issue is especially salient in judicial subjective questions, which require explicit justification rather than merely fluent generation. The challenge is therefore not only whether legal language can be generated, but whether legal conclusions are grounded in the right facts and the right normative basis.

2.3 Causality and Legal Reasoning

Causal reasoning starts from a basic distinction. Association is not causation. Compared with spurious associations induced by confounding bias or selection bias, genuine causal relations are generally more interpretable, more robust, and more useful for decision-making (Pearl & Mackenzie, 2018; Schölkopf et al., 2021). This perspective has increasingly influenced research on NLP and large language models. Kiciman et al. (2023) argued that although LLMs can sometimes produce causally plausible arguments, their underlying mechanism still mainly depends on large-scale statistical associations. This limits their reliability on tasks requiring stable causal judgment. More recent surveys examined the intersection of causality and LLMs from several angles. These include causal analyses of LLM weaknesses, causal methods for improving LLM reasoning, and the use of LLMs in traditional causal inference tasks (Liu et al., 2025; Ma, 2025).

In legal AI, however, explicit causal reasoning remains underexplored. Existing work more often focuses on fairness, debiasing, explanation generation, or rule application. It rarely incorporates causal reasoning directly into judicial subjective-question answering. Most methods treat case facts as a whole and generate answers directly. They do not explicitly distinguish legally decisive facts from correlated background details. Our work addresses this gap. We introduce key legal fact identification as an explicit intermediate layer and use it to connect fact selection, rule grounding, and answer generation in a unified reasoning framework.

3 Methodology

3.1 Problem Formulation

Given a judicial subjective question, our goal is to generate a legally sound answer based on the case materials and the legal issue to be resolved. Formally, each sample is represented as

$$x = (F, Q),$$

where F denotes the full case narrative and Q denotes the legal question. The final target is a reference answer A.

Traditional LLM-based legal reasoning methods typically take (F, Q) as input and directly generate A in a single step. However, such direct generation may rely excessively on surface-level semantic associations in the case description, rather than on the legally decisive facts that materially determine the legal conclusion. As a result, the generated answer may be distracted by irrelevant factual details, making the reasoning process less interpretable and less robust.

To address this issue, we introduce an explicit intermediate reasoning structure. Instead of directly mapping (F, Q) to the final answer, we decompose the reasoning process into two stages with four internal components:

$$(F, Q) \to M \to K, \qquad (K, Q) \to R \to Y,$$

where M denotes the extracted legal facts, K denotes the identified key legal facts, R denotes the retrieved legal rules or statutory references, and Y denotes the structured answer consisting of fact finding, legal analysis, and final answer. Although the framework contains four reasoning components, it is implemented and evaluated as a unified two-stage pipeline: Stage 1 transforms the raw case narrative into key legal facts, and Stage 2 transforms key legal facts into the final structured legal response.

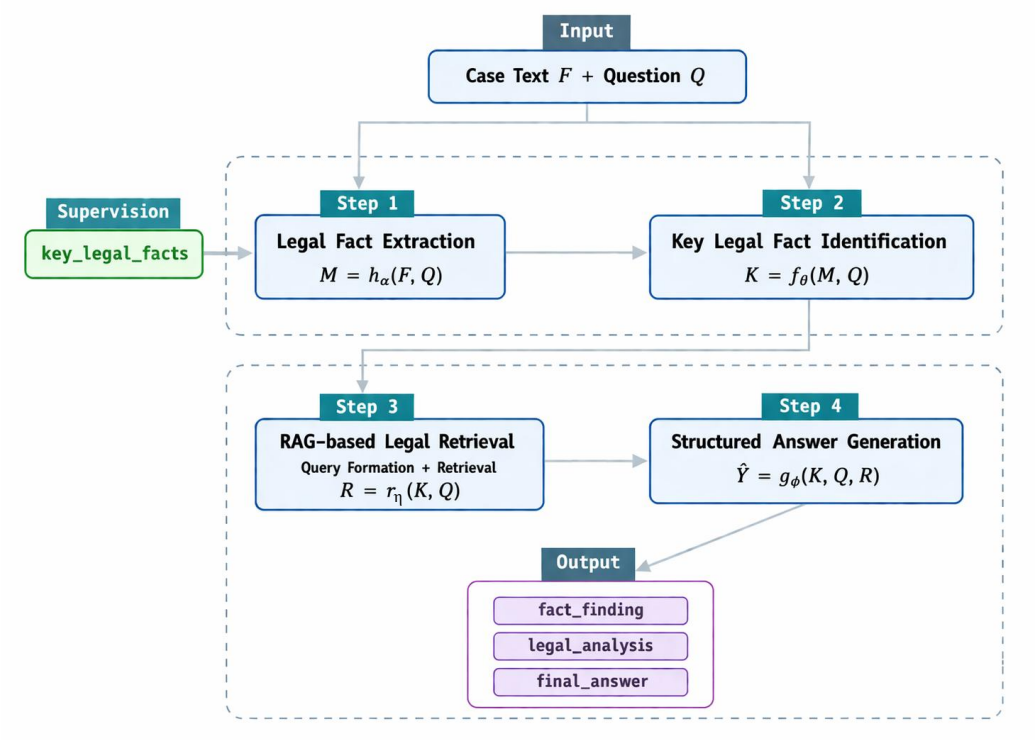

Figure 2: Overview of the proposed causality-inspired legal reasoning framework. Stage 1 transforms the raw case narrative into key legal facts, and Stage 2 retrieves relevant legal rules based on the identified key legal facts and generates the final structured legal response.

3.2 A Causal View of Legal Reasoning

Our method is motivated by a simple causal intuition: not all facts mentioned in a case contribute equally to the legal conclusion. Some facts are legally decisive, while others function only as background information, contextual description, or narrative detail. From this perspective, legal reasoning can be understood as the process of distinguishing legally decisive facts from facts that are only superficially associated with the final answer.

This intuition is closely related to a basic idea in causal reasoning: if one factual element were changed while the other core conditions remained approximately fixed, and such a change would materially alter the legal assessment, the path of rule application, or the final conclusion, then that element should be treated as a key legal fact. By contrast, if modifying a fact would not substantially change the legal answer, then that fact is less central to the reasoning process.

Accordingly, rather than allowing the model to rely indiscriminately on all details in the case narrative, we explicitly guide it to focus on the subset of facts that materially shape legal judgment. In our framework, causal reasoning is not implemented as formal causal graph construction or strict causal effect estimation. Instead, it is operationalized as a structured fact-selection principle: the model is encouraged to identify those factual elements whose variation would plausibly change the legal outcome.

3.3 Stage I: Legal Fact Extraction and Key Legal Fact Identification

The first stage transforms the raw case narrative into a compact representation of legally decisive facts. This stage contains two components.

Step 1: Legal Fact Extraction. Given the full case narrative F and the legal question Q, the model first extracts a set of legal facts:

$$M = h_\alpha(F, Q),$$

where h_α denotes the legal fact extraction function. The purpose of this step is not to answer the legal question directly, but to convert the original narrative into a more structured factual layer by filtering out purely emotional, decorative, or weakly relevant descriptions.

Step 2: Key Legal Fact Identification. Based on the extracted legal facts M and the legal question Q, the model further identifies the subset of key legal facts:

$$K = f_\theta(M, Q),$$

where f_θ denotes the key fact identification function. The model is encouraged to select those facts that satisfy the following principle: if a fact were removed, altered, or replaced while the other core conditions remained approximately unchanged, the legal assessment or final conclusion would likely change accordingly.

The output of Stage 1 is therefore not a final legal answer, but a structured intermediate representation of the case. This intermediate layer serves two purposes. First, it reduces the influence of irrelevant or weakly relevant details in the original narrative. Second, it makes the subsequent reasoning process more transparent, because the final answer can be traced back to an explicitly selected set of key legal facts.

3.4 Stage II: Rule Retrieval and Answer Generation

After identifying the key legal facts, we use them as the basis for legal retrieval and answer generation. This stage also contains two components.

Step 3: Rule Retrieval via RAG. Given the key legal facts K and the legal question Q, the system first forms a retrieval-oriented legal query and then retrieves the most relevant legal rules, statutory provisions, or normative references:

$$R = r_\eta(K, Q),$$

where r_η denotes the retrieval function. Unlike direct-answer baselines that rely only on the raw case narrative, this step performs legal retrieval based on the filtered, legally decisive facts, thereby making rule grounding more targeted and less vulnerable to irrelevant case details.

Step 4: Structured Answer Generation. The model then generates the final structured legal response conditioned on the key legal facts, the legal question, and the retrieved legal rules:

$$\hat{Y} = g_\phi(K, Q, R),$$

where g_ϕ denotes the answer generation function. The output $\hat{Y}$ consists of three parts:

$$\hat{Y} = (\hat{y}_f, \hat{y}_l, \hat{y}_a),$$

where $\hat{y}_f$ denotes fact finding, $\hat{y}_l$ denotes legal analysis, and $\hat{y}_a$ denotes the final answer. In evaluation, the main target remains the reference answer A, while the structured output provides additional interpretability for the reasoning process.

This design has two major advantages. First, it improves interpretability by explicitly grounding answer generation in both key legal facts and retrieved legal rules. Second, it improves robustness by reducing the chance that the model will be distracted by irrelevant factual details in the original case description.

Overall, the proposed framework transforms legal reasoning from one-step answer generation into a two-stage causality-inspired process:

$$(F, Q) \rightarrow M \rightarrow K \rightarrow R \rightarrow Y.$$

By explicitly modeling legal fact extraction, key legal fact identification, rule retrieval, and structured answer generation, our method provides a simple yet effective way to improve the interpretability and robustness of legal reasoning for judicial subjective questions.

3.5 IMPLEMENTATION OVERVIEW

We implement the proposed framework as a two-stage pipeline with four internal components. In Stage 1, the model first extracts legal facts from the full case narrative and then identifies key legal facts from the extracted factual layer. In Stage 2, the system retrieves relevant legal rules based on the identified key legal facts and the legal question, and then generates the final structured answer consisting of fact finding, legal analysis, and final answer.

The two stages are trained separately. Stage 1 is supervised by `key_legal_facts`, while Stage 2 is supervised by `reference_answer`. During inference, the model first predicts legal facts and key legal facts, then uses the predicted key legal facts to retrieve relevant legal rules, and finally generates the structured response. More detailed implementation settings, including prompt design, training strategy, and inference procedure, are provided in the appendix.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset. We conduct experiments on the CAIL2025 judicial subjective-question dataset. The dataset contains 84 judicial subjective questions covering multiple legal domains, including criminal law, civil law, administrative law, commercial law, criminal procedure law, civil procedure law, and theoretical law. Each sample contains a case description and a question description. In addition, standard reference answers and detailed scoring rubrics are provided for evaluation under expert guidance.

For the proposed framework, the raw model input consists only of the case narrative F and the legal question Q. Unlike conventional settings that directly provide legal provisions as part of the input, our framework obtains legal rules through a retrieval step based on the identified key legal facts. Following the proposed two-stage framework, the dataset is used for final answer generation under a causality-inspired legal reasoning setting.

Knowledge base and retrieval setting. To support the second-stage retrieval process, we adopt the same legal knowledge base construction setting as the original comparison framework for this task. The knowledge base includes a statutory and regulatory database containing effective laws and judicial interpretations, as well as an empirical knowledge base containing mock questions and historical case materials. During retrieval, we use a hybrid retrieval mechanism combining BM25 and dense embedding retrieval, followed by a cross-encoder reranking step. The top-k is set to 15 in the retrieval stage, and the top-5 reranked legal chunks are retained as the context input to the generator.

Task setting. The primary task is *judicial subjective answer generation*, in which the model must generate a legally reasoned answer for the given case and question. Under our framework, the model first identifies legally decisive facts, then retrieves relevant legal rules, and finally produces a structured legal answer. Our main evaluation focuses on final answer quality under this complete pipeline setting.

Comparison setting. We report two types of results. First, we present the results in the provided comparison table as *reported baselines*, including GPT-4o, Claude-4.5, DeepSeek-V3.1, Doubao-1.6-thinking, KIMI-K2, Qwen3-235B-a22b, and GLM4.6. These numbers are taken as reported results and are used only as external reference. Second, we evaluate our framework on eight models, namely KIMI-K2, Doubao, Qwen, DeepSeek-V3.2, GLM, GPT-5, Gemini-3.1-Flash-Lite, and Claude-Haiku-4.5. This second group constitutes the main in-framework comparison in this paper. For all compared models, we adopt the same two-stage reasoning decomposition. When model accessibility permits, the two stages are implemented with stage-wise supervised fine-tuning; otherwise, they are instantiated through the same staged prompting and inference pipeline.

Pipeline protocol. We evaluate the proposed framework in a stage-wise manner. In Stage 1, the model first identifies legal facts and then further identifies key legal facts from the case narrative and question. In Stage 2, the system retrieves relevant legal rules based on the predicted key legal

Table 1: Reported baseline results on the CAIL2025 dataset. These numbers are taken from the provided comparison table and are used as external reference rather than as a fully controlled in-framework benchmark.

Model	ROUGE			BLEU			Score
	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-N	
GPT-4o	16.50	7.04	15.81	21.12	15.24	8.71	73.27
Claude-4.5	18.60	7.52	17.54	22.68	15.93	8.80	70.34
DeepSeek-V3.1	13.65	4.60	13.30	32.16	23.65	14.24	71.23
Doubao-1.6-thinking	22.96	11.76	22.15	35.14	26.71	17.16	68.47
KIMI-K2	23.87	9.21	22.28	35.03	25.53	15.21	**78.95**
Qwen3-235B-a22b	18.73	7.61	17.61	22.88	16.96	10.26	54.00
GLM4.6	23.34	13.81	22.78	28.09	21.06	12.91	74.87

facts and the legal question, and then generates the final structured answer. During testing, gold key legal facts are not provided. The final performance therefore reflects the joint effectiveness of intermediate fact selection, retrieval grounding, and answer generation.

Evaluation metrics. We adopt the same evaluation protocol as the benchmark setting. For automatic text evaluation, we report ROUGE-1, ROUGE-2, ROUGE-L, BLEU-1, BLEU-2, and BLEU-N. More importantly, we report the task-specific *Score*, which follows the judicial examination grading mode and evaluates whether the generated answer covers the legally required scoring points defined in expert-verified rubrics. Compared with standard text-overlap metrics, this metric better reflects legal accuracy and practical usefulness in judicial subjective examinations.

4.2 Main Results

Reported baseline results. Table 1 presents the reported baseline results from the provided comparison table. Among them, KIMI-K2 achieves the best *Score*, reaching 78.95, followed by GLM4.6 with 74.87 and GPT-4o with 73.27. At the same time, the table also shows that strong ROUGE or BLEU scores do not necessarily translate into the best judicial score. For example, Doubao-1.6-thinking achieves the strongest ROUGE and BLEU values in several columns, yet its *Score* is only 68.47.

In-framework comparison. Table 2 reports the main comparison under our framework. Claude-Haiku-4.5 achieves the highest *Score* of 88.53, followed by DeepSeek-V3.2 with 88.15 and Gemini-3.1-Flash-Lite with 87.35. GPT-5 and KIMI-K2 both reach 84.90, while Qwen also performs competitively with 84.55. These results show that the proposed framework yields strong final-answer performance across multiple backbones.

Text overlap versus legal scoring. A consistent pattern across both tables is that lexical-overlap metrics and judicial scoring do not move in parallel. Some models achieve stronger ROUGE or BLEU scores without obtaining the best *Score*, while other models achieve the best *Score* without dominating the overlap-based metrics. This suggests that lexical similarity is not a sufficient proxy for legal reasoning quality in judicial subjective questions.

Interpretation of the comparison. Taken together, the two tables support the necessity of evaluating legal reasoning from both text similarity and judicial scoring perspectives. ROUGE and BLEU are useful for measuring surface-level overlap, whereas *Score* is more closely aligned with the substantive grading logic of judicial subjective examinations. Since the reported baseline table is not re-run under our controlled pipeline, we do not interpret cross-table differences as a strict head-to-head causal improvement. The main controlled comparison in this paper is the in-framework comparison together with the ablation study below.

Table 2: Main comparison of models evaluated under our framework on the CAIL2025 dataset.

Model	ROUGE			BLEU			Score
	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-N	
KIMI-K2	16.29	8.05	15.35	21.95	15.91	9.37	84.90
Doubao	17.69	10.27	16.62	21.39	16.10	10.13	82.10
Qwen	14.63	7.04	14.22	14.17	10.69	6.67	84.55
DeepSeek-V3.2	15.94	7.78	15.29	17.79	13.19	7.99	88.15
GLM	18.10	9.35	17.09	19.03	14.06	8.39	80.64
GPT-5	19.30	8.16	18.92	14.39	10.62	6.23	84.90
Gemini-3.1-Flash-Lite	20.79	12.07	19.86	16.33	12.21	7.39	87.35
Claude-Haiku-4.5	13.53	6.42	13.29	12.73	9.33	5.49	**88.53**

Table 3: Category-wise scoring performance (%) of models evaluated under our framework on the CAIL2025 dataset across seven legal domains. A missing entry indicates that the model did not produce a scorable output for that domain under the current evaluation setting.

Model	Civil law	Criminal law	Administrative law	Commercial law	Criminal-procedure law	Civil procedure law	Theoretical law
KIMI-K2	86.00	75.83	94.08	78.05	75.51	94.64	94.74
Doubao	92.86	50.00	83.70	96.94	–	93.55	94.74
Qwen	80.00	70.00	90.13	82.52	93.88	94.64	90.35
DeepSeek-V3.2	76.00	91.25	86.18	90.24	87.35	86.61	87.72
GLM	88.00	67.50	84.87	74.80	93.88	88.39	86.84
GPT-5	84.00	90.42	83.55	80.08	82.65	92.86	82.46
Gemini-3.1-Flash-Lite	88.00	88.33	80.00	84.15	90.82	83.93	94.74
Claude-Haiku-4.5	84.00	95.42	86.32	79.92	86.73	86.61	94.74

4.3 Category-wise Analysis

Performance across legal domains. Table 3 reports category-level *Score* across seven legal domains for the models evaluated under our framework, namely civil law, criminal law, administrative law, commercial law, criminal procedure law, civil procedure law, and theoretical law. The results show substantial variation across domains. Claude-Haiku-4.5 achieves the best result in criminal law, KIMI-K2 leads in administrative law, civil procedure law, and theoretical law, and Doubao performs particularly strongly in civil law and commercial law. Qwen and GLM reach the joint best result in criminal procedure law, while Gemini-3.1-Flash-Lite also performs strongly across several domains and reaches 94.74 in theoretical law. A missing entry indicates that the model did not produce a scorable output for that domain under the current evaluation setting. Overall, these findings suggest that judicial subjective reasoning performance is not uniform across legal domains, and that domain-level analysis provides a more fine-grained view than overall averages alone.

4.4 Ablation Study

Overall ablation results. Table 4 reports the ablation results of the proposed framework with DeepSeek-V3.2 as the backbone. The full model achieves the best *Score* of 88.15. Removing RAG reduces *Score* to 82.88, a drop of 5.27 points. Removing key legal fact identification causes a larger drop to 76.82, corresponding to an 11.33-point decline. When both key legal fact identification and RAG are removed, *Score* further decreases to 73.27. These results show that both modules contribute to final judicial scoring, while key legal fact identification plays the more critical role in the current setting.

Ablation interpretation. An additional observation is that some ablated variants obtain slightly higher ROUGE or BLEU values than the full model. For example, the *w/o Key Fact* setting yields higher ROUGE and BLEU scores than the full model, yet its *Score* is much lower. This again indicates that overlap-based metrics do not fully capture the legal validity of generated answers. The main gain of the full framework lies not in maximizing surface-level similarity, but in improving whether the model identifies legally decisive facts, grounds the answer in the appropriate legal basis, and covers the required scoring points in the final response.

Table 4: Ablation study of the proposed framework with DeepSeek-V3.2 as the backbone.

Setting	R-1	R-2	R-L	B-1	B-2	B-N	Score
DeepSeek-V3.2	15.94	7.78	15.29	17.79	13.19	7.99	**88.15**
w/o RAG	16.02	8.36	15.32	18.36	13.55	8.11	82.88
w/o Key Fact	17.00	8.67	16.34	18.67	13.93	8.62	76.82
w/o Key Fact & RAG	14.39	7.45	13.77	18.60	13.80	8.26	73.27

4.5 Discussion

The experiments reveal a clear pattern: text-overlap metrics and judicial scoring metrics do not move in parallel. This is visible in both the external reference table and the in-framework comparison, and it is further confirmed by the ablation study. Some models or variants achieve higher ROUGE or BLEU values without obtaining the best *Score*. Conversely, the strongest models under our framework, especially Claude-Haiku-4.5 and DeepSeek-V3.2, achieve the highest judicial scores without dominating lexical-overlap metrics. This suggests that, for judicial subjective questions, legal reasoning quality cannot be adequately characterized by surface-level similarity alone.

From a methodological perspective, the ablation results provide direct support for the design of our framework. The larger performance drop caused by removing key legal fact identification indicates that fact selection is central to judicial subjective reasoning. RAG also contributes positively, but its role appears to be more supportive than foundational in the current setting. In other words, the core challenge is not merely to generate fluent legal text, but to identify the facts that should serve as the basis of legal assessment, connect them to the appropriate legal rules, and organize them into a legally supportable answer.

The category-wise results further reinforce this interpretation. No single model dominates all categories, and performance varies substantially across question types. This suggests that judicial subjective reasoning is not a homogeneous task. Different categories place different demands on fact selection, issue identification, legal rule grounding, and answer organization.

5 Conclusion

This paper studies a central problem in legal AI. For judicial subjective questions, the difficulty lies not only in generating fluent answers, but in determining which facts should ground legal assessment and norm application. To address this problem, we introduce key legal fact identification as an explicit intermediate layer and reformulate legal reasoning as a two-stage process. The model first identifies legally decisive facts and then grounds its answer in those facts together with retrieved legal rules. This design creates a clearer path from case narrative to legal conclusion.

The contribution of this framework is not to claim a full causal model of legal inference. Its contribution is to show that a control-variable perspective on fact selection can serve as a useful organizing principle for legal reasoning. In this sense, the paper speaks to a broader shift in legal NLP and LLM research. The central question is no longer only whether models can produce legally fluent text. It is whether their reasoning can be organized around the right facts and the right legal basis. Our results suggest that progress in legal AI depends not only on stronger models or larger datasets, but also on intermediate reasoning structures that better connect facts, rules, and conclusions.

This work still has clear limitations. The current evaluation mainly focuses on final-answer performance and does not yet separately measure the quality of intermediate key legal fact identification or downstream legal retrieval. The framework also depends on the quality of key-fact annotation and prediction, and its generalizability to broader legal tasks remains to be tested. Future work may develop richer intermediate representations and combine fact-centered reasoning with more formal models of statutory interpretation and multi-step judicial analysis. More broadly, the long-term goal is legal AI that is not only accurate, but also normatively grounded, fact-sensitive, and institutionally reliable.

REFERENCES

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4310–4330, 2022.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, et al. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7933–7962, 2024.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.

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

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. Large language models and causal inference in collaboration: A comprehensive survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7668–7684, 2025.

Jing Ma. Causal inference with large language model: A survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5886–5898, 2025.

Judea Pearl. Causal inference. *Causality: objectives and assessment*, pp. 39–58, 2010.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. Causality for large language models. *arXiv preprint arXiv:2410.15319*, 2024.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*, 2018.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3540–3549, 2018.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 9701–9708, 2020.

A Detailed Implementation

Task decomposition. We implement the proposed framework as a two-stage pipeline with four internal components. In Stage 1, the model first extracts legal facts from the full case narrative and then identifies the subset of key legal facts that are most legally decisive. In Stage 2, the system retrieves relevant legal rules based on the identified key legal facts and the legal question, and finally generates a structured legal response. Formally, for each sample, the process can be written as

$$M = h_\alpha(F, Q), \qquad K = f_\theta(M, Q),$$

$$R = r_\eta(K, Q), \qquad \hat{Y} = g_\phi(K, Q, R).$$

Here, F denotes the full case narrative, Q denotes the legal question, M denotes the extracted legal facts, K denotes the identified key legal facts, R denotes the retrieved legal rules, and $\hat{Y}$ denotes the final structured output. This design ensures that the final answer is explicitly grounded in intermediate fact selection and subsequent legal retrieval.

Input format. For each case, the raw input consists of two components: the case narrative and the legal question. In Stage 1, the model is prompted to read the full factual narrative, first extract legal facts, and then identify the key legal facts that are most decisive for answering the question. In Stage 2, the system uses the identified key legal facts together with the legal question to retrieve relevant legal rules through a retrieval-augmented generation procedure, and then asks the model to produce the final structured response.

Output format. The final output is organized into three parts: `fact_finding`, `legal_analysis`, and `final_answer`. The first part summarizes and identifies the key facts of the case, the second part explains the legal analysis based on the key facts and the retrieved legal rules, and the third part provides the final conclusion-oriented response to the judicial subjective question.

Supervision signals. The training process uses two levels of supervision. The first is intermediate supervision for key legal fact identification, based on the annotated field `key_legal_facts`. The second is final-answer supervision, based on the field `reference_answer`. In this way, the model is encouraged not only to generate correct answers, but also to rely on legally decisive facts during reasoning. In the current setting, the retrieval step itself is not treated as a standalone supervised target, but as an intermediate support module for answer generation.

Model setting. We adopt an instruction-following large language model as the backbone. The same backbone model is used across the pipeline for consistency. In practice, the four components are implemented through prompt-based decomposition under a unified two-stage design. This setting allows us to separately evaluate the behavior of key legal fact identification and final answer generation while keeping the overall framework coherent.

Prompt design. To align the model with the proposed causality-inspired reasoning intuition, prompts in Stage 1 explicitly instruct the model to first extract legal facts from the raw narrative and then identify those facts that materially affect the legal conclusion. The prompts emphasize that background details, narrative embellishments, and weakly related descriptions should be excluded unless they substantially influence legal judgment. In Stage 2, the system first transforms the identified key legal facts and the legal question into a retrieval-oriented query for legal search, and then instructs the model to generate the final response based primarily on the identified key legal facts and the retrieved legal rules, rather than on the full original case description.

Training strategy. We employ supervised fine-tuning in a stage-wise manner. First, the model is trained on Stage 1 using `case_text` and `question` as input and `key_legal_facts` as target output. Then, the answer generation model in Stage 2 is trained using `key_legal_facts`, `question`, and retrieved legal rules as input and `reference_answer` as target output. This design directly reflects the decomposition

$$(F, Q) \rightarrow M \rightarrow K \rightarrow R \rightarrow Y.$$

Inference procedure. During inference, the model first predicts legal facts and key legal facts for each case. These predicted key legal facts are then used to form retrieval-oriented legal queries and retrieve relevant legal rules, which are further fed into the second-stage answer generator to produce the final structured response. We do not provide gold key legal facts at test time. Therefore, the final performance reflects the combined effectiveness of legal fact extraction, key legal fact identification, legal retrieval, and answer generation.

Implementation objective. The proposed implementation is intended to test the central hypothesis of this paper: legal reasoning performance can be improved when the model is explicitly guided to identify legally decisive facts before legal retrieval and answer generation. By separating the reasoning process into two stages and four internal components, we make it possible to evaluate whether performance gains come from better identification of pivotal facts and better grounding of legal rules, rather than from stronger direct memorization of surface-level textual associations.

B PROMPTS

This appendix presents the prompts used in our two-stage legal reasoning framework. The framework contains four prompts: (1) legal fact extraction, (2) key legal fact identification, (3) RAG query generation, and (4) answer generation with key legal facts and retrieved legal rules.

B.1 PROMPT 1: LEGAL FACT EXTRACTION

The goal of the first prompt is not to directly answer the legal question, but to extract legally relevant facts from the case materials while filtering out purely narrative, emotional, or stylistic expressions.

Prompt 1: Legal Fact Extraction

You are a rigorous legal reasoning assistant. Your task is not to directly answer the question, but to first extract "legal facts" from the case materials.

[Task Requirements]
Please identify and extract factual content that may be relevant to legal analysis based on the given case materials and legal question.
Here, "legal facts" refer to:

1. facts related to the parties, acts, time, place, objects, consequences, procedures, and liabilities of the case;
2. facts that may affect the determination of legal relationships, liability, constituent elements, or the establishment of defenses;
3. facts that can provide a basis for subsequent legal application and legal judgment.

Please note:

1. Do not directly output the final answer;
2. Do not introduce new facts that do not appear in the materials;
3. Do not output purely evaluative language;
4. Keep the factual statements concise, objective, and complete as much as possible;
5. If multiple facts can be separated, list them point by point.

[Input Information]

Case materials:
{case_text}

Legal question:
{question}

[Output Requirement]
Please strictly output in the following JSON format and do not add any other content:

```
{ "legal_facts": [
   "...",
   "...",
   "..."
]
}
```

B.2 Prompt 2: Key Legal Fact Identification

This prompt is the core step of our framework. Its purpose is to identify, from the extracted legal facts, those facts that have a decisive impact on legal evaluation, rule application, or the final conclusion.

Prompt 2: Key Legal Fact Identification

You are a rigorous legal reasoning assistant. Your task is to further identify "key legal facts" from the already extracted "legal facts."

[Task Requirements]
Please identify the key legal facts that have a decisive influence on the final legal judgment based on the given legal facts and legal question.
"Key legal facts" refer to:

1. facts that have a substantial impact on the determination of legal relationships, liability, constituent elements, legal application, or the final conclusion;
2. facts such that if they were changed, removed, or replaced, the legal evaluation, applicable rules, or final conclusion of the case might change accordingly;
3. by contrast, facts that merely serve as background introduction, auxiliary explanation, or narrative detail and do not substantially affect the legal conclusion should not be treated as key legal facts.

Please note:

1. You need to select from the given "legal facts," rather than rewrite the case materials from scratch;
2. Do not classify all facts as key facts;
3. The number of key legal facts should be moderate and sufficient to support subsequent legal judgment;
4. Do not directly output the final answer;
5. Do not introduce new facts that do not appear in the materials.

[Input Information]

Legal facts:
{legal_facts}
Legal question:
{question}

[Output Requirement]
Please strictly output in the following JSON format and do not add any other content:

```
{ "key_legal_facts": [
   "...",
   "...",
   "..."
],
"reasoning": [
   "...",
   "...",
   "..."
]
}
```

B.3 Prompt 3: RAG Query Generation

This prompt serves as the retrieval-oriented transition between key legal fact identification and legal rule grounding. Its purpose is to transform the case materials, legal question, and identified key legal facts into a concise legal retrieval query that can be used for RAG-based legal search.

Prompt 3: RAG Query Generation

You are a rigorous legal reasoning assistant. Your task is to generate a legal retrieval query for RAG-based legal search based on the case materials, legal question, and the already identified key legal facts.

[Task Requirements]
Please generate a concise and retrieval-oriented legal query that can be used to search for the most relevant legal rules, statutory provisions, judicial interpretations, or legal principles.
The generated query should:

1. focus on the legally decisive facts rather than all background details;
2. reflect the core legal issue raised by the question;
3. contain legal concepts, liability issues, or constitutive elements that are useful for retrieving relevant legal rules;
4. avoid copying the full case narrative verbatim;
5. be concise, precise, and suitable for legal retrieval.

[Input Information]

Case materials:
{case_text}
Legal question:
{question}
Key legal facts:
{key_legal_facts}

[Output Requirement]
Please strictly output in the following JSON format and do not add any other content:

```
{ "retrieval_query": "..." }
```

B.4 Prompt 4: Answer Generation with Key Legal Facts

This is the final answer-generation stage and also the main point where our method differs from the baseline systems. While baseline methods typically generate answers directly from the case materials and question, our method additionally provides the identified key legal facts and the retrieved legal rules, so that the model can reason around the legally decisive facts with explicit legal grounding.

Prompt 4: Answer Generation with Key Legal Facts

You are a rigorous legal reasoning assistant. Please generate a legally grounded and clearly reasoned answer based on the case materials, legal question, the already identified key legal facts, and the retrieved legal rules.

[Task Requirements]
You need to conduct legal analysis based on the "key legal facts" and the retrieved legal rules, and provide a well-supported answer to the question.
Please follow these principles:

1. The answer should primarily focus on the key legal facts, rather than evenly relying on all background information;
2. The answer should reflect a reasoning process from facts to rules to conclusion;
3. The answer should combine legal rules with analysis, rather than giving only a conclusion;
4. Do not fabricate facts that are not grounded in the case materials;
5. Use language that is as clear, concise, and normatively appropriate as possible, in line with the style of judicial subjective questions.

[Input Information]

Case materials:
{case_text}
Legal question:
{question}
Key legal facts:
{key_legal_facts}
Retrieved legal rules:
{retrieved_rules}

[Output Field Notes]
`fact_finding`: identification and summary of the key facts in the case
`legal_analysis`: legal application and analysis based on the key facts and retrieved legal rules
`final_answer`: the final conclusion-oriented response to the judicial subjective question

[Output Requirement]
Please strictly output in the following JSON format and do not add any other content:

```
{ "fact_finding": "...",
"legal_analysis": "...",
"final_answer": "..." }
```

C DATASET EXAMPLE

To illustrate the structure of our dataset and the role of intermediate annotations, we present a representative example below. The example shows how a raw judicial subjective-question sample is transformed into legal facts, key legal facts, and a reference answer.

Table 5: Illustrative example from the judicial subjective-question dataset

Case Description

A and B got into an argument over a trivial matter. A struck B on the head with a wooden stick, causing minor injury. After the incident, A voluntarily called the police and truthfully confessed the main facts after arriving at the police station. After treatment, B filed an incidental civil compensation claim.

Question

Please analyze the legal nature of A's conduct and explain A's criminal liability and incidental civil liability.

Legal Facts

1. A and B had an argument over a trivial matter;
2. A struck B with a wooden stick;
3. the injured body part was B's head;
4. B suffered minor injury;
5. A voluntarily called the police after the incident;
6. A truthfully confessed the main facts;
7. B filed an incidental civil compensation claim.

Key Legal Facts

1. A actively committed the assault;
2. the assaulted body part was B's head, an important body area;
3. the conduct caused minor injury to B;
4. A voluntarily reported to the police and truthfully confessed after the incident;
5. B filed an incidental civil compensation claim.

Reference Answer

A intentionally struck B on the head and caused minor injury, which satisfies the constitutive elements of intentional injury and therefore entails corresponding criminal liability. Since A voluntarily reported to the police and truthfully confessed after the incident, A may be recognized as having surrendered and may receive lenient treatment according to law. As B filed an incidental civil compensation claim, A should also bear the corresponding civil compensation liability.

D ANNOTATION PROTOCOL FOR KEY LEGAL FACTS

To ensure the quality and consistency of key legal fact annotation, we adopt an expert-guided annotation protocol. The protocol is designed to identify the facts that materially influence legal assessment, rule application, and final conclusions, while filtering out background or weakly relevant details.

Table 6: Expert-guided annotation protocol for key legal facts.

Step	Stage	Description
Step 1	Raw sample review	Annotators read the original case description and question to identify the core legal issue to be resolved, such as legal qualification, liability determination, or the basis of judgment.
Step 2	Legal fact extraction	Annotators extract factual statements that may be relevant to legal analysis from the original narrative, while excluding purely emotional, decorative, or weakly related details.
Step 3	Key fact screening	From the extracted legal facts, annotators further identify the subset of facts that are legally decisive. A fact is treated as a key legal fact if changing, removing, or replacing it would likely alter the legal assessment, rule application, or final conclusion, while other core conditions remain approximately fixed.
Step 4	Answer alignment	Annotators check whether the selected key legal facts are sufficient to support the reasoning path of the reference answer, including fact finding, legal analysis, and final conclusion.
Step 5	Expert verification	The annotated key legal facts are reviewed under expert guidance to ensure consistency with legal doctrine, judicial examination standards, and the scoring logic of the reference answer.