2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

An LLM-empowered General Workflow for Legal Case Analysis: A Case Study on Elderly Laborer Protection

Yuting Wang[1*], Runliang Niu[1*], Xingyuan Min[1],

Nanfei Gu[2(✉)], Qianli Xing[1(✉)], Qi Wang[1(✉)]

[1] Jilin University, Jilin 130012, China
[2] Shanghai Jiao Tong University, Shanghai 200030, China
{wangyt5521, niurl19}@mails.jlu.edu.cn
{qiwang, qianlixing}@jlu.edu.cn

Abstract. The emergence of LLMs has revolutionized the field of legal case analysis. Existing research primarily focuses on specific issues, e.g., legal view generation and case retrieval, neglecting the universal ability of legal semantic features within texts to address multiple issues. In this paper, we develop a novel general workflow utilizing LLMs for diverse legal case analysis tasks, uncovering implicit information in the legal case datasets. Specifically, the workflow involves three stages: (1) legal experts first define a fine-grained elements framework for legal cases; (2) LLMs then extract these elements from documents and convert them into structured tables, aiming to capture special meaningful information contained in the document; (3) various questions of interest to legal experts can be addressed by selecting and analyzing relevant elements. Benefiting from LLMs' knowledge and ability in understanding and processing text, element annotation becomes scalable, allowing our workflow to handle general legal intelligence tasks. We validate the feasibility and effectiveness of our workflow on the Elderly Laborer Protection issue as a case study, exploring the factors affecting judgment outcomes from a causal perspective. Our fine-grained legal case dataset at the document level annotated by legal experts and easy-to-use workflow tools are available at SmartFive/LegalCaseAnalysis.

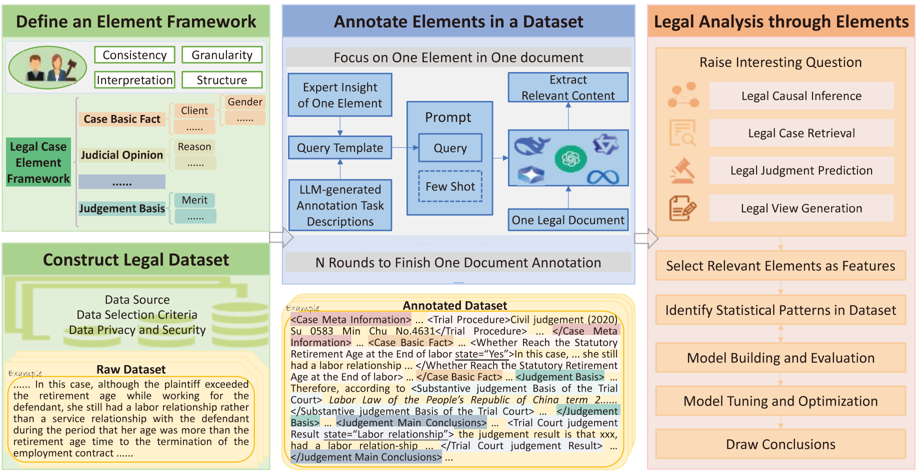
Keywords: Large Language Models, Legal Case Analysis, General Workflow.

1 Introduction

With the advancement of natural language processing (NLP) technologies, many challenges in legal intelligence have been effectively addressed. The study [1] has demonstrated that legal case elements, the specific textual components within a legal case, contain all the relevant information necessary to understand and analyze the case holistically. As a result, annotating legal cases with elements can improve the performance of legal intelligence tasks. Previous research [2] has used traditional NLP techniques to annotate legal case elements, which is not only time-consuming and costly but also suffer from coarse data annotation granularity. With the development of large language

*Equal Contribution.

models (LLMs), recent work has used LLMs to assist in the discovery [3] and annotation [4] of common facts that can support or oppose specific legal arguments. However, current methods are tailored to extract elements relevant to predefined legal questions. While effective for their intended scope, these methods lack generalizability and cannot support broader legal analysis tasks.


Define an Element Framework
Consistency
Granularity
Interpretation
Structure
Legal Case Element Framework
Case Basic Fact
Client
Gender
......
Judicial Opinion
Reason
......
Judgement Basis
Merit
Construct Legal Dataset
Data Source
Data Selection Criteria
Data Privacy and Security
Example
Raw Dataset
Annotate Elements in a Dataset
Focus on One Element in One document
Expert Insight of One Element
Query Template
LLM-generated Annotation Task Descriptions
Prompt
Query
Few Shot
Extract Relevant Content
One Legal Document
N Rounds to Finish One Document Annotation
Annotated Dataset
Legal Analysis through Elements
Raise Interesting Question
Legal Causal Inference
Legal Case Retrieval
Legal Judgment Prediction
Legal View Generation
Select Relevant Elements as Features
Identify Statistical Patterns in Dataset
Model Building and Evaluation
Model Tuning and Optimization
Draw Conclusions

...... In this case, although the plaintiff exceeded the retirement age while working for the defendant, she still had a labor relationship rather than a service relationship with the defendant during the period that her age was more than the retirement age time to the termination of the employment contract ......

<Case Meta Information> ... <Trial Procedure>Civil judgement (2020) Su 0583 Min Chu No.4631</Trial Procedure> ... </Case Meta Information> ... <Case Basic Fact> ... <Whether Reach the Statutory Retirement Age at the End of labor state="Yes">In this case, ... she still had a labor relationship ... </Whether Reach the Statutory Retirement Age at the End of labor> ... </Case Basic Fact> ... <Judgement Basis> ... Therefore, according to <Substantive judgement Basis of the Trial Court> *Labor Law of the People's Republic of China term 2......* </Substantive judgement Basis of the Trial Court> ... </Judgement Basis> ... <Judgement Main Conclusions> ... <Trial Court judgement Result state="Labor relationship"> the judgement result is that xxx, had a labor relation-ship ... </Trial Court judgement Result> ... </Judgement Main Conclusions> ...

Fig. 1. The overview of our workflow for legal case analysis using legal element annotation.

In contrast to these task-specific approaches, we propose a general and scalable workflow for legal case analysis. Our approach centers on structuring and identifying essential legal elements, enabling a unified representation of legal case content across domains. As illustrated in **Fig. 1**, our workflow includes three stages: framework definition and dataset construction, element annotation using LLMs, and downstream legal analysis. In the first stage, legal experts define a hierarchical framework for the annotation of legal cases, considering factors including the consistency and interpretation of elements, and the granularity and structure of the framework. The framework should encompass both universally applicable structural elements (e.g., case basic facts, judicial opinion, and judgment basis) and fine-grained evidence elements (e.g., client gender). Concurrently, high-quality case data are collected with careful attention to source, selection criteria, privacy and security. In the second stage, LLMs are guided to annotate these elements. The workflow leverages LLMs to extract mentions of these elements from case documents and classify them into predefined categories. To enhance model performance, the prompts will integrate insights from legal experts on elements and task descriptions generated by LLMs, and few-shot learning can be added to improve model performance. Each element in every document is annotated individually, and the fully annotated documents form the dataset. In the third stage, the annotated elements support empirical legal analysis. Selecting relevant elements enables data mining algorithms to address various interesting questions and uncover overall patterns in the dataset, thereby supporting empirical legal research from a data-driven perspective.

We validate this workflow to the "Elderly Laborers Legal Protection" problem as a case study. To investigate the usability and accuracy of using LLMs to extract elements from legal case documents, we first collaborate with legal experts to annotate a **Doc**ument-level Fine-grained Legal Case Dataset for **O**ld-age **L**aborer Issues (OLDoc). Then we study how legal experts leverage domain knowledge to drive LLMs for empirical research on a scalable paradigm, supporting mining valuable legal conclusions, such as "What factors play the most critical role in the judgment outcome?", "How does the law applicable standard affect the judgment outcome?", and so on. **Our major contributions are as follows**:

- We collaborate with legal experts to perform fine-grained annotation of a legal case dataset at the document level. This ensures an accurate representation of the original text, and elements can be selected as feature for specific legal intelligence tasks.
- We introduce a general workflow for legal case analysis based on legal case elements, which addresses the limitations of previous problem-specific methods and establishes a unified analytical framework applicable to diverse legal analysis tasks.
- We apply this workflow to the domain of elderly labor protection, defining metrics to evaluate the performance of several LLMs in annotating textual elements and investigating the key factors that influence legal judgments from causal aspect.

2 Related Work

Legal Datasets Construction and Annotation. Both AI researchers and legal experts are interested in building datasets in the legal field [5,6,7,8]. Most of the available legal datasets are developed mainly in English [9,10], and there are also some datasets in Chinese [11], German [12], and French [13]. Besides, some datasets contain multi-lingual data [14,15,16]. Several studies annotate datasets to address specific legal issues, such as contract review [17], statutory reasoning in tax law [18], tort judgements in Japan [19], and legal disputes over domain names [20], with similar work conducted in jurisdictions like Greece [21] and India [22]. But these efforts only contain metadata data, which fails to provide deeper labeling information. Thus, they cannot be directly applied to general judicial practice and fine-grained legal datasets are still very scarce.

LLMs in Legal Intelligence. With the development of LLMs, their application in the legal domain has grown, handling increasingly complex tasks. Methods like Event Grounded Generation [23] incorporate event information into criminal court views, while systems like AgentCourt [24] train lawyer agents through courtroom simulations. LawLuo [25] uses multiple LLMs for legal consultations, and research [26] applies LLMs to isolate summary judgments from UK court decisions. Other studies explore LLMs in legal advice, such as [27] on expertise concerns and [28] on intangible harms like loss of agency. However, existing research primarily focuses on specific issues, neglecting the broader potential of legal semantic features for various tasks.

Legal Analysis Workflow. Recent advancements in legal analysis workflows focus on enhancing integration, automation, and efficiency through semantic technologies. Key studies include a federated data platform for managing evidence and inquiries [29], semantic modeling for ensuring compliance in multi-jurisdictional investigations [30], and a semantic methodology for improving trial documentation and case traceability [31]. The development of Workflow Privacy Patterns (WPPs) for compliant workflows has been driven by GDPR [32]. Efforts to model China's civil litigation workflows [33] focus on data security and legal compliance. These studies demonstrate the potential of workflow automation in specific legal tasks, but they focus on individual issues and have not yet proposed a general workflow for legal analysis.

3 Methodologies

Inspired by the process of data mining, we propose a general workflow for legal intelligence tasks. The key to applicability across tasks lies in defining a hierarchical legal case element framework. This is analogous to feature definition. LLMs can annotate the legal case dataset with fine-grained elements, and modifications to the framework can be easily implemented using new prompts, ensuring flexibility and scalability in handling large volumes of data annotation. This step resembles feature extraction and normalization. Then traditional data mining algorithms can be applied to the selected features for tasks like legal causal inference.

We apply the proposed workflow to analyze legal cases involving Chinese elderly laborers, using 1,867 cases selected from the China Judgements Online dataset[1]. The average document length is 3,406 tokens, with a maximum of 13,060 tokens, showing significant variance in document length distribution. As shown in **Fig. 2**, the average length increases from the first trial to retrial instance, posing challenges for fine-grained analysis.

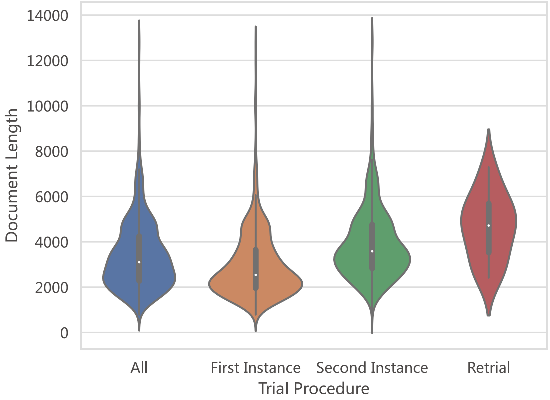


Fig. 2. Document Length Distribution.

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

Our dataset is specifically targeted at Chinese elderly laborer legal cases and has been annotated at a fine-grained level, providing a rich, domain-specific resource to support and advance future research in this area. Currently, the dataset is limited to Chinese labor dispute cases. In the future, we aim to expand the dataset to include cases from diverse legal systems and jurisdictions, enabling more comprehensive analysis and enhancing the framework's generalization capabilities to ensure broader applicability across various legal contexts.

Table 1. The structure element *case basic fact* part of Legal Element Framework for OLDoc. These elements marked by symbol √ have attribute values.

B. Case Basic Fact 案件基本事实部分			
B.5	Labor gender 劳动者性别 √	B.6	Time of birth 出生时间
B.7	Start time of labor 劳动起始时间	B.8	End time of labor 劳动结束时间
B.9	Start age of labor 开始劳动的年龄	B.10	End age of labor 结束劳动的年龄
B.11	Whether reach the statutory retirement age at the end of labor 结束劳动时是否达到法定退休年龄 √		
B.12	When do labors to reach the mandatory age for retirement 劳动者何时达到法定退休年龄 √		
B.13	Whether have a written contract 有无书面合同 √		
B.14	Reasons for refusal of arbitration by Arbitration Commission 仲裁委拒绝仲裁的理由		
B.15	Whether enjoy the benefits of the old-age insurance 有无享受养老保险待遇 √		
B.16	Kind of old-age insurance 养老保险待遇类型 √		
B.17	Review the reasons of the court of first instance 回顾一审法院的认定理由		
B.18	Review the court identification basis of first instance 回顾一审法院的认定依据		
B.19	Review the judgement basis of the court of first instance 回顾一审法院的裁判依据		
B.20	Review the judgement results of the court of first instance 回顾一审法院的裁判结果 √		
B.21	Review the reasons of the court of second instance 回顾二审法院的认定理由		
B.22	Review the court identification basis of second instance 回顾二审法院的认定依据		
B.23	Review the judgement basis of the court of second instance 回顾二审法院的裁判依据		
B.24	Review the judgement results of the court of second instance 回顾二审法院的裁判结果 √		

Element Framework Definition. Based on *Rules for the Making of Civil Judgement Documents of People's Courts* by the Supreme People's Court, legal experts define an element framework for annotating the text. It includes six structural elements (case meta information, case basic facts, judicial opinion, judgment basis, judgment main conclusions, and the end part of the judgment) to cover long paragraphs and 32 evidence elements to extract detailed phrases and sentences from the text. The element representation rule is as follows: structural elements are denoted by uppercase letters, while evidence elements are represented by numbers (e.g., E.30 refers to evidence element 30 under structural element E). Some elements and related information are shown in **Table 1**.

Table 2. Text span length statistics in OLDoc

Elements Tokens	Structure elements	Evidence elements	All elements
Answer #	10793	31514	42307
Max answer	4195.4	883.7	1728.6
Avg answer	516.3	54.5	172.3

Legal experts also manually annotate text spans in documents where predefined elements are mentioned, along with their corresponding attribute values, creating the OLDoc dataset. The text span length statistics for each element are shown in **Table 2.** Text span length statistics in OLDoc , which shows that structural elements correspond to much longer spans than evidence elements. **Fig. 3** shows the distribution of elements with attribute values.

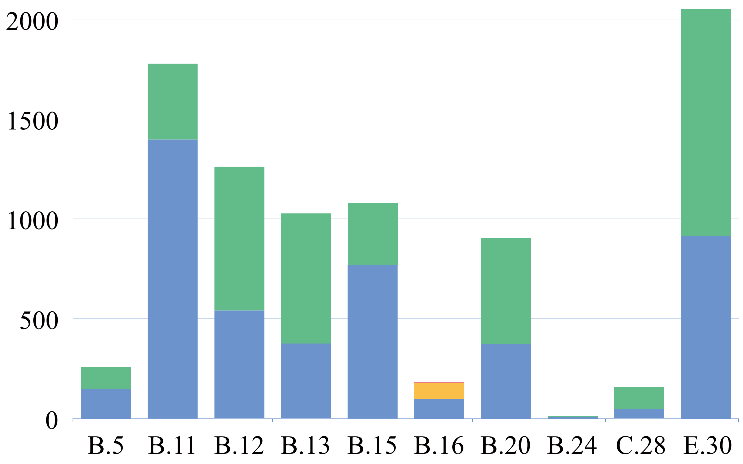


Fig. 3. Element number and attribute distribution, colored by attribute value.

Legal Case Element Annotation. We leverage LLMs to identify sentences within the documents that mention these legal elements. For legal elements with attribute values, LLMs need to classify them into predefined categories. To enhance LLM performance in the legal domain, we design element-specific prompts, combining legal expert insights with LLM-generated task descriptions, and conduct few-shot experiments.

Causal Analysis on Legal Case Elements. We chose to explore causal relationships because applying the conclusions drawn from causal analysis to other tasks is more interpretable and transparent to legal experts. To investigate factors influencing judgment outcomes and derive reasonable judgment paths, we conduct a quantitative causal estimation between judgment outcomes and related elements. The causal conclusions, derived from statistical analysis of the annotated dataset, are validated by legal experts, thereby confirming the effectiveness of the annotation process.

4 Experiment

4.1 LLMs for Legal Element Annotation

We conduct experiments with several LLMs, including GPT-4o-mini, DeepSeek-Chat, Qwen2.5-32b-AWQ, Llama3.1-8B-Chinese-Chat, GLM-4-Plus, and GLM-4-Flash.

The annotation task is divided into two parts: (1) extracting relevant text spans for elements without predefined attribute values, and (2) classifying elements with predefined attribute values into corresponding categories.

Prompt Engineering. Initially, we instructed LLMs to annotate all legal elements in a single-turn dialogue. However, due to excessive input length and element dispersion, LLMs often summarized answers and may omit elements. To address this, we switched to multi-turn dialogues, annotating one element per turn. Each element's prompt is refined using LLM-generated task descriptions and legal expert explanations, with LLM-generated explanations as a fallback. As shown in **Fig. 4**, expert knowledge and task descriptions help improve the performance of LLMs.

We also conducted few-shot experiments. Due to the high token consumption and time required to input the entire text for each dialogue, we limited the experiments to 1 and 2 shot settings. To ensure consistency, we apply the same samples and prompts across all models. The selected samples are typical legal cases rich in elements, providing comprehensive textual features for models.

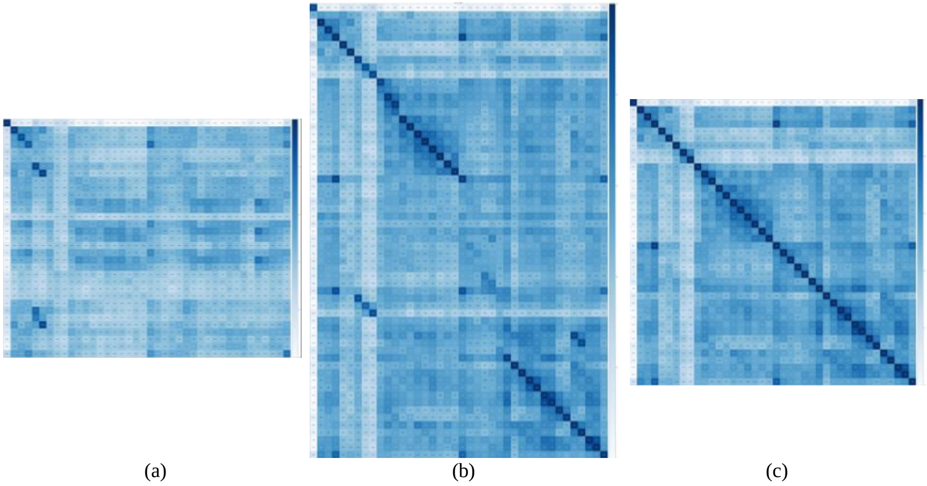


Fig. 4. The similarity matrix compares extraction results from different prompts for the same element in the same document, with reference answers shown horizontally and extraction results vertically. (a) represents prompts without expert knowledge explaining the element, (b) lacks accurate task descriptions, and (c) features prompts incorporating specialized legal knowledge and precise task descriptions.

Elements Extraction from Text. To accurately extract content related to a specific element, LLMs are instructed to avoid summarizing, rephrasing, or inferring during answer generation. However, long text inputs reduce sensitivity to these instructions, leading to occasional summaries or inferences. This situation still extracts the relevant textual information corresponding to the elements, but it introduces challenges for evaluating performance.

Traditional metrics based on lexical overlap tend to underestimate the extraction performance of generative models. To address this, we use BERTScore, which measures textual similarity using semantic embeddings from a pre-trained language model (we use bert-base-chinese). Unlike n-gram-based metrics like BLEU and ROUGE, BERTScore focuses on semantic alignment, offering a more accurate performance evaluation. We evaluate the extraction performance across 0-shot, 1-shot, and 2-shot settings. The results are presented in **Table 3**, which can serve as a benchmark for LLMs on this legal task.

Table 3. Extraction performances comparison across various models and shot numbers. The highest score is highlighted in **bold**, while the lowest score is underlined.

Score / Condition	ROUGE			BLEU			BERTScore
	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-N	
Llama-0shot	0.3223	0.3047	0.3196	0.2765	0.2705	0.2562	0.6077
Llama-1shot	0.6075	0.5768	0.6011	0.5465	0.5360	0.5033	0.8143
Llama-2shot	0.7076	0.6796	0.7011	0.6531	0.6428	0.6090	0.8716
Qwen-0shot	0.6729	0.6489	0.6698	0.5926	0.5867	0.5577	0.8647
Qwen-1shot	0.7355	0.7143	0.7314	0.6669	0.6608	0.6314	0.8983
Qwen-2shot	0.7645	0.7459	0.7600	0.7024	0.6964	0.6668	0.9089
GLM4f-0shot	0.6431	0.6149	0.6378	0.5702	0.5618	0.5367	0.8587
GLM4f-1shot	0.7226	0.6997	0.7178	0.6592	0.6520	0.6230	0.8993
GLM4f-2shot	0.7558	0.7337	0.7503	0.6946	0.6868	0.6541	0.9140
GLM4s-0shot	0.6764	0.6548	0.6738	0.6087	0.6022	0.5787	0.8807
GLM4s-1shot	0.7149	0.6937	0.7113	0.6540	0.6475	0.6237	0.8959
GLM4s-2shot	0.7682	0.7511	0.7643	0.7161	0.7103	0.6799	0.9196
GPT4o-0shot	0.6737	0.6461	0.6672	0.6013	0.5922	0.5588	0.8742
GPT4o-1shot	0.7300	0.7062	0.7237	0.6680	0.6598	0.6268	0.9092
GPT4o-2shot	0.7465	0.7223	0.7402	0.6812	0.6729	0.6397	0.9124
Deepseek-0shot	0.6363	0.6155	0.6333	0.5717	0.5651	0.5377	0.8366
Deepseek-1shot	0.7548	0.7388	0.7518	0.7019	0.6962	0.6658	0.9091
Deepseek-2shot	**0.7850**	**0.7679**	**0.7806**	**0.7379**	**0.7318**	**0.7012**	**0.9208**

Attribute Classification. LLMs are required to classify elements with predefined attributes into binary or multi-class categories. However, despite prompt instructions, LLMs often generate extraneous information during classification, which leads to underestimating performance using strict exact matching. To address this, we introduce a Flexible F1 Score, an extension of the traditional F1 Score, for a more accurate assessment of model performance.

In our evaluation, a classification is considered correct if the generated output includes the correct category, regardless of irrelevant information. For both binary and multi-class tasks, we redefine True Positives (TP), False Positives (FP), and False Negatives (FN) for a unified metric. TP refers to correctly predicted instances of attributes in the ground truth, FP to incorrect predictions, and FN to elements present in the ground truth but missing from the predictions. The evaluation metrics are calculated as follows: F1-Score considers both the precision p score and the recall r score to form a harmonic score: $\mathrm{F1} = \frac{2\mathrm{pr}}{\mathrm{p+r}}$, where $p = \frac{TP}{TP+FP}$ and $r = \frac{TP}{TP+FN}$. We evaluate the classification performance across 0-shot, 1-shot, and 2-shot settings. The results of the attribute classification experiment are presented in **Table 4**, which can serve as a benchmark for LLMs on this legal task.

4.2 Causal Estimation

To uncover factors influencing court judgment results, we design causal experiments using annotated elements to infer relationships between judgment outcomes and relevant elements. With expert knowledge, we construct a causal graph as prior knowledge, shown in **Fig. 5**. This directed acyclic graph represents possible causal relationships between variables. The goal is to verify these relationships on the predefined graph, with "E.30 Trial court judgment result" as the outcome variable Y and other relevant elements as treatment variables X_{tag}.

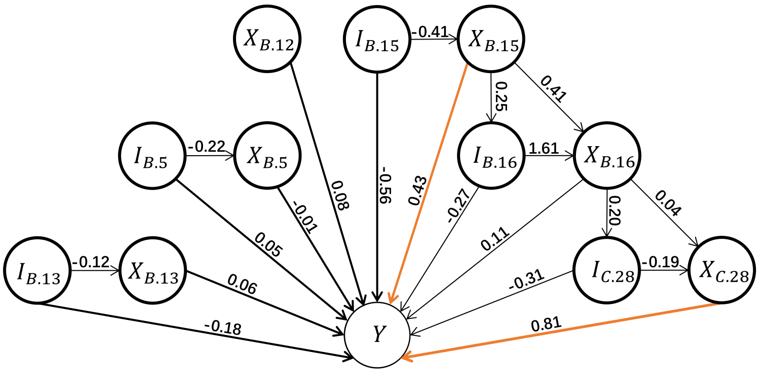


Fig. 5. The causal graph and estimate results, with ATE values marked on the edges. Colored edges represent the strongest causal relationships.

Since not every document contains all elements, we adopt an innovative causal graph structure to mitigate the missing value problem, inspired by MVPC [34]. Specifically, some indicator variables I_{tag} are introduced for treatment variables X_{tag} with missing values, $\mathrm{I_{tag}} \leftarrow \mathrm{True} \Leftrightarrow \mathrm{X_{tag}} = \mathrm{NaN}$. We use Average Treatment Effect (ATE) as a quantitative criterion for the causal effect. Let $X_T(T \subset \{1, \dots, D\})$ be the treatment variable, $p\big(X \big| do(X_T = a)\big)$ the interventional distribution, and Y the intervention target variable. Under the reference condition $X_T = b$, the ATE after applying the intervention $X_T = a$ is given by:

$$ATE(a, b) = E\big(Y(X_T = a)\big) - E\big(Y(X_T = b)\big).$$

To estimate the ATE between variables, we use the DoWhy tool [35], which implements graph-based criteria and do-calculus to calculate causal effects. We use the backdoor.linear_regression method to compute the ATE quantitatively as shown in **Fig. 5**. Excluding indicator variables I, $X_{C.28}$ has the strongest causal relationship with Y among all variables.

Table 4. Classification performance comparison. The highest score is highlighted in **bold**, while the lowest score is underlined.

	Llama			Qwen			GLM4f			GLM4s			Deepseek			GPT4o		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0-shot	0.566	0.451	0.469	0.740	0.478	0.548	0.616	0.940	0.725	0.726	0.932	0.800	0.692	0.917	0.768	0.635	**0.970**	0.746
1-shot	0.540	0.752	0.603	0.763	0.707	0.704	0.666	0.942	0.765	0.719	0.930	0.796	0.750	0.940	**0.819**	0.702	0.905	0.773
2-shot	0.652	0.879	0.726	**0.773**	0.794	0.758	0.694	0.955	0.786	0.746	0.917	0.807	0.741	0.942	0.814	0.734	0.903	0.795

5 Discussion

In this section, we first analyze the annotation results, drawing useful conclusions on utilizing LLMs. Recognizing the need to infer the attributes of elements in challenging samples, we outline our approach to designing effective prompts. Then, we evaluate the results of causal analysis using the annotated elements.

Annotation Result Analysis. The results show that ROUGE, BLEU, and BERTScore metrics strongly correlate, validating their effectiveness in evaluating annotation task. Recall (R) generally exceeds precision (P), indicating effective identification of relevant text, though there remains room for improvement in classification accuracy.

Llama3.1-8B performs the worst in both extraction and classification tasks, particularly under zero-shot conditions, constrained by its limited model size and learning capacity. While capable of general language processing, it struggles to meet the specialized precision demands of legal annotations. DeepSeek-Chat achieves the best performance in legal document annotation. Legal documents often contain dense legal terminology, intricate logical structures, and demand high-accuracy annotations, such as entity recognition and legal clause referencing. These tasks demand not only contextual understanding but also deep domain-specific knowledge. Unlike dense architecture models that rely on uniformly activated parameters and excel at general-purpose language generation, Deepseek adopts a Mixture of Experts (MoE) architecture. The architecture dynamically routes inputs to the most suitable expert networks, effectively handling scenarios that demand precise legal terminology understanding and complex reasoning.

The results highlight that few-shot learning significantly boosts model performance. Performance improves notably from 0 to 1 shot, with smaller gains observed from 1 to 2 shots. Few-shot learning enhances model generalization in low-resource scenarios, suggesting that a single example can effectively balance performance and resource constraints.

Prompt Engineering in Challenge Samples. Unlike daily communication, case judgments may go through multiple procedures, making a single sentence less credible. This often involves complex multi-hop problems. We will analyze the "Trial Court Judgment Result" element as an example.

In the document excerpt presented in **Fig. 6**, the case goes through three trial procedures: first instance, second instance, and retrial. The judgment result cannot be determined from a single sentence in the retrial; it must be considered in conjunction with the results of the first and second instances, forming a typical multi-hop question. In this case, the judgement result of the first instance is the labor relationship, but the second instance reverses it. The retrial court then reversed the second instance's ruling, reaffirming the first instance judgment. Therefore, the judgement result of the trial court (E.30) in this case is the labor relationship.

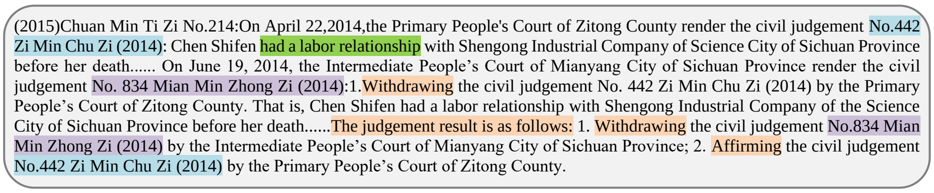

(2015)Chuan Min Ti Zi No.214:On April 22,2014,the Primary People's Court of Zitong County render the civil judgement No.442 Zi Min Chu Zi (2014): Chen Shifen had a labor relationship with Shengong Industrial Company of Science City of Sichuan Province before her death...... On June 19, 2014, the Intermediate People’s Court of Mianyang City of Sichuan Province render the civil judgement No. 834 Mian Min Zhong Zi (2014):1.Withdrawing the civil judgement No. 442 Zi Min Chu Zi (2014) by the Primary People’s Court of Zitong County. That is, Chen Shifen had a labor relationship with Shengong Industrial Company of the Science City of Sichuan Province before her death......The judgement result is as follows: 1. Withdrawing the civil judgement No.834 Mian Min Zhong Zi (2014) by the Intermediate People’s Court of Mianyang City of Sichuan Province; 2. Affirming the civil judgement No.442 Zi Min Chu Zi (2014) by the Primary People’s Court of Zitong County.

Fig. 6. A case sample with three trial procedures poses challenges in extracting judgment result.

Considering the multi-hop reasoning challenges in determining the attribution of the element "Trial court judgement result", we embed the human reasoning process into the prompt, along with a chain of thought to guide LLMs to reason through the different stages of trial procedure. After multiple iterations of experiments and adjustments, we use the following prompt:

Please extract the attribute value for the "Trial court judgment result" from the legal judgment. The possible values are "labor relationship" or "service relationship." Consider the trial procedure: For a first-instance judgment, directly output the current judgment result. For a second-instance judgment, be careful not to be influenced by the first-instance judgment result, but focus on how the second-instance handles the first-instance judgment result (upholding or correcting). For a retrial judgment, focus on how the retrial handles the first- and second-instance results (revoking or upholding). Please think step by step, find the relevant text and output the attribute value, avoiding any additional reasoning information. If no relevant content, return "0".

Verification of Inference Results. We conducted two refutation tests to verify the inference results: Add Random Common Cause (RC) and Placebo Treatment (PT). The RC test introduces a random confounder, expecting no significant change in the ATE. The PT test replaces the treatment variable with an independent random variable, expecting the ATE to converge to 0. The robustness results in **Table 5** confirm our inference is robust. Based on these results, we identified key factors influencing judgment outcomes and established a reasonable adjudication path. Our conclusions, validated by legal experts and consistent with legal facts, highlight the strong potential of our workflow in real-world judgment scenarios.

Table 5. Different treatments on missing value and ATE value of elements' relations. The ATE values reflect significant causal relationships are underlined, and the **bolded** values are the largest ATE value in that column. We report the p value of Add random common cause (RC) and Placebo treatment (PT) refute tests. All p values in the table are not statistically significant, which means that all ATE estimates in the table pass these two tests. In the table, B15→B16, B15→E30, and C28→E30, they have always maintained a prominent cause relationship in all missing value treatment situations. And C28→E30 is always the strongest intensity of causality.

Value / Relationship	Filling missing value with NaN			Filling missing value with Most Frequent			Filling missing value with NaN Add Indicator Variables		
	ATE	RC p value	PT p value	ATE	RC p value	PT p value	ATE	RC p value	PT p value
B5 → E30	-0.009	0.800	0.880	0.033	1.000	2.000	-0.011	0.940	0.980
B12 → E30	0.078	0.960	0.940	0.072	0.960	2.000	0.078	0.980	0.940
B13 → E30	0.059	0.980	0.860	0.002	0.960	2.000	0.058	0.980	0.840
B15 → B16	0.409	0.920	0.960	-0.219	1.000	2.000	0.409	0.840	0.980
B15 → E30	0.488	0.980	0.940	0.298	0.860	2.000	0.431	0.960	0.920
B16 → C28	0.035	0.960	0.920	-0.150	0.980	2.000	0.035	0.960	0.980
B16 → E30	-0.163	0.900	0.960	0.058	0.980	2.000	0.107	0.980	0.840
C28 → E30	0.851	0.980	0.960	0.455	0.920	2.000	0.807	0.940	0.800

6 Conclusion

This study presents a general workflow for legal case analysis based on legal case elements, effectively addressing the limitations of previous problem-specific methods and offering a unified analytical framework applicable to various legal analysis tasks. Collaborating with legal experts, we developed an element framework for extracting information from legal documents, instructing LLMs to annotate judgments. These extracted elements can be universally applied to tasks like case retrieval, judgment prediction, and legal view generation. To validate the workflow, we applied it to the domain of elderly laborers' protection, defining a legal case element framework including 6 structural elements and 32 evidence elements. We used LLMs to facilitate data annotation and designed metrics of the extraction and classification tasks, evaluating the LLMs performance on corresponding tasks as benchmarks. Besides, causal inference was used to identify key factors influencing judgment results at the causal level. Both the annotated datasets and workflow tools are open-sourced as benchmarks for the community, facilitating the application of large language models in law and other related fields.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Deng, C., Dou, Z., Zhou, Y., Zhang, P., Mao, K.: An element is worth a thousand words: Enhancing legal case retrieval by incorporating legal elements. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 2354–2365 (2024)
2. Wyner, A.Z.: Towards annotating and extracting textual legal case elements. Informatica e Diritto: special issue on legal ontologies and artificial intelligent techniques 19(1-2), 9–18(2010)

3. Gray, M., Savelka, J., Oliver, W., Ashley, K.: Using llms to discover legal factors. In: Legal Knowledge and Information Systems, pp. 60–71. IOS Press (2024)
4. Gray, M., Savelka, J., Oliver, W., Ashley, K.: Can gpt alleviate the burden of annotation? In: Legal Knowledge and Information Systems, pp. 157–166. IOS Press (2023)
5. Kalamkar, P., Tiwari, A., Agarwal, A., Karn, S., Gupta, S., Raghavan, V., Modi, A.: Corpus for automatic structuring of legal documents. arXiv preprint arXiv:2201.13125 (2022)
6. Zheng, L., Guha, N., Anderson, B.R., Henderson, P., Ho, D.E.: When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In: Proceedings of the eighteenth international conference on artificial intelligence and law. pp. 159–168 (2021)
7. Semo, G., Bernsohn, D., Hagag, B., Hayat, G., Niklaus, J.: Classactionprediction: A challeng ing benchmark for legal judgment prediction of class action cases in the us. arXiv preprint arXiv:2211.00582 (2022)
8. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: Jec-qa: a legal-domain question answering dataset. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 9701–9708 (2020)
9. Henderson, P., Krass, M., Zheng, L., Guha, N., Manning, C.D., Jurafsky, D., Ho, D.: Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. Advances in Neural Information Processing Systems 35, 29217–29234 (2022)
10. Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D.M., Aletras, N.: Lexglue: A benchmark dataset for legal language understanding in english. arXiv preprint arXiv:2110.00976 (2021)
11. Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Zhang, T., Han, X., Hu, Z., Wang, H., et al.: Cail2019-scm: A dataset of similar case matching in legal domain. arXiv preprint arXiv:1911.08962 (2019)
12. Urchs, S., Mitrovic, J., Granitzer, M.: Design and implementation of german legal decision corpora. In: ICAART (2). pp. 515–521 (2021)
13. Louis, A., Spanakis, G.: A statutory article retrieval dataset in french. arXiv preprint arXiv:2108.11792 (2021)
14. Chalkidis, I., Fergadiotis, M., Androutsopoulos, I.: Multieurlex–a multi-lingual and multi label legal document classification dataset for zero-shot cross-lingual transfer. arXiv preprint arXiv:2109.00904 (2021)
15. Niklaus, J., Chalkidis, I., Stürmer, M.: Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. arXiv preprint arXiv:2110.00806 (2021)
16. Aumiller, D., Chouhan, A., Gertz, M.: Eur-lex-sum: A multi-and cross-lingual dataset for long-form summarization in the legal domain. arXiv preprint arXiv:2210.13448 (2022)
17. Hendrycks, D., Burns, C., Chen, A., Ball, S.: Cuad: An expert-annotated nlp dataset for legal contract review. arXiv preprint arXiv:2103.06268 (2021)
18. Holzenberger, N., Blair-Stanek, A., Van Durme, B.: A dataset for statutory reasoning in tax law entailment and question answering. arXiv preprint arXiv:2005.05257 (2020)
19. Yamada, H., Tokunaga, T., Ohara, R., Takeshita, K., Sumida, M.: Annotation study of japanese judgments on tort for legal judgment prediction with rationales. In: Proceedings of the thirteenth language resources and evaluation conference. pp. 779–790 (2022)
20. Vihikan, W.O., Mistica, M., Levy, I., Christie, A., Baldwin, T.: Automatic resolution of domain name disputes. In: Proceedings of the Natural Legal Language Processing Workshop 2021. pp. 228–238 (2021)
21. Papaloukas, C., Chalkidis, I., Athinaios, K., Pantazi, D.A., Koubarakis, M.: Multi-granular legal topic classification on greek legislation. arXiv preprint arXiv:2109.15298 (2021)

22. Malik, V., Sanjay, R., Nigam, S.K., Ghosh, K., Guha, S.K., Bhattacharya, A., Modi, A.: Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. arXiv preprint arXiv:2105.13562 (2021)
23. Yue, L., Liu, Q., Zhao, L., Wang, L., Gao, W., An, Y.: Event grounded criminal court view generation with cooperative (large) language models. In: Proceedings of the 47th International ACMSIGIRConference on Research and Development in Information Retrieval. pp. 2221 2230 (2024)
24. Chen,G.,Fan,L.,Gong,Z.,Xie,N.,Li, Z., Liu, Z., Li, C., Qu, Q., Ni, S., Yang, M.: Agentcourt: Simulating court with adversarial evolvable lawyer agents. arXiv preprint arXiv:2408.08089 (2024)
25. Sun, J., Dai, C., Luo, Z., Chang, Y., Li, Y.: Lawluo: A chinese law firm co-run by llm agents. arXiv preprint arXiv:2407.16252 (2024)
26. Izzidien, A., Sargeant, H., Steffek, F.: Llm vs. lawyers: Identifying a subset of summary judgments in a large uk case law dataset. arXiv preprint arXiv:2403.04791 (2024)
27. Cheong, I., Xia, K., Feng, K.K., Chen, Q.Z., Zhang, A.X.: (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 2454–2469 (2024)
28. Cheong, I., Caliskan, A., Kohno, T.: Envisioning legal mitigations for llm-based intentional and unintentional harms. Administrative Law Journal (2022)
29. Stumptner, M., Mayer, W., Grossmann, G., Liu, J., Li, W., Casanovas, P., De Koker, L., Mendelson, D., Watts, D., Bainbridge, B.: An architecture for establishing legal semantic workflows in the context of integrated law enforcement. In: AI Approaches to the Complexity of Legal Systems: AICOL International Workshops 2015-2017: AICOL-VI@ JURIX 2015, AICOL-VII@ EKAW 2016, AICOL-VIII@ JURIX 2016, AICOL-IX@ ICAIL 2017, and AICOL-X@JURIX 2017, Revised Selected Papers 6. pp. 124–139. Springer (2018)
30. Mayer, W., Casanovas Romeu, P., Stumptner, M., De Koker, L., Mendelson, D.: Semantic workflows in law enforcement investigations and legal requirements (2017)
31. Di Martino, B., Colucci Cante, L., Graziano, M., D’Angelo, S., Esposito, A., Lupi, P., Am mendolia, R.: A semantic-based methodology for the management of document workflows in e-government: a case study for judicial processes. Knowledge and Information Systems pp. 1–29 (2024)
32. Robak, M., Buchmann, E.: How to extract workflow privacy patterns from legal documents. In: Information Technology for Management: Current Research and Future Directions: 17th Conference, AITM 2019, and 14th Conference, ISM 2019, Held as Part of FedCSIS, Leipzig, Germany, September 1–4, 2019, Extended and Revised Selected Papers 17. pp. 214–234. Springer (2020)
33. Zhang, T., Zeng, X., Liu, Z.: Modeling workflow for judicial business processes: A use case driven method. In: 2021 7th International Conference on Information Management (ICIM). pp. 45–56. IEEE (2021)
34. Tu, R., Zhang, C., Ackermann, P., Mohan, K., Kjellström, H., Zhang, K.: Causal discovery in the presence of missing data. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 1762–1770. Pmlr (2019)
35. Sharma, A., Syrgkanis, V., Zhang, C., Kıcıman, E.: Dowhy: Addressing challenges in express ing and validating causal assumptions. arXiv preprint arXiv:2108.13518 (2021)