Artificial Intelligence and Law
https://doi.org/10.1007/s10506-025-09460-y

ORIGINAL RESEARCH

Specialized or general AI? a comparative evaluation of LLMs' performance in legal tasks

Xue Guo[1,4] · Yuting Huang[1] · Bin Wei[2,3] · Kun Kuang[1,3] · Yiquan Wu[1,3] · Leilei Gan[1,3] · Xianshan Huang[1] · Xianglin Dong[2,3]

Accepted: 14 April 2025

Abstract
The rise of large language models (LLMs) such as ChatGPT and GPT-4 developed by OpenAI have generated significant interest in the legal domain due to their sophisticated language processing capabilities. In particular, regions like China are vigorously developing legal-specific LLMs for legal purposes. Fine-tuned with fewer parameters and based on judicial documents and Chinese case data sets, these specialized LLMs are widely expected to meet practical needs in the judicial field more effectively. However, the ability of these law-specific LLMs to perform legal tasks and their potential to outperform general LLMs has not yet been established. To fill in this research gap, we systematically evaluate a range of general and legal-specific LLMs on various legal tasks. The results show that GPT-4 maintains superior performance on most legal tasks, although legal-specific LLMs show superior performance in specific cases. This study provides insight into the factors leading to these results, hoping to enrich the discourse on the use of LLMs in the legal field.

Keywords LLMs · Legal-specific LLMs · Evaluation · Legal tasks

1 Introduction

Large language models (LLMs) based on a large scale can solve NLP tasks zero-shot without training any data for downstream tasks, but only by conditioning the model on appropriate prompts. Models, such as ChatGPT[1] and GPT-4(Achiam et al. 2023), have demonstrated remarkable performance on various natural language tasks. Nevertheless, when applying them to highly specialized and safety-critical legal domains, it is unclear whether they can also retain remarkable performance.

[1] https://chat.openai.com/

This paper is submitted for the Special Issue on Applications and Evaluation of Large Language Models in the Legal Domain.

Extended author information available on the last page of the article

Published online: 14 May 2025

Table 1 Chinese legal-specific LLMs

Model	Parameters	SFT	RLHF	BaseModel	Time
HanFei	7B	✓	✗	HanFei	2023-05
WisdomInterrogatory	7B	✓	✗	Baichuan-7B	2023-08
LawGPT-zh	6B	✓	✗	ChatGLM-6B	2023-04
FuziMingcha	6B	✓	✗	ChatGLM-6B	2023-09
ChatLaw	13/33B	✓	✗	Ziya-LLaMA-13B /Anima-33B	2023-07
Lawyer-LLaMA	13B	✓	✗	Chinese-LLaMA-13B	2023-05
LexiLaw	6B	✓	✗	ChatGLM-6B	2023-05
Lychee	10B	✓	✗	GLM-10B	2023-07
JurisLMs	–	✓	✗	GPT-2	2023-05
DISC-LawLLM	13B	✓	✗	Baichuan-13B	2023-09
BAI-Law-13B	13B	✓	✗	LLama2	2023-12
Elpis	–	✓	✓	–	2023-04

OpenAI reported that GPT-4 had passed the Uniform Bar Examination (UBE) with a score of 90, but subsequent research proved it was overstated. Furthermore, an American lawyer with over 30 years of experience faced charges of using LLMs to generate unrealistic cases, reinforcing concerns about the potential risks associated with the inappropriate application of LLMs in the legal domain. LLMs improve efficiency in the legal domain, but there are some existing problems, including inaccuracies in citing legal provisions, delays in updating datasets, and the generation of unrealistic text. To address these issues, universities and research institutions have introduced specialized LLMs for legal scenarios called legal-specific large language models, which are based on general LLMs and fine-tuned with high-quality legal data to improve the model's accuracy in legal tasks. Countries represented by China have vigorously developed their legal-specific LLMs, aiming to provide people with more secure and reliable legal aid. Many LLMs based on Chinese law have been developed, such as FuziMingcha (Wu et al. 2023), WisdomInterrogatory,[2] and ChatLaw(Cui et al. 2023), as shown in Table 1.

The legal-specific LLMs help legal workers to retrieve legal articles, analyze the focus of disputes, draft legal contracts, summarize facts, conduct reasoning, and give judgment suggestions, also helping individuals with limited legal knowledge to obtain preliminary legal advice. However, a systematic and transparent evaluation is needed to determine the extent to which legal-specific LLMs can assist humans. Additionally, it is important to consider whether legal-specific models necessarily outperform generic models for legal tasks. We address these questions by systematically evaluating several representative general and legal-specific LLMs across a range of legal tasks. To the best of my knowledge, there has been no systematic research and analysis conducted on this issue. This paper systematically compares general and legal-specific LLMs across 11 legal tasks chosen from various

[2] https://github.com/zhihaiLLM/wisdomInterrogatory

dimensions of understanding and generation, to account for the diverse abilities of LLMs in processing legal information. Experimental results indicate that GPT-4 remains the top-performing LLM in the legal domain. Nonetheless, legal-specific LLMs demonstrate potential superiority in certain specialized scenarios, such as predicting relative articles and charges.

It is worth noting that the presentation of the structure and features of investigated cases is an essential prerequisite to evaluating particular models' performance. This paper analyzes the structure of the data set used in all legal tasks. To improve the generalizability of evaluation results, we introduced datasets of other countries apart from China in the argument focus mining task. We found that the evaluation results were consistent with CAIL2020. We provide novel techniques like Chain-of-thoughts(CoT) prompting, and few-shot techniques in legal charge prediction (a sub-task of legal judgment prediction). Furthermore, we introduced human evaluation to find the shortcomings of automated evaluation. The results show that although there are some problems with automated evaluation, they do not affect the reliability of automated evaluation, which is still credible.

The contributions of this paper are as follows:

- We conduct extensive experiments comparing the performance of general LLMs with that of Chinese legal-specific LLMs, comprehensively evaluating these models' ability to process legal issues.
- Tasks based on Chinese legal scenarios are established. After discussing with legal experts, we summarized several legal tasks that are highly relevant to practical legal needs, categorizing them into understanding and generation types, and weighing the potential advantages and disadvantages of LLMs associated with such legal tasks.
- We validate that legal-specific models may not necessarily perform superior to the general model in some tasks, especially not necessarily well than GPT-4.
- We delve into the question of why there are still some general LLMs outperform Chinese legal-specific LLMs at tasks, and whether the fine-tuned process of Chinese legal-specific LLMs is a positive or negative development.
- We introduce a variety of novel techniques in legal charge prediction, such as CoT and few-shot prompting, which will provide a reference to improve the output quality of the LLMs.

This paper's structure is as follows: Sect. 2 presents a review of the research background, encompassing the development of legal-specific LLMs, LLM evaluation, and the intersection of Al&Law. Following that, Sect. 3 outlines the research methods employed in this paper, encompassing LLM selection, scenario-based legal tasks, datasets, and associated evaluation metrics. Section 4 provides a detailed account of the experimental process, where diverse models undergo evaluation across various legal tasks. This section encompasses four primary objectives: (1)Presenting all experimental results; (2) Analyzing LLM performance across diverse legal tasks and assessing their efficacy in the legal domain; (3) Identifying common issues in current LLMs and proposing potential enhancements. (4) Human evaluation is conducted to find problems in the automatic evaluation process. Finally, Sect. 5 summarizes the paper's findings and outlines future research directions.

2 Background and related work

This section reviews pertinent literature, focusing on the prevalent approach for constructing LLMs, the development of LLM evaluation, and the status of AI& Law.

2.1 Legal-specific LLMs

LLMs are distinguished by their extensive training on vast corpora, which enables them to generate high-quality text. Having billions of parameters, these models show remarkable capabilities. Moreover, LLMs are talented at few-shot learning, allowing them to execute tasks with minimal examples (Radford et al. 2019).

In the legal domain, the application of LLMs necessitates more refined processing. Unlike general-purpose LLMs, those tailored for the legal domain must not only master concrete and detailed legal knowledge but also require a more nuanced understanding of various legal tasks. To address this requirement, several legal-specific LLMs have been proposed, such as ChatLaw (Cui et al. 2023), Wisdom-Interrogatory,[3] and FuziMingcha,[4] etc. Within these legal LLMs, the mainstream methods include Continued Pre-Training (CPT), Supervised Fine-Tuning (SFT), and Retrieval-Augmented Generation (RAG). (1) CPT aims to specialize general models in legal contexts by secondary pre-training on legal corpora, such as laws, precedents, and legal documents, to enhance their understanding and analysis of legal language and concepts (Gururangan et al. 2020), thereby transforming them into models proficient in legal document analysis. (2) SFT aims to enhance model comprehension for specific legal tasks by fine-tuning datasets filled with diverse legal instructions and questions, such as event summaries and law applications and employs a CFT model to generate targeted instructions, thus improving its efficacy in legal Q&A scenarios (Zhang et al. 2023). (3) By integrating retrieval and generation with knowledge enhancement, RAG models improve the precision and trustworthiness of their legal text outputs, using relevant legal data and literature as references to inform their generated responses and ensure greater interpretability of the results through the use of legal knowledge retrieval (Lewis et al. 2020). These methods collectively contribute to the advancement of LLMs in the legal domain.

2.2 Evaluation of LLMs

Evaluation has become increasingly important to assess the capabilities of LLMs quantitatively. Given the large scale of LLMs, traditional evaluations for language models are impractical for the evaluation of LLMs, such as K-fold cross-validation

[3] https://modelscope.cn/studios/wisdomOcean/wisdomInterrogatory/summary

[4] https://github.com/irlab-sdu/fuzi.mingcha

(Rodriguez et al. 2009), Holdout cross-validation (Sapatinas 2005), Monte Carlo cross-validation (Xu and Liang 2001), Leave One Out cross-Validation (Wong 2015), Time Series cross-validation (Cerqueira et al. 2020; Bergmeir and Benítez 2012), etc.

Previous efforts have predominantly concentrated on all kinds of aspects of LLMs, aiming to evaluate the general capabilities of LLMs, such as HELM (Liang et al. 2022), which measures 7 metrics (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency) for each of 16 core scenarios, etc., and evaluates 30 prominent language models from 12 organizations around the world. Big-Bench (Srivastava et al. 2023) evaluates the performance of various tasks of LLMs, consisting of 204 tasks. The tasks involve linguistics, child development, mathematics, common sense reasoning, biology, physics, social prejudice, and other fields. Furthermore, widely adopted evaluation benchmarks comprise GLUE (Wang et al. 2018) and SuperGLUE (Wang et al. 2020a), which serve as common test sets to assess the natural language understanding of LLMs.The evaluation methods can be categorized into two types: the holistic evaluation represented by Liang (2022), which involves multiple tasks but also requires high computational costs and human costs. The other is the evaluation in specific scenarios, targeting the specific aspect of LLMs. For instance, Sun et al. (2023) evaluated the text classification ability of LLMs and introduced the CARP strategy to enhance classification performance. Similarly, Wang et al. (2023) assessed the summarization ability of LLMs, employing various prompts to guide LLMs in improving their summarization task performance.

In AI&Law, (Conrad and Zeleznikow 2015) has evaluated legal tasks and emphasized the importance of performance evaluation in this field (Savelka and Ashley 2023) compared the performance of GPT-4 to the previous generation of GPT models on three legal text annotation tasks involving diverse documents. They found that the GPT-4 is the best-performing model outperforming others on most tasks. They also discovered that analyzing court opinions to interpret legal concepts GPT-4, prompted with annotation guidelines, could perform on par with well-trained law student annotators (Savelka et al. 2023a). The other research involves how GPT-4 generates factual explanations of terms in legislation (Savelka et al. 2023b) and thematic analysis in empirical legal studies (Drápal et al. 2023).

2.3 AI&Law for legal tasks

The classical NLP tasks include reasoning (Wei et al. 2023), summarization (Sun et al. 2021), named entity recognition (Tjong Kim Sang and De Meulder 2003) and sentiment analysis (Socher et al. 2013), etc. However, in the legal domain, there are some specific features: The legal documents usually have extremely long text and complex structure, which make them difficult to comprehend. Moreover, determined by the particularity of the legal field, the accuracy and credibility of results are highly required. Researchers have created a series of language models for legal tasks based on these characteristics. Before the appearance of LLMs, almost all models were designed for a specific task (Xu et al. 2020; Li et al. 2021; Shao et al. 2020).

When the LLM era arrived, a model(LLM) based on a large scale could handle a variety of NLP tasks by conditioning the model on appropriate prompts, which is also suitable for the legal domain (Savelka et al. 2023b; Savelka and Ashley 2023).

This paper divides legal tasks into two major categories: understanding and generation. The understanding means that the model can understand legal documents' legal concepts, entities, and relationships. Its subtasks include legal judgment prediction (Chalkidis et al. 2019), judicial examination (Zhong et al. 2020), argument focus mining (Poudyal et al. 2020a), reading comprehension (Xiao et al. 2021; Zhong et al. 2020), entity recognition (Leitner et al. 2019), relation extraction (Nasar et al. 2021; Chen et al. 2020), event detection (Li et al. 2020; Cai et al. 2024), document proofreading (Li et al. 2023; Lin et al. 2023), etc. The generation tasks are used to test whether the models could generate correct and reliable sentences from source files, which include judgment summarization (Deroy et al. 2024; Jain et al. 2021, 2023), public opinion summarization(Huang et al. 2020; Hayatin et al. 2021), and court opinion (Wu et al. 2020; Ye et al. 2018). The details will be discussed in Sect. 3.

3 Experimental method

3.1 Selection of LLMs

This study aims to contrast general LLMs' efficacy with legal-specific LLMs across various legal tasks. Consequently, selecting suitable and representative models for comparative analysis is imperative. Regarding general LLMs, we opt for both multilingual and Chinese-oriented variants. The chosen models are displayed in Table 2.

ChatGPT was trained based on the GPT−3.5 series model with reinforcement learning from human feedback (RLHF) (Christiano et al. 2023).RLHF includes training a model with supervised learning, collecting comparison data based on human preferences, training a reward model, and optimizing the language model against the reward model using reinforcement learning (Ouyang et al. 2022). GPT-4(Achiam et al. 2023)is the most advanced natural language processing model in the world. It is also the generation of the GPT model series released by OpenAI, which has powerful natural language generation capabilities.

Table 2 LLMs evaluated on legal tasks

General models		Legal specific models
Multilingual	Chinese	WisdomInterrogatory, FuziMingcha
ChatGPT,GPT-4	Baichuan2(7b), GLM2(6b),Qwen,Internlm2	

Chinese-oriented LLMs are proposed to enhance the capabilities of LLMs to resolve Chinese NLP tasks. They are either pre-trained from scratch on Chinese corpora or perform SFT on Chinese instruction data. Here, we select Baichuan2(7B) (Yang et al. 2023), GLM2(6B),[5]Qwen (Bai et al. 2023)and Internlm2[6] as the test objects. They are all open-source Chinese LLMs pre-trained on Chinese corpus data sets.

As to legal-specific LLMs, we chose models that are already open-source and easy to use. WisdomInterrogatory is pre-trained on the basic model(baichuan-7B) with 40 G Chinese judicial knowledge and then performs instruction fine-tuning to improve the model's ability to communicate with users. In the fine-tuning phase, the developer performs multi-stage fact enhancement on the legal knowledge base, which is aimed at enhancing the interpretability of the results generated by the model. FuziMingcha(Wu et al. 2023) is a Chinese judicial model based on ChatGLM. It is trained on massive Chinese unsupervised judicial corpus (judgment documents, laws, regulations, etc.) and supervised judicial fine-tuning data (legal question answering, similar case retrieval). The model supports functions of law retrieval, case analysis, syllogistic reasoning, and judicial dialogue.

3.2 Taxonomy of legal tasks and data source

In this section, we will introduce different legal tasks and their corresponding datasets. Moreover, the structures of cases(chosen from datasets) will be displayed in detail. To thoroughly assess LLMs, we categorize legal tasks into understanding and generation. Understanding refers to how models capture vital legal information, such as legal concepts, elements, issues, entities, relationships, etc. This assesses the model's fundamental information-processing ability. The generation capacity determines whether the model can generate accurate and reliable legal text. Our research primarily uses data from the Challenge of AI In Law(CAIL)[7] shown in Table 3, sourced from criminal legal documents on the China Judgment Online.[8] The dataset of the CAIL can contain multiple subsets for different tasks, such as CAIL2020, which contains different subsets for argument focus mining, judicial reading comprehension, and judgment summarization tasks. Besides a series of CAIL datasets, we also used three additional datasets as supplementary: JEC-QA,[9] AC-NLG[10] and European Court of Human Rights(ECHR).[11] We have chosen and displayed the corresponding prompts in Figs. 1, 2, and 3.

[5] https://github.com/thudm/chatglm2-6b

[6] https://github.com/InternLM/InternLM

[7] http://cail.cipsc.org.cn/

[8] http://wenshu.court.gov.cn/

[9] http://jecqa.thunlp.org/

[10] https://github.com/wuyiquan/AC-NLG

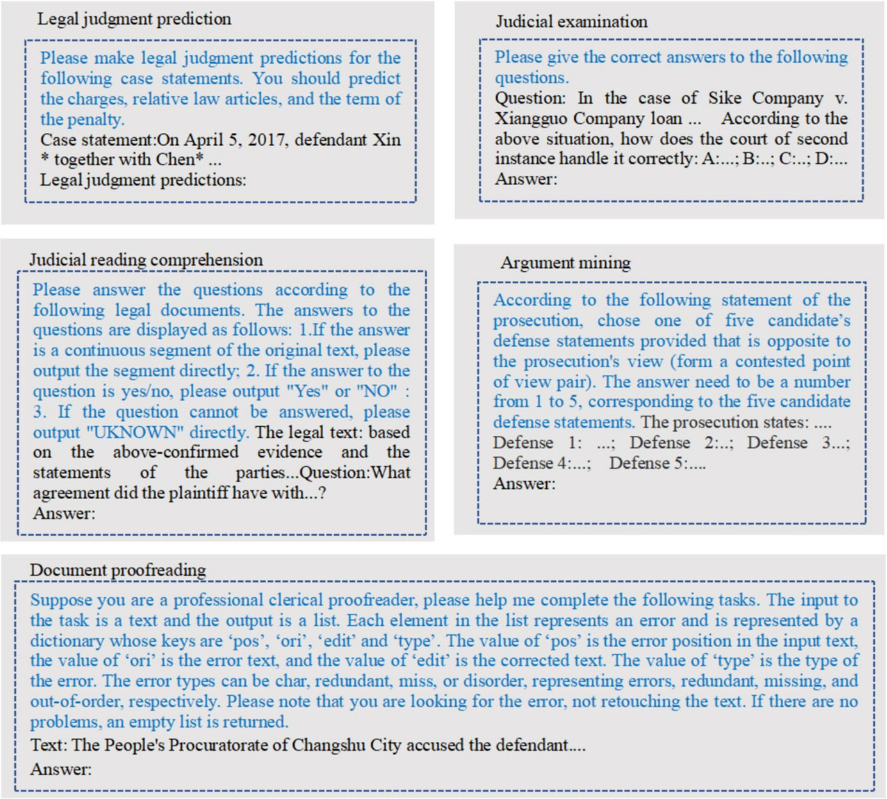
[11] http://www.di.uevora.pt/~pq/echr/

Table 3 Datasets of CAIL

Source	Legal task	Sample size	URL
CAIL2018	Legal Judgment Prediction	20000	http://cail.oss-cn-qingdao.aliyuncs.com/CAIL2018_ALL_DATA.zip
CAIL2020	Argument Focus Mining	4000	http://cail.cipsc.org.cn/task_summit.html?raceID=3&cail_tag=2020
	Judicial Reading Comprehension	8000	http://cail.cipsc.org.cn/task_summit.html?raceID=0&cail_tag=2020
	Judgment Summarization	8400	http://cail.cipsc.org.cn/task_summit.html?raceID=1&cail_tag=2020
CAIL2022	Entity Recognition	4000	http://cail.cipsc.org.cn/task_summit.htmlraceID=6cail_tag=2022
	Relation Extraction	4000	http://cail.cipsc.org.cn/task_summit.htmlraceID=6cail_tag=2022
	Event Detection	35000	http://cail.cipsc.org.cn/task_summit.htmlraceID=6&cail_tag=2022
	Document Proofreading	10000	http://cail.cipsc.org.cn/task_summit.html?raceID=2&cail_tag=2022
	Public Opinion Summarization	5000	http://cail.cipsc.org.cn/task_summit.html?raceID=4&cail_tag=2022

3.2.1 Understanding

- *Legal judgment prediction.* Given the fact description of a legal case, the judgment prediction's goal is to predict relative law articles, charges, and terms of prison. The prediction of legal judgment (LJP) has been investigated in various jurisdictions(Malik et al. 2021). It is worth noting that judges in the Chinese legal system have discretion, but usually in more complex cases that include ethical, moral, and emotional elements. The criminal cases in the data set in this paper are cases where the facts and regulations are applied explicitly, and the judge only needs to make a judgment according to the applicable provisions of the criminal law. CAIL2018(Xiao et al. 2018) is a large-scale dataset for LJP, consisting of two parts: fact description of cases and corresponding judgment results, not concerning evidence and procedural issues. The judgment result of each case includes relevant law articles, charges, and prison terms. CAIL2018 only retains cases with a single defendant to decrease the difficulty of cases and is mainly constructed on single charge cases (only contains a few cases of multiple charges, and the maximum number of charges is 4). For convenience, we'll only test the cases with a single charge. The complexity of the cases in the CAIL2018 is different: We compared the charge prediction results of high-frequency cases(such as dangerous driving, theft, fraud, intentional injury, etc.) to low-frequency cases(deforestation, obstruction of official duties, etc), and the results show that the values of the accuracy of the former are significantly higher than those of the latter. In addition, because all cases in CAIL2018 are



Legal judgment prediction

Please make legal judgment predictions for the following case statements. You should predict the charges, relative law articles, and the term of the penalty.
Case statement:On April 5, 2017, defendant Xin * together with Chen* ...
Legal judgment predictions:

Judicial examination

Please give the correct answers to the following questions.
Question: In the case of Sike Company v. Xiangguo Company loan ... According to the above situation, how does the court of second instance handle it correctly: A:...; B:..; C:..; D:...
Answer:

Judicial reading comprehension

Please answer the questions according to the following legal documents. The answers to the questions are displayed as follows: 1.If the answer is a continuous segment of the original text, please output the segment directly; 2. If the answer to the question is yes/no, please output "Yes" or "NO" : 3. If the question cannot be answered, please output "UKNOWN" directly. The legal text: based on the above-confirmed evidence and the statements of the parties...Question:What agreement did the plaintiff have with...?
Answer:

Argument mining

According to the following statement of the prosecution, chose one of five candidate's defense statements provided that is opposite to the prosecution's view (form a contested point of view pair). The answer need to be a number from 1 to 5, corresponding to the five candidate defense statements. The prosecution states: ....
Defense 1: ...; Defense 2:..; Defense 3...; Defense 4:...; Defense 5:....
Answer:

Document proofreading

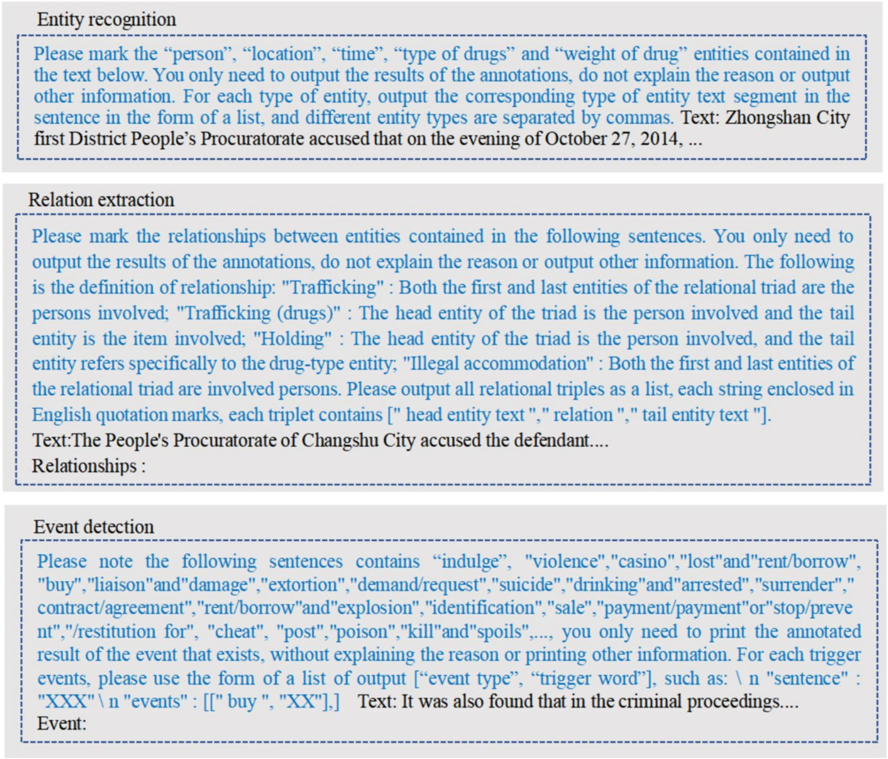
Suppose you are a professional clerical proofreader, please help me complete the following tasks. The input to the task is a text and the output is a list. Each element in the list represents an error and is represented by a dictionary whose keys are 'pos', 'ori', 'edit' and 'type'. The value of 'pos' is the error position in the input text, the value of 'ori' is the error text, and the value of 'edit' is the corrected text. The value of 'type' is the type of the error. The error types can be char, redundant, miss, or disorder, representing errors, redundant, missing, and out-of-order, respectively. Please note that you are looking for the error, not retouching the text. If there are no problems, an empty list is returned.
Text: The People's Procuratorate of Changshu City accused the defendant....
Answer:

Fig. 1 Instructions and input formats of understanding legal tasks (including judgment prediction, judicial examination, reading comprehension, argument mining, and document proofreading). Instructions are colored in blue. The model generates an answer after reading the entire input (circled by the blue dashed box)

taken from real courts,[12] most of the accused are convicted. An example of legal judgment prediction case structure is shown in Table 4, and the term of prison is calculated in months.

- *Judicial examination*.The judicial examination, administered by the Ministry of Justice, is a nationally recognized qualification test to select competent legal professionals. The examination usually comprises subjective and objective questions. In our research, only multiple-choice questions are selected as experimental data. A multiple-choice question is deemed correct only if all options are correct. JEC-QA is a large question-answering dataset in the legal domain, collecting questions from the National Judicial Examination of China (NJEC) (Zhong et al. 2020). NJEC is the legal professional certification examination for

[12] http://wenshu.court.gov.cn/



Entity recognition

Please mark the "person", "location", "time", "type of drugs" and "weight of drug" entities contained in the text below. You only need to output the results of the annotations, do not explain the reason or output other information. For each type of entity, output the corresponding type of entity text segment in the sentence in the form of a list, and different entity types are separated by commas. Text: Zhongshan City first District People's Procuratorate accused that on the evening of October 27, 2014, ...

Relation extraction

Please mark the relationships between entities contained in the following sentences. You only need to output the results of the annotations, do not explain the reason or output other information. The following is the definition of relationship: "Trafficking" : Both the first and last entities of the relational triad are the persons involved; "Trafficking (drugs)" : The head entity of the triad is the person involved and the tail entity is the item involved; "Holding" : The head entity of the triad is the person involved, and the tail entity refers specifically to the drug-type entity; "Illegal accommodation" : Both the first and last entities of the relational triad are involved persons. Please output all relational triples as a list, each string enclosed in English quotation marks, each triplet contains [" head entity text "," relation "," tail entity text "].
Text:The People's Procuratorate of Changshu City accused the defendant....
Relationships :

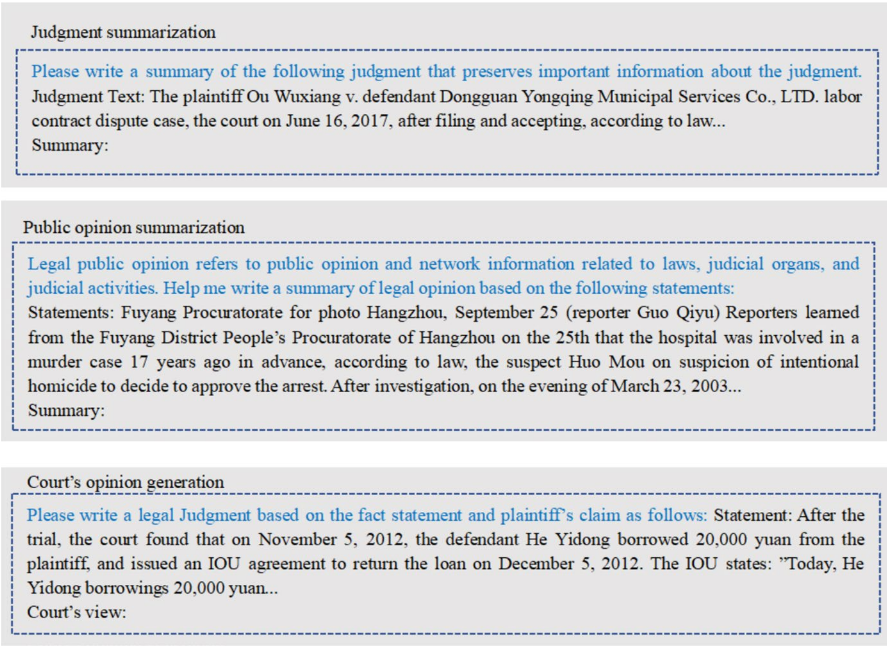
Event detection

Please note the following sentences contains "indulge", "violence","casino","lost"and"rent/borrow", "buy","liaison"and"damage","extortion","demand/request","suicide","drinking"and"arrested","surrender"," contract/agreement","rent/borrow"and"explosion","identification","sale","payment/payment"or"stop/prevent","/restitution for", "cheat", "post","poison","kill"and"spoils",..., you only need to print the annotated result of the event that exists, without explaining the reason or printing other information. For each trigger events, please use the form of a list of output ["event type", "trigger word"], such as: \ n "sentence" : "XXX" \ n "events" : [[" buy ", "XX"],] Text: It was also found that in the criminal proceedings....
Event:

Fig. 2 Instructions and input formats of understanding legal tasks(including entity recognition, relation extraction, and event detection). Instructions are colored in blue. The model generates an answer after reading the entire input (circled by the blue dashed box)

those who want to be lawyers or judges in China. There are 26,365 multiple-choice questions, which cover all legal knowledge required by the examination and provide useful extra labels for questions. An example of judicial examination case structure in JEC-QA is shown in Fig. 5.

- *Argument focus mining*. This task automatically identifies controversial points from both sides in the court record and extracts the focal disputes of the case. In this task, the model is presented with a prosecution opinion and five candidate viewpoints from the defendant, and it must automatically select one viewpoint to create a paired controversial point of dispute with the prosecution opinion. As is known, any argument may be attacked on different bases, such as questioning the premises (undermining), questioning the relationship between the premises and the conclusion (undercutting), or presenting an autonomous argument supporting the conclusion contrary to the conclusion posed by the opponent (rebuttal). The options in the dataset include all three attack relationships but are not counted or classified, which are generally treated here as relationships of attack or questioning. An example of case structure in the subset of the CAIL2020 for argument focus mining is displayed in Fig. 6. Additionally, we perform the task



Judgment summarization

Please write a summary of the following judgment that preserves important information about the judgment. Judgment Text: The plaintiff Ou Wuxiang v. defendant Dongguan Yongqing Municipal Services Co., LTD. labor contract dispute case, the court on June 16, 2017, after filing and accepting, according to law...
Summary:

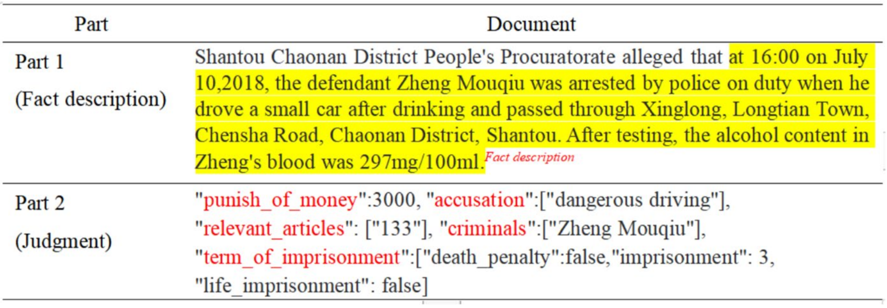
Public opinion summarization

Legal public opinion refers to public opinion and network information related to laws, judicial organs, and judicial activities. Help me write a summary of legal opinion based on the following statements:
Statements: Fuyang Procuratorate for photo Hangzhou, September 25 (reporter Guo Qiyu) Reporters learned from the Fuyang District People's Procuratorate of Hangzhou on the 25th that the hospital was involved in a murder case 17 years ago in advance, according to law, the suspect Huo Mou on suspicion of intentional homicide to decide to approve the arrest. After investigation, on the evening of March 23, 2003...
Summary:

Court's opinion generation

Please write a legal Judgment based on the fact statement and plaintiff's claim as follows: Statement: After the trial, the court found that on November 5, 2012, the defendant He Yidong borrowed 20,000 yuan from the plaintiff, and issued an IOU agreement to return the loan on December 5, 2012. The IOU states: "Today, He Yidong borrowings 20,000 yuan...
Court's view:

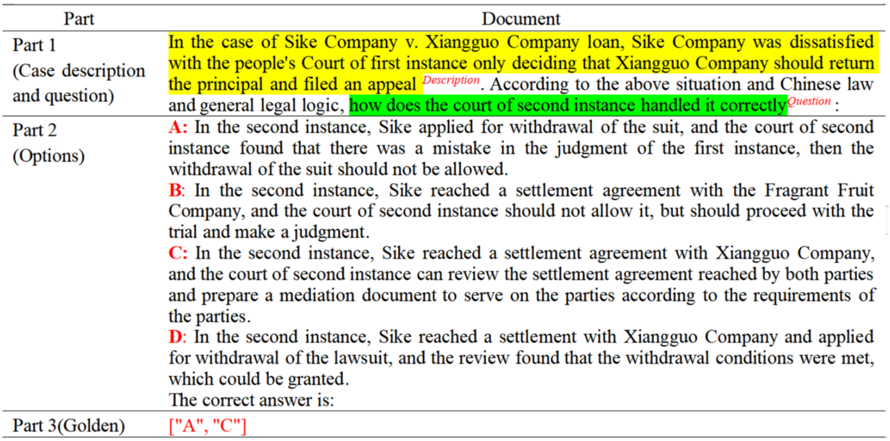
Fig. 3 Instructions and input formats of generation legal tasks(including judgment summarization, public opinion summarization, court's opinion generation). Instructions are colored in blue. The model generates an answer after reading the entire input (circled by the blue dashed box)



Part	Document
Part 1 (Fact description)	Shantou Chaonan District People's Procuratorate alleged that at 16:00 on July 10,2018, the defendant Zheng Mouqiu was arrested by police on duty when he drove a small car after drinking and passed through Xinglong, Longtian Town, Chensha Road, Chaonan District, Shantou. After testing, the alcohol content in Zheng's blood was 297mg/100ml.*Fact description*
Part 2 (Judgment)	"punish_of_money":3000, "accusation":["dangerous driving"], "relevant_articles": ["133"], "criminals":["Zheng Mouqiu"], "term_of_imprisonment":["death_penalty":false,"imprisonment": 3, "life_imprisonment": false]

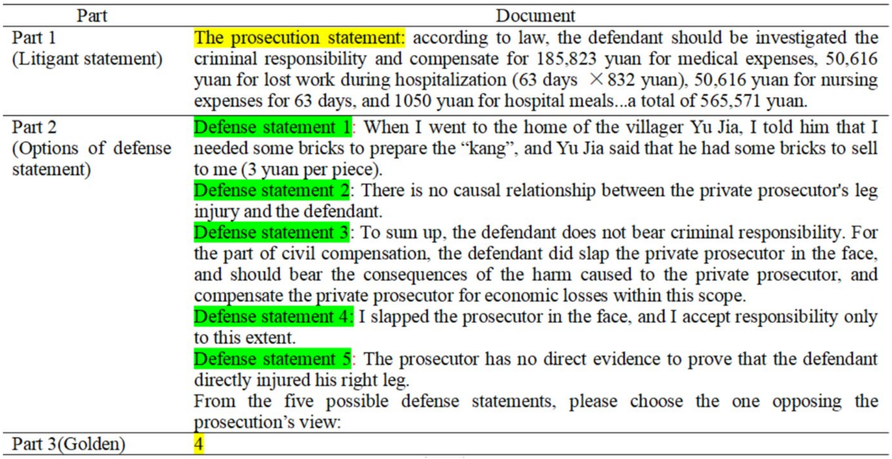
Fig. 4 An example of legal judgment prediction case structure

on datasets apart from China to confirm the generalizability of our evaluation results. An annotated corpus of 42 decisions of the European Court of Human Rights (ECHR) documents have been annotated for argument mining(Poudyal et al. 2020b). The corpus is annotated in terms of three types of clauses useful in argument mining: premise, conclusion, and non-argument parts of the text. The premise and conclusion belong to arguments. The goal of argument-focused mining on this dataset is to predict if a clause belongs to an argument or not.



Part	Document
Part 1 (Case description and question)	In the case of Sike Company v. Xiangguo Company loan, Sike Company was dissatisfied with the people's Court of first instance only deciding that Xiangguo Company should return the principal and filed an appeal ^{Description}. According to the above situation and Chinese law and general legal logic, how does the court of second instance handled it correctly^{Question} :
Part 2 (Options)	A: In the second instance, Sike applied for withdrawal of the suit, and the court of second instance found that there was a mistake in the judgment of the first instance, then the withdrawal of the suit should not be allowed. B: In the second instance, Sike reached a settlement agreement with the Fragrant Fruit Company, and the court of second instance should not allow it, but should proceed with the trial and make a judgment. C: In the second instance, Sike reached a settlement agreement with Xiangguo Company, and the court of second instance can review the settlement agreement reached by both parties and prepare a mediation document to serve on the parties according to the requirements of the parties. D: In the second instance, Sike reached a settlement with Xiangguo Company and applied for withdrawal of the lawsuit, and the review found that the withdrawal conditions were met, which could be granted. The correct answer is:
Part 3(Golden)	["A", "C"]

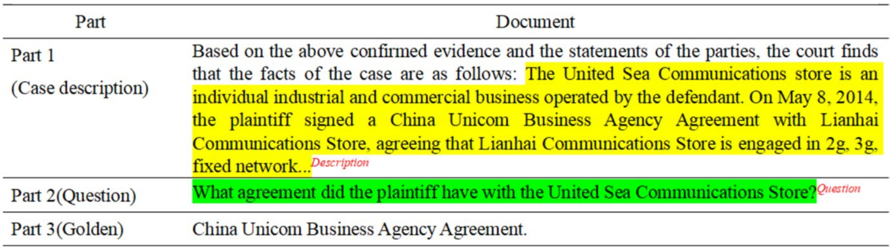
Fig. 5 An example of judicial examination case structure



Part	Document
Part 1 (Litigant statement)	The prosecution statement: according to law, the defendant should be investigated the criminal responsibility and compensate for 185,823 yuan for medical expenses, 50,616 yuan for lost work during hospitalization (63 days ×832 yuan), 50,616 yuan for nursing expenses for 63 days, and 1050 yuan for hospital meals...a total of 565,571 yuan.
Part 2 (Options of defense statement)	Defense statement 1: When I went to the home of the villager Yu Jia, I told him that I needed some bricks to prepare the "kang", and Yu Jia said that he had some bricks to sell to me (3 yuan per piece). Defense statement 2: There is no causal relationship between the private prosecutor's leg injury and the defendant. Defense statement 3: To sum up, the defendant does not bear criminal responsibility. For the part of civil compensation, the defendant did slap the private prosecutor in the face, and should bear the consequences of the harm caused to the private prosecutor, and compensate the private prosecutor for economic losses within this scope. Defense statement 4: I slapped the prosecutor in the face, and I accept responsibility only to this extent. Defense statement 5: The prosecutor has no direct evidence to prove that the defendant directly injured his right leg. From the five possible defense statements, please choose the one opposing the prosecution's view:
Part 3(Golden)	4

Fig. 6 An example of argument focus mining case structure

- *Judicial reading comprehension*. Extracting relevant content from the given judgment document to answer the corresponding question, the answer can take three forms: a fragment of the document, “yes" or “no", or a refusal to answer (unknown), which have been illustrated in our prompt. In addition, the answer may involve some syntactical transformation, semantic paraphrasing, or terminology not mentioned by the prompt, but we do not consider these due to the difficulty of evaluation. An example of the case structure in the subset of the CAIL2020 dataset for the judicial reading comprehension task is shown in Fig. 7.



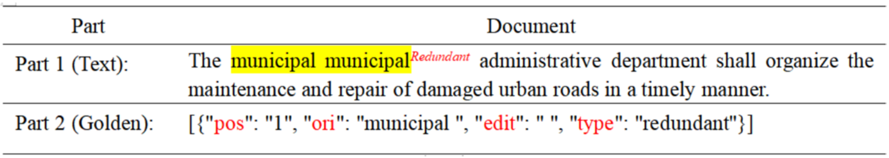
Part	Document
Part 1 (Case description)	Based on the above confirmed evidence and the statements of the parties, the court finds that the facts of the case are as follows: The United Sea Communications store is an individual industrial and commercial business operated by the defendant. On May 8, 2014, the plaintiff signed a China Unicom Business Agency Agreement with Lianhai Communications Store, agreeing that Lianhai Communications Store is engaged in 2g, 3g, fixed network...Description
Part 2(Question)	What agreement did the plaintiff have with the United Sea Communications Store?Question
Part 3(Golden)	China Unicom Business Agency Agreement.

Fig. 7 An example of judicial reading comprehension case structure

Table 4 Definitions of errors in legal texts to be proofread

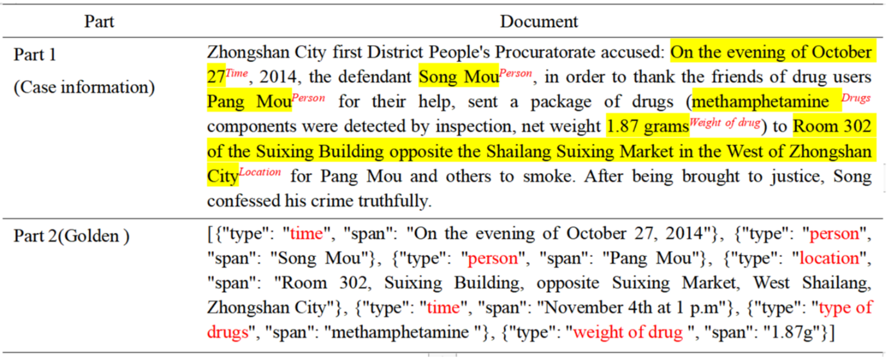
Error Categories	Definitions
Typos	Word is misspelled
Redundancies	The text covers explicit repetitions of words
Omissions	The absence of critical elements in a paragraph or sentence(only consider grammatical omissions)
Disorder	Words appear in the wrong place, resulting in semantic or grammatical errors in the sentence

- *Document proofreading*. Legal documents serve as a medium for judicial organs and citizens to exercise legal rights, necessitating utmost accuracy in their textual content. Document proofreading aims to automatically detect and correct errors in legal documents, such as typos, redundancies, omissions, and disorder. Here, we only define redundancy as the text that covers explicit repetitions of words. Omissions represent the absence of critical elements in a paragraph or sentence. Specifically, it can be a lack of complete grammatical information or a complete chain of evidence. However, given the complexity of the problem, we will only consider grammatical omissions here. Finally, we define disorder as words appearing in the wrong place, resulting in semantic or grammatical errors in the sentence. Definitions of these errors in legal texts are shown in Table 4. An example of the case structure in the subset of the CAIL2022 dataset for the document proofreading task is shown in Fig. 8.
- *Entity recognition*. Given a judgment document, extract entity information corresponding to pre-defined entity types such as plaintiffs, defendants, victims, criminal suspects, etc. Additionally, depending on the cause of action, entities may also include collusion, criminal tools, monetary amounts, related items, etc. In AI&Law, entities and their relationships are the vital elements of the legal ontology. The data sets we used in entity recognition, relation extraction, and event detection are all selected and verified by legal experts. Overlapping categories, supercategories, and subcategories cannot be processed here, because they are not included in the data set. An example of the case structure in the



Part	Document
Part 1 (Text):	The municipal municipal^{Redundant} administrative department shall organize the maintenance and repair of damaged urban roads in a timely manner.
Part 2 (Golden):	[{"pos": "1", "ori": "municipal ", "edit": " ", "type": "redundant"}]

Fig. 8 An example of document proofreading case structure

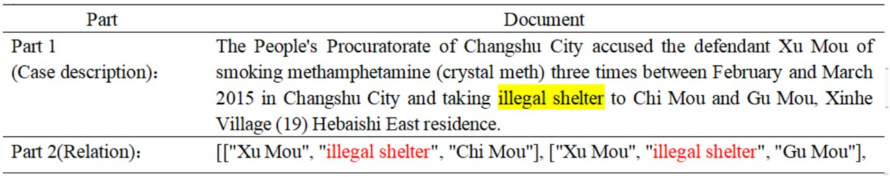


Part	Document
Part 1 (Case information)	Zhongshan City first District People's Procuratorate accused: On the evening of October 27^{Time}, 2014, the defendant Song Mou^{Person}, in order to thank the friends of drug users Pang Mou^{Person} for their help, sent a package of drugs (methamphetamine ^{Drugs} components were detected by inspection, net weight 1.87 grams$^{Weight\ of\ drug}$) to Room 302 of the Suixing Building opposite the Shailang Suixing Market in the West of Zhongshan City^{Location} for Pang Mou and others to smoke. After being brought to justice, Song confessed his crime truthfully.
Part 2(Golden)	[{"type": "time", "span": "On the evening of October 27, 2014"}, {"type": "person", "span": "Song Mou"}, {"type": "person", "span": "Pang Mou"}, {"type": "location", "span": "Room 302, Suixing Building, opposite Suixing Market, West Shailang, Zhongshan City"}, {"type": "time", "span": "November 4th at 1 p.m"}, {"type": "type of drugs", "span": "methamphetamine "}, {"type": "weight of drug ", "span": "1.87g"}]

Fig. 9 An example of document entity recognition case structure

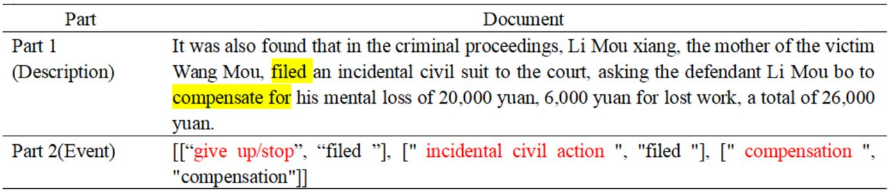
subset of the CAIL2022 datasets for document entity recognition is shown in Fig. 9.

- *Relation extraction*. Texts may imply triples consisting of three elements: a subject, an object, and the relationship between them. Relation extraction, also known as triple extraction, involves extracting these triples. Usually, each relation involves a distinct subject and object entity type. For example, in drug-related cases, entities include the person's name, place name, time, type, and weight of the drug. The relationships include selling (to sell the drug to a person), selling the drug, possession, and providing shelter (for someone). An example of the case structure in the subset of CAIL2022 for the relation extraction task is shown in Fig. 10.
- *Event detection*. The objective of the event detection task is to identify trigger words indicative of a particular type of event in a legal document. In legal documents, incidents are frequently associated with particular causes of action. These behavioral descriptions contain trigger words that determine the event category. An example of the case structure in the subset of CAIL2022 for the event detection task is shown in Fig. 11.



Part	Document
Part 1 (Case description):	The People's Procuratorate of Changshu City accused the defendant Xu Mou of smoking methamphetamine (crystal meth) three times between February and March 2015 in Changshu City and taking illegal shelter to Chi Mou and Gu Mou, Xinhe Village (19) Hebaishi East residence.
Part 2(Relation):	[["Xu Mou", "illegal shelter", "Chi Mou"], ["Xu Mou", "illegal shelter", "Gu Mou"],

Fig. 10 An example of relation extraction case structure

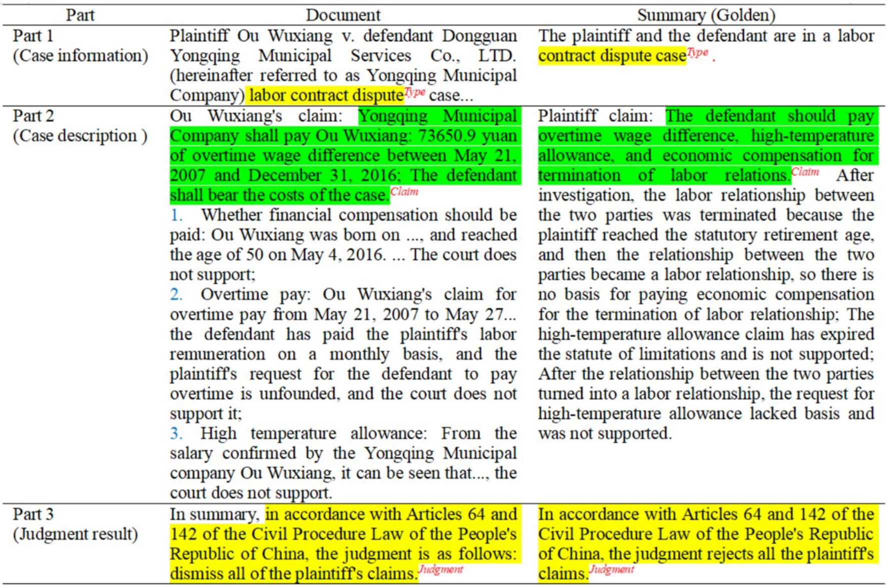


Part	Document
Part 1 (Description)	It was also found that in the criminal proceedings, Li Mou xiang, the mother of the victim Wang Mou, filed an incidental civil suit to the court, asking the defendant Li Mou bo to compensate for his mental loss of 20,000 yuan, 6,000 yuan for lost work, a total of 26,000 yuan.
Part 2(Event)	[["give up/stop", "filed "], [" incidental civil action ", "filed "], [" compensation ", "compensation"]]

Fig. 11 An example of event detection case structure

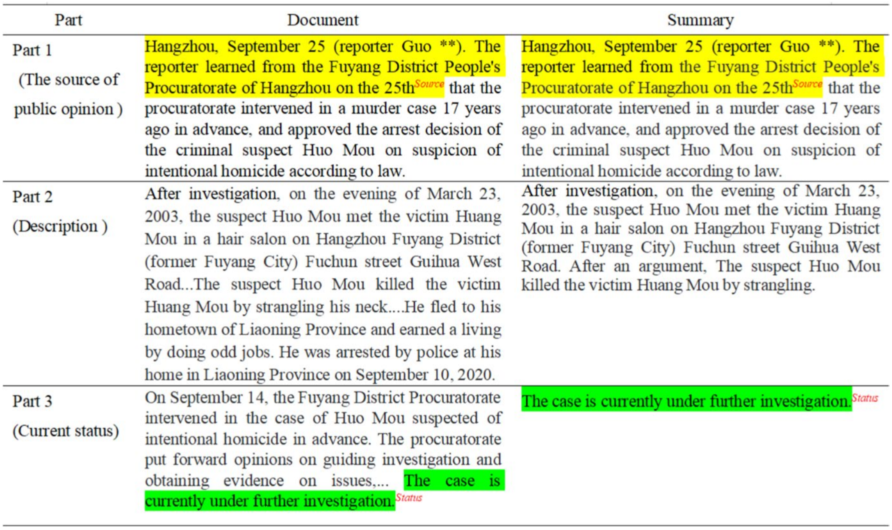
3.2.2 Generation

- *Judgment summarization*. As the concise version of a long text, the summary could be realized by NLP technologies without losing key information. Judgment documents are an important carrier to record the public trial activities, reasons, legal basis, and final judgment results. The generation of the legal judgment summary is to compress the content of legal documents, preserving vital details, including event elements, argument issues, relative articles, and the final results, etc. An example of the case structure in the subset of the CAIL2020 dataset for judgment summarization is shown in Fig. 12.
- *Public opinion summarization*. Legal public opinion encompasses public sentiments regarding legal issues. It involves analyzing public legal events to generate concise summaries of legal public opinion in Chinese. Like other summary tasks, it aims to preserve essential details of public sentiment. An example of the case structure in the subset of the CAIL2022 dataset for the public opinion summary task is shown in Fig. 13.
- *Court's opinion generation*.The task aims to produce textual judgment outcomes based on case facts, whose essence is a text generation task. Cases are categorized into criminal and civil types. Criminal judgments typically include crucial details such as relevant articles, charges, fines, and prison sentences. In civil cases, the judgment either upholds or dismisses the plaintiff's claim. Wu et al. (2020) built AC-NLG dataset containing more than 40,000 cases based on private lending. They split legal documents by keywords into three parts: description of cases, plaintiff's claims, and the final court's view. These datasets are used for the judicial examination and the court opinion generation, respectively. An example of the case structure in the AC-NLG dataset is shown in Fig. 14.



Part	Document	Summary (Golden)
Part 1 (Case information)	Plaintiff Ou Wuxiang v. defendant Dongguan Yongqing Municipal Services Co., LTD. (hereinafter referred to as Yongqing Municipal Company) labor contract dispute^{Type} case...	The plaintiff and the defendant are in a labor contract dispute case^{Type}.
Part 2 (Case description)	Ou Wuxiang's claim: Yongqing Municipal Company shall pay Ou Wuxiang: 73650.9 yuan of overtime wage difference between May 21, 2007 and December 31, 2016; The defendant shall bear the costs of the case.^{Claim} 1. Whether financial compensation should be paid: Ou Wuxiang was born on ...., and reached the age of 50 on May 4, 2016. ... The court does not support; 2. Overtime pay: Ou Wuxiang's claim for overtime pay from May 21, 2007 to May 27... the defendant has paid the plaintiff's labor remuneration on a monthly basis, and the plaintiff's request for the defendant to pay overtime is unfounded, and the court does not support it; 3. High temperature allowance: From the salary confirmed by the Yongqing Municipal company Ou Wuxiang, it can be seen that..., the court does not support.	Plaintiff claim: The defendant should pay overtime wage difference, high-temperature allowance, and economic compensation for termination of labor relations.^{Claim} After investigation, the labor relationship between the two parties was terminated because the plaintiff reached the statutory retirement age, and then the relationship between the two parties became a labor relationship, so there is no basis for paying economic compensation for the termination of labor relationship; The high-temperature allowance claim has expired the statute of limitations and is not supported; After the relationship between the two parties turned into a labor relationship, the request for high-temperature allowance lacked basis and was not supported.
Part 3 (Judgment result)	In summary, in accordance with Articles 64 and 142 of the Civil Procedure Law of the People's Republic of China, the judgment is as follows: dismiss all of the plaintiff's claims.^{Judgment}	In accordance with Articles 64 and 142 of the Civil Procedure Law of the People's Republic of China, the judgment rejects all the plaintiff's claims.^{Judgment}

Fig. 12 An example of judgment summarization structure



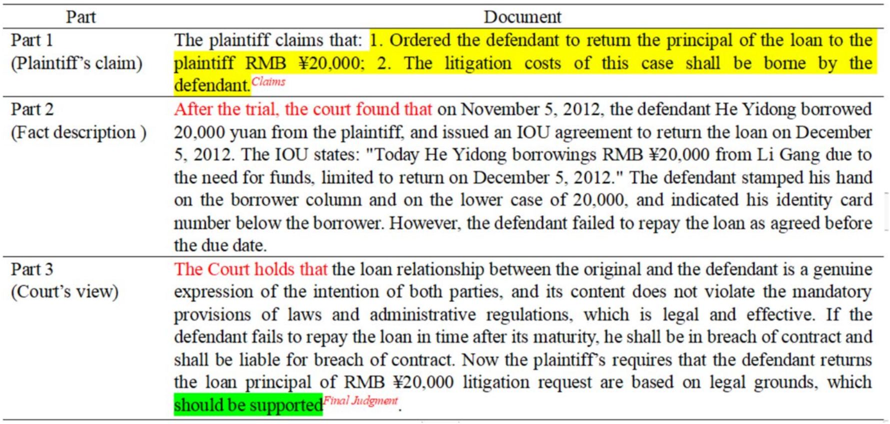
Part	Document	Summary
Part 1 (The source of public opinion)	Hangzhou, September 25 (reporter Guo **). The reporter learned from the Fuyang District People's Procuratorate of Hangzhou on the 25th^{Source} that the procuratorate intervened in a murder case 17 years ago in advance, and approved the arrest decision of the criminal suspect Huo Mou on suspicion of intentional homicide according to law.	Hangzhou, September 25 (reporter Guo **). The reporter learned from the Fuyang District People's Procuratorate of Hangzhou on the 25th^{Source} that the procuratorate intervened in a murder case 17 years ago in advance, and approved the arrest decision of the criminal suspect Huo Mou on suspicion of intentional homicide according to law.
Part 2 (Description)	After investigation, on the evening of March 23, 2003, the suspect Huo Mou met the victim Huang Mou in a hair salon on Hangzhou Fuyang District (former Fuyang City) Fuchun street Guihua West Road...The suspect Huo Mou killed the victim Huang Mou by strangling his neck....He fled to his hometown of Liaoning Province and earned a living by doing odd jobs. He was arrested by police at his home in Liaoning Province on September 10, 2020.	After investigation, on the evening of March 23, 2003, the suspect Huo Mou met the victim Huang Mou in a hair salon on Hangzhou Fuyang District (former Fuyang City) Fuchun street Guihua West Road. After an argument, The suspect Huo Mou killed the victim Huang Mou by strangling.
Part 3 (Current status)	On September 14, the Fuyang District Procuratorate intervened in the case of Huo Mou suspected of intentional homicide in advance. The procuratorate put forward opinions on guiding investigation and obtaining evidence on issues,... The case is currently under further investigation.^{Status}	The case is currently under further investigation.^{Status}

Fig. 13 An example of public opinion summary structure

3.3 Metric

We employ several metrics to measure the results generated by LLMs on different tasks. Additionally, human evaluation is used to discover the problems of automated evaluation. They are displayed as follows:

- **Accuracy:** Accuracy performs an exact match between the model prediction and the gold answer. It means the proportion of correctly predicted samples over the total samples.
- **F1**: Both precision and recall are considered so that they can reach the highest level at the same time to achieve a balance.
- **Distance**: This function computes the logarithm of the difference between the predicted term of imprisonment and the gold answer, then normalize it to the space between 0 and 1.
- **Match**: It is used to calculate the matching degree of two texts. If they match exactly, the result is 1, otherwise, it is 0.
- **ROUGE-1**: The most common evaluation metric for generation tasks is ROUGE(Recall-oriented Understudy for Gisting Evaluation), which is mainly based on recall(Lin 2004). ROUGE-1 is used to measure how well the summary generated by the system matches the words contained in the reference summary.
- **ROUGE-2**: It is consistent with the basic principle of ROUGE-1, but the comparison object is expanded to two adjacent words from the generated summary and the reference one.
- **ROUGE-L**: ROUGE-L considers sentence-level structure similarity naturally and automatically identifies the longest co-occurring sequence n-grams to compare the extracted and gold answers.
- **Human Evaluation**: We conduct a human evaluation to discover the problems of the automated evaluation by analyzing the results generated by different LLMs, and given the large scale of these results, we simplified the process. For different legal tasks, we sample some cases and present the corresponding results(generated by different LLMs) to 30 professionals with legal backgrounds. These professionals include faculty members from law schools, as well as doc-



Part	Document
Part 1 (Plaintiff's claim)	The plaintiff claims that: 1. Ordered the defendant to return the principal of the loan to the plaintiff RMB ¥20,000; 2. The litigation costs of this case shall be borne by the defendant.$^{\text{Claims}}$
Part 2 (Fact description)	After the trial, the court found that on November 5, 2012, the defendant He Yidong borrowed 20,000 yuan from the plaintiff, and issued an IOU agreement to return the loan on December 5, 2012. The IOU states: "Today He Yidong borrowings RMB ¥20,000 from Li Gang due to the need for funds, limited to return on December 5, 2012." The defendant stamped his hand on the borrower column and on the lower case of 20,000, and indicated his identity card number below the borrower. However, the defendant failed to repay the loan as agreed before the due date.
Part 3 (Court's view)	The Court holds that the loan relationship between the original and the defendant is a genuine expression of the intention of both parties, and its content does not violate the mandatory provisions of laws and administrative regulations, which is legal and effective. If the defendant fails to repay the loan in time after its maturity, he shall be in breach of contract and shall be liable for breach of contract. Now the plaintiff's requires that the defendant returns the loan principal of RMB ¥20,000 litigation request are based on legal grounds, which should be supported$^{\text{Final Judgment}}$.

Fig. 14 An example of the court's opinion generation structure

Table 5 The legal tasks, corresponding datasets and metrics

Type	Task	Source	Metric
Understanding	Legal Judgment Prediction	CAIL2018	Accuracy,F1
	Judicial Examination	JEC-QA	Exact Match
	Argument Focus Mining	CAIL2020,ECHR	Accuracy
	Judicial Reading Comprehension	CAIL2020	Match
	Entity Recognition	CAIL2022	F1
	Relation Extraction	CAIL2022	
	Event Detection	CAIL2022	
	Document Proofreading	CAIL2022	F1
Generation	Judgment Summarization	CAIL2020	ROUGE-1, ROUGE-2, ROUGE-L
	Public Opinion Summarization	CAIL2022	
	Court Opinion Generation	AC-NLG	

toral and master's students who have passed China's Unified Qualification Exam for Legal Professionals. It is worth noting that some results can be directly judged right or wrong according to the standard(golden) answers in objective legal tasks, such as legal judgment prediction, judicial examination, argument focus mining (multiple choice question), etc. However, some results are paragraphs of text that cannot be merely judged by comparison with the golden answer, such as judgment summarization, public opinion summarization, court's opinion generation tasks, etc., and we can conduct human evaluations following four perspectives: correctness, integrality, fluency, and redundancy.

4 Experimental results and discussion

4.1 Experiment setting

This section presents and analyzes the experimental results of various legal task categories. All results are rounded to two decimal places. Table 5 displays all selected legal tasks, their corresponding experimental datasets, and metrics. We use the models' default parameters(temperature and top) for all text-generating tasks. There are no restrictions on the length of the input and output text. The zero-shot prompts for all tasks displayed in Figs. 1, 2, and 3. Examples taken from the best-performing models for each task are shown in Appendix A.

4.2 Understanding

- Judgment Prediction. The goal of judgment prediction is to predict relative law articles, charges, and terms of prison, which we evaluated respectively. The results are displayed in Table 6. WisdomInterrogatory demonstrates the highest

F1 in relevant article prediction and FuziMingcha has superior performance in charge prediction. We classify the term of prison into three categories: fixed-term imprisonment, life imprisonment, and death penalty. Firstly, we deal with this as a classification problem and calculate the accuracy for each model. Then we used the *Distance* function to calculate the difference between the predicted result and the golden answer in fixed-term imprisonment cases. Finally, the score for each model in prison term prediction is calculated as follows:

$$Score = Accuracy * 0.5 + Distance * 0.5 \tag{1}$$

The reason why we set the coefficient to 0.5 is that these two parts are equally important. The results are shown in Table 6, and Baichuan2(7b) has the highest score.

- Judicial examination. The result of this task is shown in Table 6 as well. We can see that Internlm2 has shown the best performance in this task, and GPT-4 follows it. Unfortunately, they all fail to achieve satisfactory results with scores of only 0.13 and 0.12. We hypothesize that the LLMs do not have an excellent ability to resolve multiple-choice questions.

- Reading comprehension. To simplify the comprehension task, the responses are restricted to three forms: "yes or no," "unknown," or verbatim text excerpts. The results are displayed in Table 7. In reading comprehension tasks, the results show that GPT-4 and ChatGPT all have high scores with 0.83 and 0.71, respectively.
- Argument focus mining. The evaluation results of different LLMs on the CAIL2020 and ECHR datasets are shown in Table 7 and Table 8, respectively. Although we used different metrics on the two datasets, the final results are the same: GPT-4 was the best model for this task, but not with ideal metric scores, only with 0.40(accuracy) and 0.51(F1), respectively. We analyzed the results generated by GPT-4 on specific cases of CAIL2020 and found some possible

Table 6 Performance of different LLMs on judgment prediction and judicial examination

Models	Judgment prediction			Judicial examination (Exact match)
	Relative articles (F1)	Charges (Accuracy)	Term (Score)	
GPT-4	0.33	0.67	0.62	0.12
ChatGPT	0.49	0.69	0.81	0.09
GLM2(6b)	0.16	0.51	0.50	0.11
Baichuan2(7b)	0.37	0.80	0.91	0.05
Qwen	0.44	0.53	0.24	0.09
Internlm2	0.47	0.68	0.87	0.13
WisdomInterrogatory	0.68	0.69	0.64	0.11
FuziMingcha	0.38	0.83	0.72	0.03

Table 7 Performance of different LLMs on reading comprehension, argument mining, and document proofreading tasks

Models	Reading comprehension (Match)	Argument mining (Accuracy)	Document proofreading (F1)
GPT-4	0.83	0.40	0.59
ChatGPT	0.71	0.26	0.51
GLM2(6b)	0.55	0.21	0.14
Baichuan2(7b)	0.66	0.29	0.08
Qwen	0.66	0.23	0.29
Internlm2	0.64	0.21	0.43
WisdomInterrogatory	0.40	0.20	0.38
FuziMingcha	0.57	0.23	0.24

Table 8 The argument mining results on ECHR dataset

Models	Argument mining		
	Precision	Recall	F1
GPT-4	0.39	0.82	0.51
ChatGPT	0.31	0.64	0.41
GLM2(6b)	0.31	0.93	0.45
Baichuan2(7b)	0.25	0.92	0.35
Qwen	0.27	0.96	0.41
Internlm2	0.34	0.75	0.46
WisdomInterrogatory	0.25	0.77	0.37
FuziMingcha	0.25	0.83	0.38

causes. Take the case shown in Fig. 6 for example, the golden answer is defense statement 4, but the answer generated by GPT-4 is defense statement 2. We compare these options and find that they are very similar, but Defense Statement 2 is much shorter. We hypothesize that the GPT-4 handles this task much more simply and intuitively. This may be influenced by the training datasets as well.

- Document proofreading. The model that performed best in this task was GPT-4 but with a low value of F1. Take the case in Fig. 8 as an example, 'pos' generated by GPT-4 is '0' but the golden answer is '1', meaning that the positioning of GPT-4 is not accurate enough in this task.
- Information extraction. The subtasks for information extraction involve entity recognition, relation extraction, and event detection. The experimental results are shown in Table 9. We can see that ChatGPT has the highest score in entity recognition. Different results between LLMs may be due to the different understanding of the scope of entities. For relation extraction, the best model is Qwen. GPT-4 has the best performance in event detection.

4.3 Generation

In this section, we test all selected LLMs on three legal generation tasks involving judgment summarization, public opinion summarization, and court opinion. The evaluation metric ROUGE Score(RS) can be designed as:

$$RS = ROUGE_1 * 0.2 + ROUGE_2 * 0.4 + ROUGE_L * 0.4 \tag{2}$$

Where $ROUGE_1$, $ROUGE_2$, and $ROUGE_L$ stand for ROUGE-1, ROUGE-2,and ROUGE-L respectively.The overall zero-shot performance of each model is shown in Table 10. Table 10 presents the scores of different LLMs on legal generation tasks in zero-shot situations, demonstrating the differences in executing different tasks. The experimental results we observed are described as follows:

Firstly, FuziMingcha has the best performance in public opinion summarization and owns the highest average score on the generative tasks. Then, Chinese-oriented general LLMs also show good performance: Qwen has the highest score in judgment summarization, and Baichuan2(7b) performs best in the court's opinion generation. Finally, the performance of most models on the court's opinion generation task is inferior to the other summary tasks (as seen in Table 10), which may be due to the character of the task itself. Court's opinion generation requires making a correct judgment on the case, deepening the difficulty of this work.

Overall, their accuracy and reliability are critical due to the particularity of legal text, which is a challenge to LLMs. From the experimental results, the ability of LLMs in legal text generation needs to be improved, and people can not completely rely on the legal text generated by LLMs. In addition, although we use ROUGE for calculation and evaluation, the test work is still insufficient, and some details should be considered, such as the completeness of the generated text and the fluency of the language.

Table 9 Performance of different LLMs on information extraction tasks

Models	Information extraction (F1)		
	Entity recognition	Relation extraction	Event detection
GPT-4	0.62	0.48	0.58
ChatGPT	0.70	0.32	0.51
GLM2(6b)	0.30	0.23	0.13
Baichuan2(7b)	0.44	0.00	0.30
Qwen	0.61	0.76	0.34
Internlm2	0.67	0.00	0.39
WisdomInterrogatory	0.16	0.16	0.16
FuziMingcha	0.16	0.05	0.12

Table 10 Scores(%) of different LLMs on legal generation tasks

Models	Judgment summarization	Public opinion summarization	Court's opinion	Average
GPT-4	22.31	19.66	18.36	20.11
ChatGPT	23.71	25.64	20.73	23.36
GLM2(6b)	27.33	28.58	18.51	24.80
Baichuan2(7b)	26.45	26.48	23.24	25.39
Qwen	39.73	28.35	22.88	30.32
Internlm2	26.98	23.78	19.05	23.27
WisdomInterrogatory	26.62	20.26	20.78	22.55
FuziMingcha	24.61	57.31	16.74	32.88

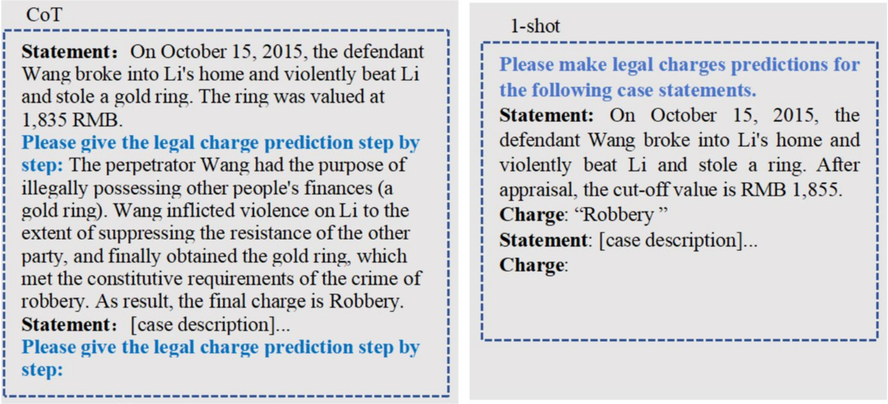
4.4 Chain-of-Thought prompting and few-shot techniques

Chain-of-Thought(CoT) prompting improves the ability of LLMs to perform complex reasoning by a series of intermediate reasoning steps (Wei et al. 2023). According to whether there are manual demonstrations, current CoT prompting methods can be divided into two main categories: manual-CoT and zero-Shot-CoT (Qin et al. 2023). In this paper, we only perform CoT reasoning with manually designed demonstrations(manual-CoT) on charge prediction task, a subtask of legal judgment prediction. We guide the model in producing results by giving the reasoning process of a case step by step. Few-shot learning can rapidly generalize to new tasks containing only a few samples with supervised information, has emerged as an effective learning method, and shows great potential(Wang et al. 2020b). We designed CoT and 1-shot prompts as shown in Fig. 15. The results are present in Table 11, and the GPT-4 performs best. The performance of legal-specific LLMs is not outstanding, which may be attributed to the fact that they have not been trained on the CoT or 1-shot cases. However, all the best results of LLMs in CoT and 1-shot can not be comparable to the performance of FuziMingcha in 0-shot charge prediction shown in Table 6.

4.5 Human evaluation

According to section 3.4, the results generated by the model can be divided into two main categories: some can be objectively compared to standard answers(correct if matching, otherwise incorrect) and the others require subjective judgment. We found that automated evaluation aligns with human evaluation in objective comparisons but encounters issues in subjective judgments. However, when it comes to subjective judgment, some typical problems emerge. Then, these issues are detailed below:

- *Correctness*. Automated evaluation may deem some results incorrect, while human evaluation considers them right. For example, a result generated by



CoT

Statement: On October 15, 2015, the defendant Wang broke into Li's home and violently beat Li and stole a gold ring. The ring was valued at 1,835 RMB.
Please give the legal charge prediction step by step: The perpetrator Wang had the purpose of illegally possessing other people's finances (a gold ring). Wang inflicted violence on Li to the extent of suppressing the resistance of the other party, and finally obtained the gold ring, which met the constitutive requirements of the crime of robbery. As result, the final charge is Robbery.
Statement: [case description]...
Please give the legal charge prediction step by step:

1-shot

Please make legal charges predictions for the following case statements.
Statement: On October 15, 2015, the defendant Wang broke into Li's home and violently beat Li and stole a ring. After appraisal, the cut-off value is RMB 1,855.
Charge: "Robbery "
Statement: [case description]...
Charge:

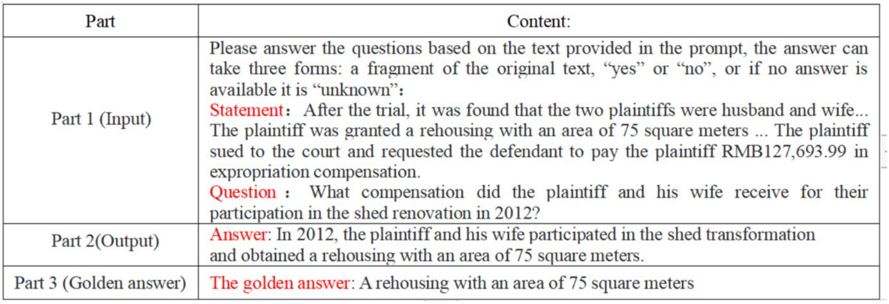
Fig. 15 An illustration of CoT and 1-shot prompts for the legal charge prediction task. Instructions are colored in blue, and the model will generate the prediction after reading the entire input (circled by the blue dashed box)

Table 11 The performance of LLMs on legal charge prediction with CoT and 1-shot prompt(Accuracy)

LLMs	CoT	1-shot
GPT-4	0.80	0.76
ChatGPT	0.70	0.69
GLM2(6b)	0.63	0.43
Baichuan2(7b)	0.53	0.52
Qwen	0.79	0.29
Internlm2	0.80	0.53
WisdomInterrogatory	0.68	0.44
FuziMingcha	0.59	0.54

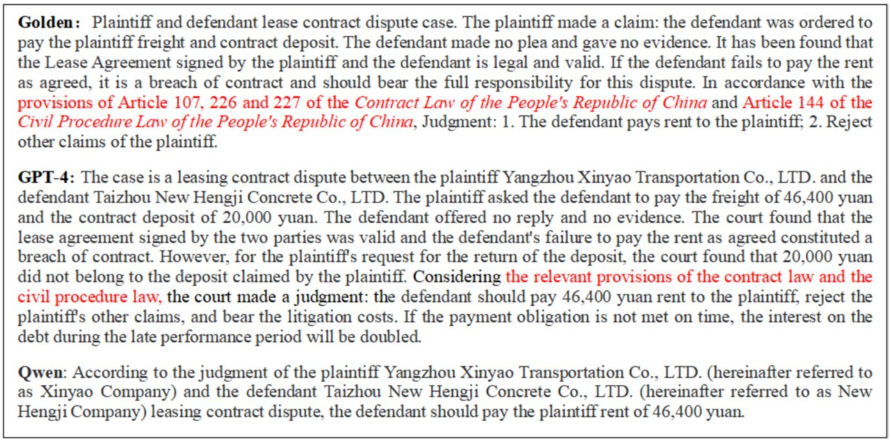
ChatGLM2 in the reading comprehension task is shown in Fig. 16. After reading the legal statement, the question is "What compensation did the plaintiff and his wife receive for their participation in the shed renovation in 2012?" The standard(golden) answer is "A rehousing with an area of 75 square meters". The answer generated by ChatGLM2 is "In 2012, the plaintiff and his wife participated in the shed transformation and obtained a rehousing with an area of 75 square meters", which should be judged correct. However, the automated evaluation concluded that the model output was wrong, potentially due to evaluation metrics focusing on overlapping words or phrases. This is also a relatively common problem when measuring the similarity between texts.

- *Integrality*. In judgment summarization, models sometimes omit crucial legal terms affecting the judgment's integrity. GPT-4 and Qwen's summarizations illustrate this problem as shown in Fig. 17. Additionally, some models oversimplify judgments, leading to poor integrity, as seen in Qwen's summarization in Fig. 17.



Part	Content:
Part 1 (Input)	Please answer the questions based on the text provided in the prompt, the answer can take three forms: a fragment of the original text, “yes” or “no”, or if no answer is available it is “unknown”: Statement：After the trial, it was found that the two plaintiffs were husband and wife... The plaintiff was granted a rehousing with an area of 75 square meters ... The plaintiff sued to the court and requested the defendant to pay the plaintiff RMB127,693.99 in expropriation compensation. Question ： What compensation did the plaintiff and his wife receive for their participation in the shed renovation in 2012?
Part 2(Output)	Answer: In 2012, the plaintiff and his wife participated in the shed transformation and obtained a rehousing with an area of 75 square meters.
Part 3 (Golden answer)	The golden answer: A rehousing with an area of 75 square meters

Fig. 16 An illustration of incorrectness resulting from the reading comprehension task



Golden： Plaintiff and defendant lease contract dispute case. The plaintiff made a claim: the defendant was ordered to pay the plaintiff freight and contract deposit. The defendant made no plea and gave no evidence. It has been found that the Lease Agreement signed by the plaintiff and the defendant is legal and valid. If the defendant fails to pay the rent as agreed, it is a breach of contract and should bear the full responsibility for this dispute. In accordance with the provisions of Article 107, 226 and 227 of the *Contract Law of the People's Republic of China* and Article 144 of the *Civil Procedure Law of the People's Republic of China*, Judgment: 1. The defendant pays rent to the plaintiff; 2. Reject other claims of the plaintiff.

GPT-4: The case is a leasing contract dispute between the plaintiff Yangzhou Xinyao Transportation Co., LTD. and the defendant Taizhou New Hengji Concrete Co., LTD. The plaintiff asked the defendant to pay the freight of 46,400 yuan and the contract deposit of 20,000 yuan. The defendant offered no reply and no evidence. The court found that the lease agreement signed by the two parties was valid and the defendant's failure to pay the rent as agreed constituted a breach of contract. However, for the plaintiff's request for the return of the deposit, the court found that 20,000 yuan did not belong to the deposit claimed by the plaintiff. Considering the relevant provisions of the contract law and the civil procedure law, the court made a judgment: the defendant should pay 46,400 yuan rent to the plaintiff, reject the plaintiff's other claims, and bear the litigation costs. If the payment obligation is not met on time, the interest on the debt during the late performance period will be doubled.

Qwen: According to the judgment of the plaintiff Yangzhou Xinyao Transportation Co., LTD. (hereinafter referred to as Xinyao Company) and the defendant Taizhou New Hengji Concrete Co., LTD. (hereinafter referred to as New Hengji Company) leasing contract dispute, the defendant should pay the plaintiff rent of 46,400 yuan.

Fig. 17 Judgment summarizations generated by GPT-4 and Qwen

- *Redundancy*. In generative tasks, models’ outputs may be longer than golden answers, often containing redundant information. This problem can be evident in GPT-4’s summarization results shown in Fig. 17.
- *Fluency*. All models can produce fluent text, and we haven’t found any mistakes at present.

All of the above are problems we found through human evaluation. Although there are some problems with automated evaluation, they do not affect the reliability of automated evaluation results.

4.6 Comparison and discussion

4.6.1 Comparison of general LLMs and legal-specific LLMs

We compare the performance of the general LLMs with legal-specific LLMs on various legal tasks. The results show that the performance of the general models (GPT-4 and ChatGPT) in most legal tasks is better than that of the legal-specific LLMs. However, legal-specific LLMs outperform others on some special tasks, such as judgment prediction. We describe the results and possible reasons as follows:

1. *GPT-4 is still the best model.* GPT-4 achieves the best performances in many legal tasks, and this result is consistent with the other LLMs evaluation researches(Kalyan 2024; Fei et al. 2023). As we all know, GPT-4 has trillions of parameters and is trained on more diverse data, enabling it to excel in various scenarios. Although ChatGPT performs well, it is inferior to GPT-4 and even to certain Chinese-oriented general models in some special legal tasks. This could be attributed to Chinese pre-training and fine-tuning of Chinese-oriented general LLMs, enhancing the model's proficiency in processing Chinese text.
2. *Legal-specific LLMs perform well.* In judgment prediction tasks, both WisdomInterrogatory and FuziMingcha demonstrate excellent performance. Specifically, WisdomInterrogatory outperformed all other models in related article prediction. This phenomenon is primarily attributed to the secondary pre-training and fine-tuning of legal-specific LLMs on legal knowledge. Pre-training intends to integrate legal knowledge into the general LLMs, encompassing legal documents, judicial cases, legal question-answer data, and more. Notably, in certain tasks, such as court's opinion summarization, the performance of legal-specific LLMs falls short compared to Chinese-oriented general models. We suspect that training on legal-specific datasets may somewhat restrict the language processing capabilities of legal-specific LLMs, and they are sensitive to the design of prompts.
3. *Chinese-oriented general LLMs need to be improved.* Chinese-oriented general LLMs are trained on Chinese corpora or fine-tuned using Chinese datasets. They exhibit two simultaneous characteristics: (1) Compared to GPT-4, their smaller parameter scale limits their capabilities in legal tasks; (2) They are not trained or infused with legal knowledge, so they are inferior to legal-specific LLMs. However, while Chinese-oriented general LLMs appear to excel at some legal tasks, there are some factors not taken into account that could influence the experimental results, such as the amount of data for the test, the structures of the prompt, etc.

In summary, though most LLMs show some capability in handling legal tasks, they have low scores in many tasks, even the top-performing model GPT-4. The experimental results indicate that (1) The training of LLMs is not enough; (2) The quantity and quality of the knowledge infusion are not high enough, so the expected goal is not reached in many tasks; (3) Further analysis is needed for the underlying reasons. Additionally, our research is still very preliminary, because prompts, the dataset, large model parameters, and other issues will affect the experiment results. In the future, we will improve the experiment to promote the application of LLMs in the legal field with a more scientific evaluation reference.

4.6.2 Common flaws

During our experiments, we noticed that the results were unsatisfactory across various tasks and did not meet the requirements of real-world legal work. We identified that both legal-specific and general large models had issues with hallucinations, confusing terms, catastrophic forgetting, and ethical issues.

1. *Hallucination*. In our experiments, we often confront a phenomenon called hallucination, where models generate fictitious or misleading content based on inadequate or incorrect information. This issue of hallucination is notably more acute in LLMs, presenting a significant risk in the application of such models for legal tasks that demand high levels of precision, coherence, and interpretability. As a result, the hallucinatory content produced by large models in legal tasks necessitates thorough examination and validation by legal experts to ensure the reliability and accuracy of the provided predictive results. Some research has found approaches to eliminate the issue of hallucination, such as adding the augmented module to LLM(Savelka et al. 2023b).
2. *Confusing terms*. There are many confusing charges and articles in legal scenarios. For example, the charge descriptions of insulting and defamation both involve damage to the reputation of others, but the difference is that the crime of insult can use real facts and violence, while the crime of libel uses fabricated facts without violence. This phenomenon will degrade the performance of the LLMs on some specific legal tasks, such as judgment prediction.
3. *Catastrophic forgetting*. In our experiments, we also encountered "Catastrophic Forgetting," which refers to the phenomenon where a model loses knowledge of previously learned tasks when it is trained on new tasks in a continuous learning setting. This issue becomes apparent in large model evaluations on legal tasks, where models lose their grasp on legal knowledge or cases they had learned before, once trained with new data about specific legal cases. While the model may show enhanced performance in that specific domain, it could simultaneously forget critical knowledge from other legal areas, like criminal law or contract law.
4. *Ethical issues*. We employed the publicly available CAIL datasets, widely used across numerous experimental studies. While we ensure data protection and privacy of individuals in legal documents in our experiments, many current experimental studies involve data that contain personal information, exposing them to significant risks of privacy breaches. Consequently, researchers must ensure that their generated texts do not reveal personal privacy details and that the outcomes are devoid of biases related to gender, age, race, and other factors.

4.6.3 Limitations.

This work is a study on the zero-shot learning ability of LLMs on representative legal tasks having several limitations.

1. In this paper, we only test limited models, but more models with larger scales should be tested. Besides, the one-shot learning ability of models should be tested in all legal tasks in further research. Finally, we only show the performance of different LLMs on legal tasks without providing solutions.
2. In the current experimental setup, we focused on a limited selection of tasks within the understanding and generation categories, not covering the full spectrum of legal tasks. Consequently, this limitation prevented a comprehensive comparison of the capabilities between specialized and general LLMs.
3. We employ primary evaluation metrics such as Accuracy, F1, Distance, and ROUGE. However, legal professionals also value the relevance and completeness of the responses provided by LLMs. For example, Logicians have developed logical criteria for evaluating legal tasks, extending beyond the standard of validity to include relevance, sufficiency, and acceptability (RSA triangle)(Blair 2012).
4. The performance of legal-specific LLMs heavily depends on the general LLMs they are based on. In our investigation, we developed these specialized LLMs using Chinese general LLMs as their base. Since ChatGPT−4.0 outperforms these base models in tasks like information extraction, it also surpasses the legal-specific LLMs in most tasks. This connection implies that the performance of legal-specific LLMs in our study is fundamentally linked to how well the Chinese general LLMs perform, highlighting the critical role of the base models in shaping the success of their specialized counterparts.
5. The evaluation of LJP is insufficient. We only perform single-charge predictions, not considering multi-charge cases, which may artificially boost LLM performance and narrow real-world applicability.
6. The results generated by LLMs are not flexible. To simplify the reading comprehension task, the responses are restricted to three forms: "yes or no," "unknown," or verbatim text excerpts, excluding paraphrased or inferential responses, which are not completely consistent with practical scenarios. This also limits our realization of the ability of LLMs to solve reading comprehension problems.
7. The absence of annotation regarding argument structure. The data sets selected for the argument focus mining task lack annotation of attack types, such as undermining, rebuttal, and undercutting, which is a weakness. Given the importance of argument mining in legal AI, this task needs further improvements.
8. The need of future ontology refinement. In AI&law, entities and their relationships are elements of the legal ontology. Some legal tasks, such as entity recognition, relation extraction, and event detection, should presuppose a rich ontology of legal entities and the relationships between them. However, these relationships, such as overlapping categories, supercategories, and subcategories, are not included in our data set, which are important and necessary to be identified in legal practice. For future ontology refinement, some data sets that include a rich ontology of legal entities and the relationships should be constructed.

5 Conclusion

LLMs have demonstrated significant advantages in NLP tasks, and evaluating them in specific domains is crucial in LLM research. This paper assesses the zero-shot learning abilities of different LLMs across various legal tasks. We chose representative models from both general and legal-specific categories. This paper's notable aspect is the categorization of legal tasks into understanding and generating, followed by testing the model across various dimensions and levels. The experimental results show that (1) GPT-4 takes the lead on most legal tasks, and the Chinese-oriented LLMs and the legal-specific LLMs perform well on some particular legal tasks; (2) The model size, training data, and fine-tuning contribute to the performance of LLMs on legal tasks; (3) The current LLMs are unable to give dependable legal aids, their scores on most tasks are poor. In further work, we will improve the structure of prompts. The researchers have demonstrated that the performance of the LLMs can be effectively improved by changing the structure of the prompts(Yu et al. 2023). For example, the reasoning ability of the model as well as its interpretability can be enhanced by using the chain-of-thought (CoT) prompts(Wei et al. 2023). In the legal field, we can also build CoT prompts for legal tasks. We can incorporate some designed demonstrations into legal prompts or split a complex task into multiple simple subtasks through multiple prompt processes. We intend to broaden our exploration by incorporating experiments on legal reasoning and legal Q&A, to delve deeper into the LLMs' capacity for logical reasoning and their proficiency in handling knowledge-intensive Q&A within the legal context. Furthermore, to better mirror real-world evaluation, we will incorporate subjective standards similar to those used by legal professionals to evaluate tasks in practice. These include: (1) Correctness: assessing whether the generated content is accurate; (2) Fluency: evaluating if the sentences in the legal text are smooth and clear; (3) Completeness: determining whether the generated content covers all essential information in the document.

Appendix

Details of task instruction

See Tables 12, 13, 14, 15, 16, 17, 18, 19, 20, 21 and 22.

Table 12 An example of judgment prediction task

Instruction: Please make legal judgment predictions for the following case statements. You should predict the charges, relative law articles, and the term of the penalty.

Input: Shantou Chaonan District People's Procuratorate alleged that at 16:00 on July 10, 2018, the defendant Zheng Mouqiu was arrested by police on duty when he drove a small car after drinking and passed through Xinglong, Longtian Town, Chensha Road, Chaonan District, Shantou. After testing, the alcohol content in Zheng's blood was 297 mg/100 ml.

Answer: "charge":["dangerous driving"], "relevant articles":[133], "Term of penalty":
"36 months"

Table 13 An example of judicial examination task

Instruction: Please give the correct answers to the following questions.

Input: In the case of Sike Company v. Xiangguo Company loan, Sike Company was dissatisfied with the People's Court of First Instance only deciding that Xiangguo Company should return the principal and filed an appeal. According to the above situation, how does the court of the second instance handle it correctly:

A: In the second instance, Sike applied for withdrawal of the suit, and the court of the second instance found that there was a mistake in the judgment of the first instance, then the withdrawal of the suit should not be allowed.

B: In the second instance, Siko reached a settlement agreement with the Fragrant Fruit Company, and the court of the second instance should not allow it but should proceed with the trial and make a judgment.

C: In the second instance, SiKO reached a settlement agreement with Xiangguo Company, and the court of the second instance can review the settlement agreement reached by both parties and prepare a mediation document to serve the parties according to the requirements of the parties.

D: In the second instance, Sike reached a settlement with Xiangguo Company and applied for withdrawal of the lawsuit, and the review found that the withdrawal conditions were met, which could be granted.Question: According to Chinese law and general legal logic, the correct answer is:

Answer: C

Table 14 An example of the reading comprehension

Instruction: Please answer the questions according to the following legal documents. The answers to the questions are displayed as follows: 1. If the answer is a continuous segment of the original text, please output the segment directly; 2. If the answer to the question is yes/no, please output "Yes" or "NO": 3. If the question cannot be answered, please output "UKNOWN" directly.

Input: The legal text: Based on the above-confirmed evidence and the statements of the parties, the court finds that the facts of the case are as follows: The United Sea Communications store is an individual industrial and commercial business operated by the defendant. On May 8, 2014, the plaintiff signed a "China Unicom Business Agency Agreement" with Lianhai Communications Store, agreeing that Lianhai Communications Store is engaged in 2 g, 3 g, fixed network...Question: What agreement did the plaintiff have with the United Sea Communications Store?

Answer: China Unicom Business Agency Agreement.

Table 15 An example of argument focus mining task

Instruction: According to the following statement of the prosecution, please select the opposite of the prosecution’s view from the five candidate defense statements provided. That is, the disputed view can be formed. The answers given must be a number from 1 to 5, corresponding to the five candidate defense statements.

Input: The prosecution statement: according to law, the defendant should be investigated the criminal responsibility and compensated for 185,823 yuan for medical expenses, 50,616 yuan for lost work during hospitalization (63 days × 832 yuan), 50,616 yuan for nursing expenses for 63 days, and 1050 yuan for hospital meals...a total of 565,571 yuan. **Defense statement 1**: When I went to the home of the villager Yu Jia, I told him that I needed some bricks to prepare the “kang”, and Yu Jia said that he had some bricks to sell to me (3 yuan per piece). **Defense statement 2**: There is no causal relationship between the private prosecutor’s leg injury and the defendant.**Defense statement 3**: To sum up, the defendant does not bear criminal responsibility. For the part of civil compensation, the defendant did slap the private prosecutor in the face, and should bear the consequences of the harm caused to the private prosecutor, and compensate the private prosecutor for economic losses within this scope. **Defense statement 4**: I slapped the prosecutor in the face, and I accept responsibility only to this extent. **Defense statement 5**: The prosecutor has no direct evidence to prove that the defendant directly injured his right leg. From the five possible defense statements, please choose the one opposing the prosecution’s view:

Answer: 2

Table 16 An example of document proofreading task

Instruction: Suppose you are a professional clerical proofreader, please help me complete the following tasks. The input to the task is a text, and the output is a list. Each element in the list represents an error and is represented by a dictionary whose keys are ‘pos’, ‘ori’, ‘edit’, and ‘type’. The value of ‘pos’ is the error position in the input text, the value of ‘ori’ is the error text, and the value of ‘edit’ is the corrected text. The value of ‘type’ is the type of the error. The error types can be typos, redundancies, omissions, or disorder, representing errors, redundant, missing, and out-of-order, respectively. Please note that you are looking for the error, not retouching the text. If there are no problems, an empty list is returned.

Input: The municipal municipal administrative department shall organize the maintenance and repair of damaged urban roads promptly.

Answer: [{‘pos’: 0, ‘ori’: ‘municipal’, ‘edit’: ‘ ’, ‘type’: ‘redundant’}].

Table 17 An example of entity recognition task

Instruction: Please mark the “person”, “location”, “time”, “type of drugs” and “weight of drug” entities contained in the text below. You only need to output the results of the annotations, do not explain the reason or output other information. For each type of entity, output the corresponding type of entity text segment in the sentence in the form of a list, and different entity types are separated by commas.

Input: Zhongshan City first District People’s Procuratorate accused: On the evening of October 27, 2014, the defendant Song Mou, to thank the friends of drug user Pang Mou for their help, sent a package of drugs (methamphetamine components were detected by inspection, net weight 1.87 gs) to Room 302 of the Suixing Building opposite the Shailang Suixing Market in the West of Zhongshan City for Pang Mou and others to smoke. After being brought to justice, Song confessed his crime truthfully.

Answer: “time”: [“On the evening of October 27, 2014”,“November 4th at 1 p.m”], “person”:[“Song Mou”, “Pang Mou”], “location”: [“Room 302, Suixing Building, opposite Suixing Market, West Shailang, Zhongshan City”], “type of drugs”: [“methamphetamine”], “weight of drug”: [“1.87g”]

Table 18 An example of relation extraction task

Instruction: Please mark the relationships between entities contained in the following sentences. You only need to output the results of the annotations, do not explain the reason or output other information. The following is the definition of a relationship: “Trafficking”: Both the first and last entities of the relational triad are the persons involved “Trafficking (drugs)”: The head entity of the triad is the person involved and the tail entity is the item involved “Holding”: The head entity of the triad is the person involved, and the tail entity refers specifically to the drug-type entity “Illegal accommodation”: Both the first and last entities of the relational triad are involved persons Please output all relational triples as a list, each string enclosed in English quotation marks, and each triplet contains [“ head entity text”,“ relation”,“tail entity text ”].

Input: The People’s Procuratorate of Changshu City accused the defendant Xu Mou of smoking methamphetamine (crystal meth) three times between February and March 2015 in Changshu City and taking illegal shelter to Chi Mou and Gu Mou, Xinhe Village (19) Hebaishi East residence.

Answer: [[“Xu Mou”, “illegal shelter”, “Chi Mou”], [“xu”, “illegal shelter”, “Gu Mou”], [“Xu Mou”, “hold”, “methamphetamine”]]

Table 19 An example of event detection task

Instruction:Please note the following sentences contains “indulge”, “violence”, “casino”, “lost” and “rent/borrow”, “buy”, “liaison” and “damage”, “extortion”, “demand/request”, “suicide”, “drinking” and “arrested”, “surrender”, “contract/agreement”, “rent/borrow” and “explosion”, “identification ”, “sale”,“payment/payment” or “stop/prevent”, “restitution for”, “cheat”, “post”, “poison”, “kill” and “spoils”, “employment”, “limited/detention”, “death”, “to inform/remind”, “possession/hide” and “help/assistance”, “conflict”, “traffic accident”, “landslide”, “to sell Drugs ”, “kidnapping”, “natural disaster”, “confession” and “illegal driving” and “fire” and “profit”, “prostitution” and “physical”, “give up/stop”, “spreading” and “provocative/play”, “trapped”, “armed/gun”, “goons” and “collusion” and “left”, “introduce/referral”, “drug” and “abusive”, “plunder”, “reject/resist”, “harm person”, “report”, “offer” and “threat/force”, “financing”, “convention”, “seize property”, “sell”, “credit” and “poisoning”, “pretend”, “misappropriation of property”, “speech conflict”, “gambling”, “altered”, “looted”, “instigating/instigated”, “a bribe ”, “smuggling”, “leaked information”, “tracking”, “damaging property”, “flood”, “agree/accept” and “knowing”, “stun” and “sell”, “rape” and “fake”, “back when”, “alarm/report”, “kidnapping” and “dry” and “search/to seize”, “prostitution” and “stolen property” and “compensation”, “guaranty”, “rent / loan ”, “obscene” and “accident”, “understanding”, “injury”, “recommended”, “manufacturing”, “fire”, “transport/”,“ inviting/drum up ”, “organization/arrangements”, “house/door” event. You only need to print the annotated result of the event that exists, without explaining the reason or printing other information. For each trigger event, please use the form of a list of output [“event type”, “trigger word”]

Input: It was also found that in the criminal proceedings, Li Mouxiang, the mother of the victim Wang Moumou, filed an incidental civil suit to the court, asking the defendant Li Moubo to compensate for his mental loss of 20,000 yuan, 6,000 yuan for lost work, a total of 26,000 yuan.

Answer: [[“give up/stop”, “filed”], [“incidental civil action ”, “lift”], [“compensation ”, “compensation”]]

Table 20 An example of judgment summarization task

Instruction: Please write a summary of the following judgment that preserves important information about the judgment.

Input: Plaintiff Ou Wuxiang v. defendant Dongguan Yongqing Municipal Services Co., LTD. (hereinafter referred to as Yongqing Municipal Company) labor contract dispute case, the court on June 16, 2017, after filing and accepting, according to law, the application of summary procedures, public hearing. 1. Ou Wuxiang's claim: Yongqing Municipal Company shall pay Ou Wuxiang: 73650.9 yuan of overtime wage difference between May 21, 2007, and December 31, 2016; The defendant shall bear the costs of the case. 2. Whether financial compensation should be paid: Ouwuxiang was born on May 5, 1966, and reached the age of 50 on May 4, 2016. The labor relationship between the two parties was terminated because Ouwuxiang reached the statutory retirement age, and the relationship between the two parties turned into a labor relationship on May 5, 2016. Ouwuxiang filed an arbitration on May 26, 2017, and demanded that Yongqing Municipal Company pay economic compensation for the termination of the labor relationship. This court does not support; 3. Overtime pay: Ouwuxiang's claim for overtime pay from May 21, 2007, to May 27, 2015, has expired the limitation of arbitration, and the Court does not support it; From May 28, 2015, to May 4, 2016, it can be seen from the payroll confirmed by Ouwuxiang's signature provided by Yongqing Municipal Company that Ouwuxiang has been in attendance for 22 days a month and there is no overtime work. Ouwuxiang's request for overtime pay lacks basis and is not supported by the court. It is a labor relationship after May 5, 2016, during which the plaintiff and the defendant have a labor relationship, and the defendant has paid the plaintiff's labor remuneration every month, and the plaintiff's request for the defendant to pay overtime is unfounded, and the court does not support it. 4 High-temperature allowance: Ouwuxiang's request for a high-temperature allowance from 2007 to 2014 has expired the statute of limitations, and the court does not support it; From the salary confirmed by the Yongqing Municipal company Ouwu Xiang, it can be seen that the Yongqing Municipal company has paid Ouwu Xiang 2015 annual high-temperature allowance, the relationship between the two sides has turned into a labor relationship since May 5, 2016, Ouwu Xiang requires Yongqing Municipal company to pay 2016 high-temperature allowance, the lack of basis, the court does not support. In summary, according to the provisions of Article 64 and Article 142 of the Civil Procedure Law of the People's Republic of China, the judgment is as follows: All claims of plaintiff Ou Wuxiang are rejected. Please write a summary of the judgment:

Answer: The case involves a labor contract dispute between the plaintiff Ou Wuxiang and the defendant Dongguan Yongqing Municipal Services Co., LTD. The plaintiff seeks payment of the difference in overtime pay for the period from May 21, 2007, to December 31, 2016, and costs. After hearing the case, the court found that the plaintiff's claim lacked basis, because the overtime pay and high-temperature allowance it requested had expired or did not exist in fact, and after the termination of the labor relationship, the two parties turned into labor relations, and the defendant had paid the plaintiff monthly labor remuneration. Based on the above reasons, the court ruled that all the claims of the plaintiff were dismissed.

Table 21 An example of public opinion summarization task

Instruction: Legal public opinion refers to public opinion and network information related to laws, judicial organs, and judicial activities. Help me write a summary of legal opinion based on the following statements:

Input: Fuyang Procuratorate for photo Hangzhou, September 25 (reporter Guo Qiyu) Reporters learned from the Fuyang District People's Procuratorate of Hangzhou on the 25th that the hospital was involved in a murder case 17 years ago in advance, according to law, the suspect Huo Mou on suspicion of intentional homicide to decide to approve the arrest. After investigation, on the evening of March 23, 2003, the suspect Huo Mou in Hangzhou Fuyang District (former Fuyang city) Fuchun Street GUI Hua West Road guest hair salon after identifying the victim Huang Mou, Huang Mou will be taken to a rental near the hair salon for sex, after the dispute between the two sides due to the problem of prostitution, the suspect Huo Mou is the use of neck and other means to kill the victim Huang Mou. Huo was captured by police. After the incident, the suspect Huo Moumou fled Fuyang overnight and fled back to his hometown of Liaoning Province after several turns to earn a living by doing odd jobs. He was arrested by police at his home in Liaoning Province on September 10, 2020. On September 14, the Fuyang District Procuratorate intervened in advance in the case of Huo Mou suspected of intentional homicide, put forward opinions on guiding investigation and obtaining evidence on issues such as the norms of evidence collection and the direction of continued investigation, and then approved the arrest of Huo mou by the law on suspicion of intentional homicide. The case is currently under further investigation.

Answer: The People's Procuratorate of Fuyang District of Hangzhou intervened in a murder case 17 years ago and approved the arrest of the suspect Huo Mou on suspicion of intentional homicide. It is understood that on the evening of March 23, 2003, Huo Mou in a hair salon in Fuyang District of Hangzhou, and the victim Huang Mou after a dispute, took violent means to kill him. After the incident, Huo fled to his hometown of Liaoning Province and was arrested by police on September 10, 2020. The Fuyang District Procuratorate intervened in the case in advance on September 14, and after putting forward guidance on the norms of evidence collection and other issues, approved the arrest of Huo Mou suspected of intentional homicide following the law. At present, the case is under further investigation.

Table 22 An example of court's opinion generation task

Instruction: Please write a legal Judgment based on the fact statement and plaintiff's claim as follows:

Input Statement: After the trial, the court found that on November 5, 2012, the defendant He Yidong borrowed 20,000 yuan from the plaintiff, and issued an IOU agreement to return the loan on December 5, 2012. The IOU states: "Today, He Yidong borrowings 20,000 yuan from Li Gang due to the need for funds, which will be returned only on December 5, 2012, at 20,000 yuan." The defendant stamped his hand on the borrower column and on the lower case of \20,000 and indicated his identity card number below the borrower. However, the defendant failed to repay the loan as agreed before the due date. The plaintiff claims: 1. The defendant is ordered to repay the principal of the loan of the plaintiff RMB 20,000; 2. The defendant shall bear the litigation costs of this case. Generate the court's opinion (to respond effectively to the plaintiff's claim) in the form that the beginning is the court's opinion:

Answer: The Court held that according to the contents of the IOU, the defendant He Yidong borrowed RMB 20,000 from the plaintiff and explicitly agreed to return it on December 5, 2012. However, the defendant's failure to repay the loan as agreed after maturity constitutes a breach of contract. Therefore, the court supported the plaintiff's first claim, ordering the defendant to return the plaintiff's loan principal of 20,000 yuan. As for the plaintiff's second claim, the request for the defendant to bear the litigation costs, the court considers it reasonable, but it needs to decide according to the specific circumstances.

Funding This work was supported by the Zhejiang Province's Lingyan R&D Project [Grant No. 2024C01259].

References

Achiam J, Adler S, Sandhini Agarwal ea (2023) Gpt-4 technical report. arxiv:2303.08774

Bai J, Bai S, Chu Y, et al (2023) Qwen technical report. arxiv:2309.16609

Bergmeir C, Benítez JM (2012) On the use of cross-validation for time series predictor evaluation. Inf Sci 191:192–213

Blair JA (2012) Relevance, acceptability and sufficiency today. Science 5:87–100. https://doi.org/10.1007/978-94-007-2363-4_8

Cai Z, Kung PN, Suvarna A, et al (2024) Improving event definition following for zero-shot event detection. In: Ku LW, Martins A, Srikumar V (eds) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 2842–2863, arxiv:2403.02586

Cerqueira V, Torgo L, Mozetič I (2020) Evaluating time series forecasting models: an empirical study on performance estimation methods. Mach Learn 109:1997–2028

Chalkidis I, Androutsopoulos I, Aletras N (2019) Neural legal judgment prediction in english. In: Annual Meeting of the Association for Computational Linguistics, https://doi.org/10.18653/v1/P19-1424

Chen Y, Sun Y, Yang Z, et al (2020) Joint entity and relation extraction for legal documents with legal feature enhancement. In: Proceedings of the 28th International Conference on Computational Linguistics, pp 1561–1571

Christiano P, Leike J, Brown TB, et al (2023) Deep reinforcement learning from human preferences. arxiv:1706.03741

Conrad J, Zeleznikow J (2015) The role of evaluation in ai and law: an examination of its different forms in the ai and law journal. Science. https://doi.org/10.1145/2746090.2746116

Cui J, Li Z, Yan Y, et al (2023) Chatlaw: open-source legal large language model with integrated external knowledge bases. arxiv:2306.16092

Deroy A, Ghosh K, Ghosh S (2024) Ensemble methods for improving extractive summarization of legal case judgments. Artifi Intell Law 32:231–289. https://doi.org/10.1007/s10506-023-09349-8

Drápal J, Westermann H, Savelka J (2023) Using large language models to support thematic analysis in empirical legal studies. arxiv:2310.18729

Fei Z, Shen X, Zhu D, et al (2023) Lawbench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289

Gururangan S, Marasović A, Swayamdipta S, et al (2020) Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964

Hayatin N, Ghufron KM, Wicaksono GW (2021) Summarization of covid-19 news documents deep learning-based using transformer architecture. TELKOMNIKA (Telecommunication Computing Electronics and Control) 19(3):754–761

Huang Y, Yu Z, Guo J et al (2020) Legal public opinion news abstractive summarization by incorporating topic information. Int J Mach Learn Cybern 11:2039–2050

Jain D, Borah MD, Biswas A (2021) Summarization of legal documents: where are we now and the way forward. Comput Sci Rev 40:100388

Jain D, Borah MD, Biswas A (2023) A sentence is known by the company it keeps: improving legal document summarization using deep clustering. Artifi Intell Law 5:1–36

Kalyan KS (2024) A survey of gpt-3 family large language models including chatgpt and gpt-4. Nat Lang Process J 6:100048. https://doi.org/10.1016/j.nlp.2023.100048

Leitner E, Rehm G, Moreno-Schneider J (2019) Fine-grained named entity recognition in legal documents. In: International conference on semantic systems, pp 272–287, https://doi.org/10.1007/978-3-030-33220-4_20

Lewis P, Perez E, Piktus A et al (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst 33:9459–9474

Li D, Zhao Q, Chen J et al (2021) Adan: an intelligent approach based on attentive neural network and relevant law articles for charge prediction. IEEE Access 99:1–1

Li Q, Zhang Q, Yao J, et al (2020) Event extraction for criminal legal text. In: 2020 IEEE International Conference on Knowledge Graph (ICKG), pp 573–580, https://doi.org/10.1109/ICBK50248.2020.00086

Li Y, Huang H, Ma S, et al (2023) On the (in) effectiveness of large language models for chinese text correction. arXiv preprint arXiv:2307.09007

Liang P, Bommasani R, Lee T, et al (2022) Holistic evaluation of language models. arXiv preprint arXiv:2211.09110

Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp 74–81, https://aclanthology.org/W04-1013

Lin W, Han M, Jin T (2023) Multi-stage legal instrument grammatical error correction via seq2edit and data augmentation. In: Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing, pp 215–220

Malik V, Sanjay R, Nigam SK, et al (2021) Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. arXiv preprint arXiv:2105.13562

Nasar Z, Jaffry SW, Malik MK (2021) Named entity recognition and relation extraction: state-of-the-art. ACM Comput Surv 54(1):1–39

Ouyang L, Wu J, Jiang X, et al (2022) Training language models to follow instructions with human feedback. arXiv e-prints

Poudyal P, Šavelka J, Ieven A, et al (2020a) Echr: Legal corpus for argument mining. In: Proceedings of the 7th Workshop on Argument Mining, pp 67–75, https://api.semanticscholar.org/CorpusID:227230459

Poudyal P, Savelka J, Ieven A, et al (2020b) ECHR: Legal corpus for argument mining. In: Cabrio E, Villata S (eds) Proceedings of the 7th Workshop on Argument Mining, pp 67–75, https://aclanthology.org/2020.argmining-1.8

Qin C, Zhang A, Zhang Z, et al (2023) Is chatgpt a general-purpose natural language processing task solver? https://doi.org/10.18653/v1/2023.emnlp-main.85, arxiv:2302.06476

Radford A, Wu J, Child R, et al (2019) Language models are unsupervised multitask learners. https://api.semanticscholar.org/CorpusID:160025533

Rodriguez JD, Perez A, Lozano JA (2009) Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell 32(3):569–575

Sapatinas T (2005) Discriminant analysis and statistical pattern recognition. J R Stat Soc Ser A Stat Soc 168(3):635–636. https://doi.org/10.1111/j.1467-985X.2005.00368_10.x

Savelka J, Ashley KD (2023) The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. Front Artifi Intell 6:63. https://doi.org/10.3389/frai.2023.1279794

Savelka J, Agarwal A, Bogart C, et al (2023a) Can generative pre-trained transformers (gpt) pass assessments in higher education programming courses? In: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1, p 117-123, https://doi.org/10.1145/3587102.3588792

Savelka J, Ashley KD, Gray MA, et al (2023b) Explaining legal concepts with augmented large language models (gpt-4). arxiv:2306.09525

Shao Y, Mao J, Liu Y, et al (2020) Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In: International Joint Conference on Artificial Intelligence, https://api.semanticscholar.org/CorpusID:267909336

Socher R, Perelygin A, Wu JY, et al (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Empirical Methods in Natural Language Processing

Srivastava A, Rastogi A, Rao A, et al (2023) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint https://doi.org/10.48550/arXiv.2206.04615

Sun X, Li X, Li J, et al (2023) Text classification via large language models. In: Bouamor H, Pino J, Bali K (eds) Findings of the Association for Computational Linguistics: EMNLP 2023, pp 8990–9005, https://doi.org/10.18653/v1/2023.findings-emnlp.603

Sun Y, Yang F, Wang X et al (2021) Automatic generation of the draft procuratorial suggestions based on an extractive summarization method: Bertslca. Math Probl Eng. https://doi.org/10.1155/2021/3591894

Tjong Kim Sang EF, De Meulder F (2003) Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp 142–147, https://aclanthology.org/W03-0419

Wang A, Singh A, Michael J, et al (2018) Glue: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp 353–355, https://doi.org/10.18653/v1/W18-5446

Wang A, Pruksachatkun Y, Nangia N, et al (2020a) SuperGLUE: A stickier benchmark for general-purpose language understanding systems. arxiv:1905.00537

Wang J, Liang Y, Meng F, et al (2023) Zero-shot cross-lingual summarization via large language models. In: Proceedings of the 4th New Frontiers in Summarization Workshop, pp 12–23

Wang Y, Yao Q, Kwok JT et al (2020) Generalizing from a few examples: a survey on few-shot learning. ACM Comput Surv 53(3):254

Wei J, Wang X, Schuurmans D, et al (2023) Chain-of-thought prompting elicits reasoning in large language models. arxiv:2201.11903

Wong TT (2015) Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recogn 48(9):2839–2846

Wu S, Liu Z, Zhang Z, et al (2023) fuzi.mingcha. https://github.com/irlab-sdu/fuzi.mingcha

Wu Y, Kuang K, Zhang Y, et al (2020) De-biased court's view generation with causality. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 763–780

Xiao C, Zhong H, Guo Z, et al (2018) Cail2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478

Xiao C, Hu X, Liu Z et al (2021) Lawformer: a pre-trained language model for Chinese legal long documents. AI Open 2:79–84

Xu N, Wang P, Chen L, et al (2020) Distinguish confusing law articles for legal judgment prediction. arxiv:2004.02557

Xu QS, Liang YZ (2001) Monte carlo cross validation. Chemom Intell Lab Syst 56(1):1–11

Yang A, Xiao B, Wang B, et al (2023) Baichuan 2: open large-scale language models. arxiv:2309.10305

Ye H, Jiang X, Luo Z, et al (2018) Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. arXiv preprint arXiv:1802.08504

Yu F, Quartey L, Schilder FA (2023) Exploring the effectiveness of prompt engineering for legal reasoning tasks. In: Annual Meeting of the Association for Computational Linguistics, https://api.semanticscholar.org/CorpusID:259848886

Zhang S, Dong L, Li X, et al (2023) Instruction tuning for large language models: a survey. arXiv preprint arXiv:2308.10792

Zhong H, Xiao C, Tu C, et al (2020) Jec-qa: a legal-domain question answering dataset. In: Proceedings of the AAAI conference on artificial intelligence, pp 9701–9708

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Xue Guo[1,4] · Yuting Huang[1] · Bin Wei[2,3] · Kun Kuang[1,3] · Yiquan Wu[1,3] · Leilei Gan[1,3] · Xianshan Huang[1] · Xianglin Dong[2,3]

✉ Bin Wei
binwei@zju.edu.cn

✉ Kun Kuang
kunkuang@zju.edu.cn

Xue Guo
guoxue@sxu.edu.cn

Yuting Huang
yutinghuang@zju.edu.cn

Yiquan Wu
wuyiquan@zju.edu.cn

Leilei Gan
leileigan@zju.edu.cn

Xianshan Huang
huangxs@zju.edu.cn

Xianglin Dong
xianglingdong@zju.edu.cn

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, China

[2] Guanghua Law School, Zhejiang University, Hangzhou, China

[3] Law&AI Lab, Zhejiang University, Hangzhou, China

[4] College of Automation and Software,Shanxi University, Shanxi, China