ASP2LJ : An Adversarial Self-Play Laywer Augmented Legal Judgment Framework

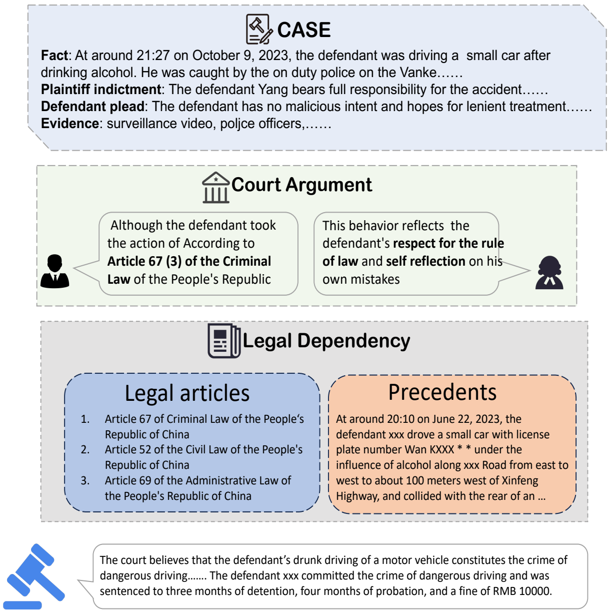
Anonymous ACL submission

Abstract

Legal Judgment Prediction (LJP) aims to predict judicial outcomes, including relevant legal articles, terms, and fines, leveraging advancements in artificial intelligence and Large Language Models (LLMs). However, despite such progress, LJP faces two key challenges: (1)Data Labeling: Current datasets, derived from authentic cases, suffer from high human annotation costs and imbalanced distributions, leading to model performance degradation. (2)Lawyer's Improvement: Existing systems focus on enhancing judges' decision-making but neglect the critical role of lawyers in refining arguments, which limits overall judicial accuracy. To address these issues, we propose an Adversarial Self-Play Lawyer Augmented Legal Judgment Framework, called ASP2LJ, which integrates a controversy-aware case generation module to tackle long-tailed data distributions and an adversarial self-play mechanism to enhance lawyers' argumentation skills. Our framework enables a judge to reference evolved lawyer's arguments, improving the objectivity, fairness, and rationality of judicial decisions. We also introduce RareCases, a benchmark for rare legal cases in China, and demonstrate the effectiveness of our approach on the SimuCourt dataset. Experimental results show significant improvements, with a 9% increase in legal article generation accuracy over AgentsCourt and 14% over GPT-4 on average. Our contributions include a novel integrated framework, a rare-case benchmark, and publicly releasing datasets and code to support further research in automated judicial systems.

1 Introduction

Legal Judgment Prediction (LJP) aims to predict judgment results of a legal case, including the relevant legal articles, terms, fines, and other related aspects(Cui et al., 2022). With the advancement of artificial intelligence, an increasing number of studies have been proposed to assist humans in Legal Judgment Prediction (LJP). Especially in recent years, the emergence of LLMs has made further progress in this field, substantially improving the performance of LJP tasks.



Fact: At around 21:27 on October 9, 2023, the defendant was driving a small car after drinking alcohol. He was caught by the on duty police on the Vanke……
Plaintiff indictment: The defendant Yang bears full responsibility for the accident……
Defendant plead: The defendant has no malicious intent and hopes for lenient treatment……
Evidence: surveillance video, poljce officers,……

Although the defendant took the action of According to **Article 67 (3) of the Criminal Law** of the People's Republic

This behavior reflects the defendant's **respect for the rule of law** and **self reflection** on his own mistakes

1. Article 67 of Criminal Law of the People's Republic of China
2. Article 52 of the Civil Law of the People's Republic of China
3. Article 69 of the Administrative Law of the People's Republic of China

At around 20:10 on June 22, 2023, the defendant xxx drove a small car with license plate number Wan KXXX * * under the influence of alcohol along xxx Road from east to west to about 100 meters west of Xinfeng Highway, and collided with the rear of an ...

The court believes that the defendant's drunk driving of a motor vehicle constitutes the crime of dangerous driving……. The defendant xxx committed the crime of dangerous driving and was sentenced to three months of detention, four months of probation, and a fine of RMB 10000.

Figure 1: Based on a real case, the lawyers argue and the judge make judgment according to legal dependency.

In order to achieve improved performance, the current work primarily involves collecting relevant cases from government public websites to conduct supervised training. However, despite these advancements, the current LJP task faces two major challenges: (1) **Data Labeling**: The current datasets are derived from authentic cases which incurs substantial human expenditures. Besides, the distribution of the data adheres to the 80/20 rule and some cases are ignored so that there is not sufficient data to train, which results in the degradation of the model. (2) **Lawyer's Improvement**: Recently some works try to introduce a simulated court to help the judge improve the final judgment accuracy. However, they just focus on the judge while the lawyers don't get full improvement, which limits the performance of the judge's final judgment.

As illustrated in Figure 1, in real-world legal practice, the lawyers start to argue based on a provided case, and the judge should reference law articles and precedents to make his judgment. In common law system, judicial precedents can serve as binding authorities for court decisions, whereas in civil law system, precedents are often utilized for interpretative guidance or reference purposes. Besides, judges' adjudicative capabilities are enhanced through accumulating experience from handling numerous cases. It is obvious that legal precedents play an important role in judgment. However, the current distribution of cases exhibits a long-tail characteristic, meaning that certain types of cases represent a very small proportion of the overall cases and a human judge is puzzled when he meets such cases. For instance, in cases involving copyright infringement of AI-generated images, there is a lack of relevant legal articles or precedents that can serve as appropriate references, making it challenging for judges to reach a well-informed decision. For automated judicial systems, the issue of data distribution leads to a lack of generalizability in models when dealing with rare cases. As indicated in Table 2, even state-of-the-art models exhibit performance degradation when processing infrequent cases compared to more common ones. While (Wang et al., 2024) attempts to address this issue through LLM-generated cases, these methods still require manual judgment annotation, limiting their scalability and practical applicability. These limitations highlight the need for improved approaches to improve the automated judicial system's capability in managing rare cases. Besides, the role of legal professionals, particularly lawyers, is equally critical in the judicial process. Through their arguments, lawyers can present case analysis and relevant legal references and help organize the facts of the case, provide legal perspectives, demonstrate the effect of evidence, and identify potential points of contention, which contributes to the development of balanced judicial judgment. According to the arguments, the judge can make an accurate judgment. Empirical studies have demonstrated that judicial outcomes are influenced by the quality of legal arguments presented, leading to more equitable rulings. Nevertheless, current research in this domain faces limitations, either neglecting the role of legal arguments or being constrained by insufficient real-world data for optimizing legal argumentation. Therefore, enhancing the capabilities of lawyers to provide valuable references for judicial decision-making presents another significant challenge.

To address these challenges, we propose an Adversarial Self-Play Laywer Augmented Legal Judgment framework, which enables the judge to reference the augmented lawyers' arguments and improve the objectivity, fairness, and rationality of judicial decisions of the judgment. In order to address the issue of real cases' distribution, we propose a controversy-aware complex case generation module. Our approach incorporates a case-court pipeline to mitigate the challenges posed by the long-tailed case distribution and facilitate the accumulation of judicial experience. To demonstrate the effectiveness of our method, we introduce a benchmark called RareCases which encompasses rare cases in China. All the cases are sampled from the China Judgements Online.[1] Furthermore, in order to enhance the proficiency of lawyers, we propose an adversarial self-play mechanism for lawyer agents where the plaintiff and defendant lawyers engage in case analysis and confrontation, iteratively accumulating agent experience and improving their legal analysis capabilities. The system integrates lawyers' argumentative content with judicial decision-making modules to support more objective, impartial and reasonable adjudication. The experimental results demonstrate that our framework effectively enhances the capabilities of the agents and exhibits strong performance on our proposed benchmark for rare cases.

Our contributions are as follows:

- We propose an integrated framework that incorporates lawyer arguments iteratively enhancement into judicial decision-making processes, enabling judges to better understand the case and make more accurate predictions.
- We introduce RareCases, a legal benchmark including the main rare cases, which provides an approach to assess the legal capacity of current LLMs.
- We demonstrate the effectiveness of our framework by conducting experiments on SimuCourt, a public data set in China. Experimental results show that our framework outperforms the existing methods in various aspects.

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

Impressively, in legal article generation, we get a 14% increase higher than GPT-4, indicating the utility of the proposed framework in supporting legal decision-making processes.

- To enable further research, we will release our datasets and code publicly.

2 Related Work

Legal Artificial Intelligence is a rapidly growing field that has gathered significant interest among researchers, encompassing various tasks such as legal case retrieval (LCR), statutory article retrieval (SAR), and legal judgment prediction (LJP).

2.1 Legal AI

Prior to the advent of LLMs, legal tasks were predominantly addressed using conventional artificial intelligence techniques. CAIL(Xiao et al., 2018) was established as a well-known annual Chinese legal AI competition, featuring tasks like LCR and LJP, which attracts widespread participation from legal AI researchers. Some studies(Niklaus et al., 2021) focus on the legal language varies in different countries, trying to construct benchmarks to evaluate the concurrent models' capacity in dealing with different language. Some works(Chalkidis et al., 2020; Douka et al., 2021; Limsopatham, 2021) try to introduce specific retriever models like Bert into legal tasks

2.2 LLM+Law

Following the introduction of ChatGPT, numerous studies have explored integrating LLMs into legal tasks, yielding promising results. Due to LLM's strong performance on reasoning, (Yao et al., 2023) combines LLM with legal knowledge. For instance, LawBench(Fei et al., 2023) comprises approximately 20 tasks focused on legal memory, understanding, and application. GEAR(Qin et al., 2024) introduces a methodology that constructs a hierarchical structure of legal articles, thereby augmenting the model's interpretative capabilities. (Zhou et al., 2024) proposes LawGPT by fine-tuning Chinese-LLaMA with Chinese legal knowledge. Additionally, several works(Li et al., 2023; Pipitone and Alami, 2024; Feng et al., 2024; Hou et al., 2024; Gao et al., 2024) have contributed to enhancing the retrieval capabilities of legal systems. LLMs can further improve their performance through retrieval-augmented generation (RAG). Concurrently, some researchers have explored using LLMs to tackle legal entrance exam questions(Kim et al., 2024) but the performance is not satisfying, indicating the large challenge in legal field. Other studies such as (Wu et al., 2023) have demonstrated that combining LLMs with domain-specific legal models can enhance LJP performance, while (Qin et al., 2024) introduced GEAR, a novel framework integrating LCR, SAR, and LJP.

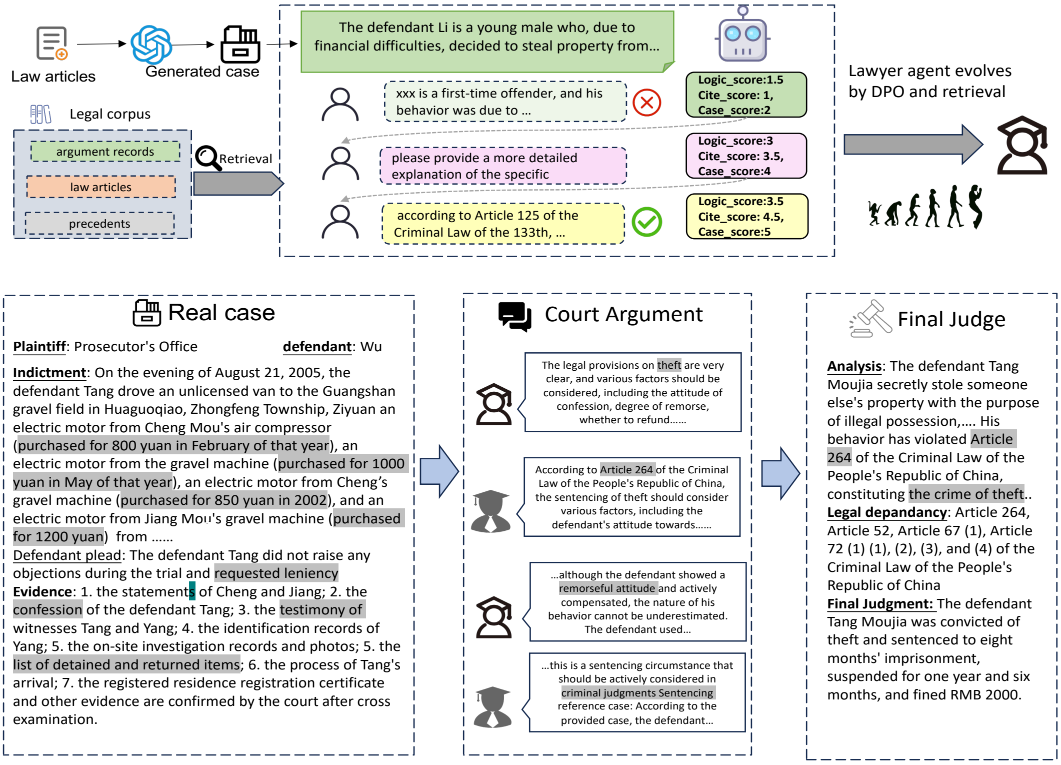
2.3 Legal Agent

With the advent of LLM-based agents, researchers try to simulate courtroom environments using these agents. For example, (Chen et al., 2024a) employs agents to engage in debates and generate extensive records to refine their capabilities. Similarly, (He et al., 2024) proposes a framework where lawyer agents argue, and the judge retrieves relevant legal articles, precedents, and papers to ensure the accuracy of the final judgment. However, these works overlook the critical role of lawyers which results in suboptimal performance, leaving room for further improvement.

In this paper, we propose an agent-based framework that integrates court argumentation and judgment prediction. Furthermore, we introduce a lawyer agent evolution mechanism to assist judges in making more accurate judicial predictions. In this framework, lawyer agents present arguments, which are evaluated by the judge through a scoring system. The lawyers then refine their arguments based on this feedback, leading to continuous improvement in the quality and structure of their legal statements which enhances the overall effectiveness of the legal argumentation and LJP process.

3 Method

We propose an adversarial self-play lawyer augmented legal judgment framework that combines the adversarial argument process with the continuous enhancements of lawyer capabilities and the judge's Legal Judgment Prediction. This framework is designed to provide the judge with a more comprehensive understanding of cases, facilitating his impartial and accurate decision-making. The conceptual structure of this approach is illustrated in Figure2.



The defendant Li is a young male who, due to financial difficulties, decided to steal property from...

xxx is a first-time offender, and his behavior was due to ...

please provide a more detailed explanation of the specific

according to Article 125 of the Criminal Law of the 133th, ...

Plaintiff: Prosecutor's Office **defendant**: Wu

Indictment: On the evening of August 21, 2005, the defendant Tang drove an unlicensed van to the Guangshan gravel field in Huaguoqiao, Zhongfeng Township, Ziyuan an electric motor from Cheng Mou's air compressor (purchased for 800 yuan in February of that year), an electric motor from the gravel machine (purchased for 1000 yuan in May of that year), an electric motor from Cheng’s gravel machine (purchased for 850 yuan in 2002), and an electric motor from Jiang Mou's gravel machine (purchased for 1200 yuan) from ......

Defendant plead: The defendant Tang did not raise any objections during the trial and requested leniency

Evidence: 1. the statements of Cheng and Jiang; 2. the confession of the defendant Tang; 3. the testimony of witnesses Tang and Yang; 4. the identification records of Yang; 5. the on-site investigation records and photos; 5. the list of detained and returned items; 6. the process of Tang's arrival; 7. the registered residence registration certificate and other evidence are confirmed by the court after cross examination.

The legal provisions on theft are very clear, and various factors should be considered, including the attitude of confession, degree of remorse, whether to refund......

According to Article 264 of the Criminal Law of the People's Republic of China, the sentencing of theft should consider various factors, including the defendant's attitude towards......

...although the defendant showed a remorseful attitude and actively compensated, the nature of his behavior cannot be underestimated. The defendant used...

...this is a sentencing circumstance that should be actively considered in criminal judgments Sentencing reference case: According to the provided case, the defendant...

Analysis: The defendant Tang Moujia secretly stole someone else's property with the purpose of illegal possession,.... His behavior has violated Article 264 of the Criminal Law of the People's Republic of China, constituting the crime of theft..

Legal depandancy: Article 264, Article 52, Article 67 (1), Article 72 (1) (1), (2), (3), and (4) of the Criminal Law of the People's Republic of China

Final Judgment: The defendant Tang Moujia was convicted of theft and sentenced to eight months' imprisonment, suspended for one year and six months, and fined RMB 2000.

Figure 2: Lawyer Evolution: Before a lawyer’s speech, he will retrieve some corpus; after that, an evaluator scores and give some explanations to improve his speech

3.1 Court Pipeline

To enhance the capabilities of lawyers within our framework, we implement a multi-stage approach that utilizes LLM to simulate court proceedings. Our pipeline is structured into three distinct phases: (1) legal case generation, (2) court argumentation, and (3) lawyer evolution which aims to systematically improve the argumentation and reasoning skills of lawyer agents.

3.1.1 Legal Case Generation

Previous approaches primarily rely on real cases as the foundation for legal arguments. However, this methodology presents several limitations. First, a large portion of existing legal cases lacks controversy, offering limited opportunities for lawyers to fully apply their professional expertise. Second, the distribution of legal cases follows a long-tail pattern, which means that certain types of cases are rare and difficult for LLMs to generalize, hindering their ability to learn and address related legal issues effectively. Consequently, lawyer agents often face challenges in leveraging prior experience when encountering such infrequent cases.

To overcome these limitations, we propose a pipeline that utilizes LLM to automatically generate simulated legal cases. We collect a comprehensive collection of Chinese legal articles spanning criminal, civil, and administrative law. For each case generation, a subset of articles is randomly selected to serve as the legal foundation and an LLM is then instructed to generate cases that include key components such as evidence, relevant legal articles, factual descriptions, reasoning, and judgments based on the selected articles. By permuting and combining these articles randomly, the LLM can produce different legal scenarios. Besides, to ensure the quality and complexity of the generated cases, we employ a rejection sampling strategy. Cases that are overly simplistic or excessively anomalous will be discarded. This process ensures that the cases are sufficiently complex and debatable, providing a robust platform for lawyers to demonstrate and refine their argumentative skills.

Through this automated case generation process, lawyers can continuously engage in arguments based on a diverse and evolving dataset. As the number of cases increases, lawyers gain more opportunities to develop their expertise, leading to progressive improvement in their capabilities. This approach not only addresses the limitations of relying solely on real-world cases but also creates a dynamic environment for the continuous evolution

Model	Legal Articles			Civil and Admini.			Criminal			Case Analysis		
	P	R	F	P	R	F	Charge	Prison term	Fine	Correct	Logic	Concise
						First						
LawGPT	0.11	0.05	0.07	0.21	0.43	0.3	0.69	0.13	0.09	0.33	0.46	0.39
Qwen1.5-7B-Chat	0.16	0.10	0.11	0.27	0.50	0.34	0.88	0.21	0.25	0.75	0.75	0.76
GPT-3.5	0.24	0.09	0.13	0.29	0.56	0.36	0.85	0.30	0.25	0.67	0.71	0.71
GPT-4	0.27	0.11	0.14	**0.33**	0.60	0.41	0.89	0.31	0.27	0.83	0.78	0.83
AgentsCourt	0.32	0.16	0.22	0.31	0.57	**0.44**	0.88	0.31	0.19	0.74	0.77	0.74
ours	**0.42**	**0.27**	**0.27**	0.31	**0.67**	0.41	**0.91**	**0.33**	**0.28**	**0.83**	**0.88**	**0.86**
						Second						
LawGPT	0.09	0.04	0.06	0.21	0.59	0.33	0.58	0.06	0.18	0.28	0.20	0.33
Qwen1.5-7B-Chat	0.20	0.16	0.16	0.29	0.65	0.38	0.85	0.21	0.22	0.39	0.42	0.42
GPT-3.5	0.22	0.09	0.12	0.45	0.76	0.5	0.82	0.19	0.28	0.4	0.42	0.43
GPT-4	0.21	0.13	0.15	0.38	**0.78**	0.47	0.88	0.22	0.27	**0.7**	**0.71**	**0.7**
AgentsCourt	0.27	0.28	0.27	0.40	0.76	0.45	0.83	0.20	0.29	0.58	0.66	0.64
ours	**0.35**	**0.3**	**0.31**	**0.56**	**0.78**	**0.57**	**0.87**	**0.23**	**0.32**	0.46	0.52	0.47

Table 1: Overall performance of SimuCourt and baselines in the first and second instance experimental settings.

of lawyer agents.

3.1.2 Court Argumentation

The argument process in the simulated court follows the procedures of real court trials. In the initial phase, the plaintiff's lawyer must meticulously prepare the complaint based on the facts of the case and legal grounds, clearly stating the claims, facts, and reasons. The defendant's lawyer, on the other hand, must respond to the content of the complaint by addressing factual determinations and legal applications, thereby formulating a defense statement. Once the formal argument stage begins, both sides take turns presenting their arguments over three rounds, with each round consisting of several core components:

Statement Both lawyers must articulate their own standpoints and legal claims. This serves as the foundation of the argument and the starting point for subsequent arguments.

Retort The plaintiff's lawyer must counter the arguments presented by the defendant in the previous round, pointing out logical flaws or errors in the legal application. The defendant's lawyer, in turn, must respond to the plaintiff's rebuttals while identifying weaknesses in the plaintiff's arguments.

Legal Citations When presenting their arguments, both lawyers must support their claims with relevant legal evidence, such as legal articles, judicial interpretations, and precedents. This not only tests their legal research skills but also their ability to extract relevant information and engage in logical reasoning.

Summarization Lawyers' statements must be concise and accurate, without distorting or misrepresenting the facts of the case. Through the description, understanding, and reasoning of the case, the judge can gain a clear and intuitive understanding of the matter.

After each round of arguments, the judge provides the lawyers with an opportunity to reflect on and revise their statements. The judge evaluates the lawyers' performance based on multiple dimensions, including legal citations, reasoning, and factual descriptions, and offers constructive feedback. The lawyers then refine their arguments based on this feedback, thereby improving the quality of their presentations. Through this multi-round, multi-faceted argument, the factual details of the case are fully revealed, and the points of contention are clearly presented. This process not only enhances the quality of the lawyers' arguments but also aids the judge in rendering a fair and just verdict.

3.1.3 Lawyer Agent Evolution

The judge's understanding of a case is partially shaped by the argumentative skills of the legal representatives. The manner in which lawyers construct their arguments—through their form, structure, and content—significantly aids the judge in comprehending the case details. Effective legal representatives typically deliver arguments that are clear, logically structured, and supported by relevant legal citations and precedents, along with precise descriptions of the case's substantive relationships. Developing such skills necessitates ongoing practice and learning through simulated case scenarios. Despite this, there has been limited focus in prior research on refining the content of courtroom arguments, which is crucial for thorough case analysis. Our objective is to improve the argumentative proficiency of lawyers, enabling them to present case information in a more organized and comprehensive manner. To this end, we introduce a method for lawyer capacity enhancement that facilitates continuous learning and refinement of debating abilities, thereby enriching the data available for the judge's legal judgment prediction task.

Previous research has paid limited attention to the automated evaluation of court arguments. To address this gap, we introduce a subjective evaluation metric tailored to assess the quality of lawyers' arguments, with the aim of identifying higher-quality presentations through a structured scoring system. The scoring framework focuses on three key dimensions: (1) the ability to accurately understand and cite relevant legal articles and precedents; (2) the logical coherence and organization of the argument; and (3) the depth and comprehensiveness of case analysis. Each dimension is scored on a scale of 0 to 5, yielding a total possible score of 15 points. Beginning in the first round, after each lawyer presents their argument, the content is evaluated using this metric, and constructive feedback is provided to guide improvements. The lawyer agent then refines its argument based on the feedback. This iterative process is repeated three times, with the highest-scoring argument selected as the final submission.

3.2 Judgment Prediction

The lawyers who have evolved through simulated cases can engage in arguments on real cases and the judge can reference the generated records. During the case judgment process, the judge agent not only considers the arguments presented by the lawyers but also utilizes an advanced legal retrieval system to search for relevant cases and legal articles. This retrieval process primarily relies on two authoritative legal databases: for criminal cases, the judging agent uses the LeCardv2 database, which contains a large number of representative Chinese criminal cases; for civil and administrative law cases, the SimuCourt's legal-KB database is employed, which includes 6.5 million legal cases and complete legal provisions from 2017 to 2022, with 50,000 selected legal cases forming the core retrieval corpus of the system. This dual retrieval mechanism ensures the accuracy and comprehensiveness of case adjudication.

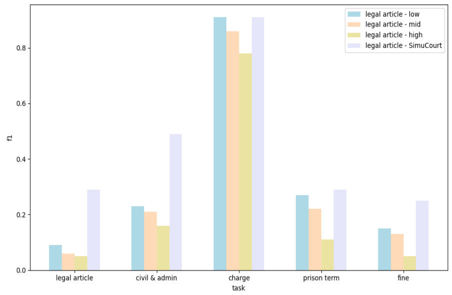


Figure 3: Models' performance between rare cases and common cases. Rare cases are divided into low, mid and high, which represents their rarity.

After completing the analysis of argument records and legal retrieval, the judge agent enters the case prediction phase. This phase mainly involves two core tasks: predicting the judgment outcome and determining the legal basis. In the prediction of criminal cases, the judge agent needs to accurately determine whether the defendant's charges are established and predict the corresponding fines and prison terms based on the circumstances of the case. For civil and administrative cases, the judge agent needs to predict specific judgment outcomes, such as the validity of contracts and the division of liability for compensation. In terms of determining the legal basis, the judging agent must accurately identify and cite the relevant legal articles to ensure that the judgment is well-founded.

4 Experiment

In this section, we start to evaluate the performance of our framework in downstream tasks. We will elaborate on our benchmark, experiment design, and result analysis.

4.1 Benchmark

We adopt the SimuCourt benchmark of previous work, AgentsCourt, as the main part of our experiments. SimuCourt is a Chinese benchmark consisting of 420 cases which encompasses objective evaluations and subjective analyses, including first-instance and second-instance cases. Besides, in order to evaluate the current models' capacity of handling rare cases, we propose our benchmark called RareCases, which consists of 180 rare cases encompassing civil, administrative and criminal law. These legal cases are divided into high-rare, mid-rare and low-rare by their rarity.

Model	Legal Articles			Civil and Admini.			Criminal			Case Analysis		
	P	R	F	P	R	F	Charge	Prison term	Fine	Correct	Logic	Concise
LawGPT	0.12	0.02	0.04	0.09	0.11	0.09	0.56	0.09	0.02	0.03	0.06	0.14
Qwen1.5-7B-Chat	0.15	0.05	0.06	0.21	0.23	0.21	0.82	0.15	0.06	0.13	0.19	0.24
GPT-3.5	0.17	0.05	0.07	0.31	0.24	0.26	0.82	0.22	0.08	0.38	0.42	**0.57**
GPT-4	0.23	0.10	0.14	**0.35**	0.28	0.28	**0.84**	**0.26**	0.09	0.42	**0.47**	0.44
AgentsCourt	0.20	0.04	0.07	0.33	0.29	0.25	0.82	0.24	0.07	0.41	0.39	0.47
ours	**0.25**	**0.14**	**0.16**	0.33	**0.31**	**0.31**	**0.84**	0.22	**0.11**	**0.43**	0.44	0.52

Table 2: Overall performance of our RareCases and baselines.

4.2 Settings

Models. We adopt Qwen1.5-7B-Chat as base model and simultaneously compare it with lawGPT(Zhou et al., 2024), GPT-3.5-turbo-1106 and gpt-4-1106-preview. For subsequent retrieval and optimization, Qwen1.5-7B-Chat is used as the base model.

Baselines. We compare our method with the following baselines:

(1) Vanilla. We choose Qwen1.5-7B-Chat, GPT3.5-turbo-1106, GPT-4-1106-preview as vanilla models, and the base model of our ASP2LJ is Qwen1.5-7B-Chat.

(2) LawGPT. LawGPT is Chinese-LLaMA-7B fune-tuned on a dataset of 300,000 legal question-answer pairs.

(3) AgentsCourt. An LLM agent framework. They improve the judge's performance by introducing argument data and retrieving several law articles, precedents and law papers.

Retriever. BGE-m3(Chen et al., 2024b) is an advanced retriever proposed by BAAI, which leads to superior performances on multi-lingual retrieval, cross-lingual retrieval, and multi-lingual long-document retrieval tasks, while in legal tasks, sparse retriever BM25 is in common use for its relevance scoring algorithm. In this paper, we adopt a hybrid retrieval method to search for argument records, cases, and legal articles. Regarding argument records and cases, due to the context limitations of Qwen1.5-7B-Chat, we use BM25 for retrieval to obtain 100 candidate documents and then use BGE-M3 for reranking. Finally, 1 document is selected as the retrieved document. For legal articles, we use BM25 to retrieve 1000 legal articles as candidates and then use bge-m3 as the reranker. Eventually, 30 legal articles are selected.

Finetune. As shown in Figure 2, each statement is assessed and scored according to evaluation metrics. Every case undergoes three rounds of dialogue, with each round being evaluated three times. From these evaluations, we select the highest and lowest scores to conduct DPO (Differential Privacy Optimization) training. Besides, we fine-tune the BGE-m3 with our generated data. Specifically, since our case generation is based on legal articles, each case is associated with several gold legal articles. Consequently, we establish an index based on the fundamental facts of each case, utilizing the gold legal provisions as positive samples and an equivalent number of highly similar non-gold legal articles as negative samples. This approach facilitates the DPO training, thereby enhancing the legal provision retrieval capability.

4.3 Results

Table 1 and Table 2 present the main experimental results on SimuCourt and our RareCourt, respectively.

4.3.1 SimuCourt

As Figure 1 shows, our method outperforms other methods overall. Compared with the vanilla model, the performance has improved.

Criminal Prediction. As for crime prediction, We extract the charge, prison term and fine by regular expressions. Among all the results, AgentsCourt achieves the best during the baselines, indicating the importance of argumentation and retriever. Although our method achieves the best, we don't have an obvious advance. There is still a large room for improvement.

Civil and Administrative Prediction. In the area of civil and administrative laws, our indicators comprehensively surpass those of the vanilla Qwen1.5-7B-Chat and slightly exceed those of GPT-3.5. Due to the relatively flexible judgment results in civil and administrative laws, we use GPT-4o as an analyzer to extract key points and compare them with the key points of the reference answers as evalua-

Model	Legal Articles	Judgement Results			
		Civil and Admini.	Charge	Prison term	Fine
ASP2LJ	0.16	0.31	0.84	0.22	0.11
w/o Court argument	0.13	0.23	0.79	0.17	0.08
w/o Lawyer Evolution	0.14	0.26	0.81	0.19	0.08
w/o Retriever	0.09	0.24	0.82	0.21	0.10

Table 3: Ablation Experiment on RareCases

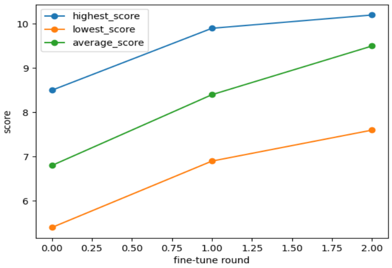


Figure 4: Score of lawyer agent by fine-tuning rounds

tion indicators. This also reflects that there is still a huge gap between current judgment and evaluation work in civil and administrative laws.

Law Articles. It is seen that the average accuracy of charge prediction by our method exceeds that of the vanilla Qwen1.5-7B-Chat by 20%, and exceeds that of GPT-3.5 and GPT-4 by 16% and 14% respectively. In terms of prison terms, our method also exceeds GPT-4 by an average of about 5% on average. Regarding fines, it exceeds the vanilla Qwen-1.5-7B-Chats by about 12%.

4.3.2 RareCases

As shown in Table 2, our work outperforms other models twice. Our framework enables Qwen1.5-7B-Chat to reach 25%, 14%, 16% in legal articles, respectively. It is also obvious that all the models exhibit performance degradation in various degree. Especially in civil and administrative cases, the performance has a large decline, averaging 17%, which indicates the difficulty of models to judge rare cases and there is a large room for improvement.

4.4 Ablation and Analysis

Ablation As demonstrated in the table, retrieval plays a pivotal role in the task of legal provision generation, enhancing the F1 score by 7%. In terms of judicial outcomes, the retrieval of precedents can also assist the model in adjudicating cases. Furthermore, the analysis of the original case debates enables the model to better comprehend the cases, thereby improving the accuracy of the judgment outcomes. The evolution of the lawyer agent will elevate the quality of discourse, consequently augmenting the understanding of the cases.

rare cases As illustrated in Figure 3, the performance of the model progressively declines with the increasing rarity of cases. With the exception of ’charge’, all other metrics fall significantly below those of simuCourt, indicating that the model’s capability to handle rare cases is insufficient and there is considerable room for improvement.

Fine-tune Iteratively As illustrated in Figure 4, we direct Qwen1.5-7B-Chat to produce argument records and conduct DPO training utilizing this data. During the initial round, 400 records are generated and Qwen1.5-7B-Chat undergoes fine-tuning. In the subsequent round, Qwen1.5-7B-Chat-dpo is employed to create an additional 400 records, followed by the fine-tuning of Qwen1.5-7B-Chat-dpo. The efficacy of these models is ultimately assessed based on their scores across 20 cases. Ultimately, we compiled the highest scores, the lowest scores, and the average scores of their arguments across these 20 cases. It is evident that, as the tuning iterations progress, all three categories of scores have improved and gradually stabilized.

5 Conclusion

We conduct a thorough analysis of our framework’s performance. In our framework, lawyer agents can evolve and the judge can benefit from the evolution. To deal with the legal cases’ long-tail distribution, we propose a method to gather legal cases by generating legal cases based on legal articles. Then We fine-tune the Qwen1.5-7B-Chat with the generated data to gain a better performance. The experimental results show that our method enables a weak model, Qwen1.5-7B-Chat, to surpass powerful models like GPT-4. Besides, the proposed benchmark, RareCases, also indicates that there is still an improvement room in LJP task.

6 Limitations

In this work, we primarily introduce an approach to generate cases automatically and propose a benchmark encompassing rare cases. Despite our contribution, there is still some limitations. We just focus on Chinese laws while there are still various cases which are much different, leaving room to explore. We only focus on SimuCourt and our RareCases benchmark without evaluating other well-know datasets like LAiW(Dai et al., 2024), LawBench or CAIL. We just study LJP task, overlooking other tasks like LCR or SAR. We plan to further explore the legal tasks in future studies.

References

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. 2024a. Agentcourt: Simulating court with adversarial evolvable lawyer agents.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. 2024. Laiw: A chinese legal large language models benchmark.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. JuriBERT: A masked-language model adaptation for French legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models.

Yi Feng, Chuanyi Li, and Vincent Ng. 2024. Legal case retrieval: A survey of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Gao, Chaojun Xiao, Zhenghao Liu, Huimin Chen, Zhiyuan Liu, and Maosong Sun. 2024. Enhancing legal case retrieval via scaling high-quality synthetic query-candidate pairs.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416, Miami, Florida, USA. Association for Computational Linguistics.

Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2024. Clerc: A dataset for legal case retrieval and retrieval-augmented analysis generation.

Yeeun Kim, Young Rok Choi, Eunkyung Choi, Jinhwan Choi, Hai Jin Park, and Wonseok Hwang. 2024. Developing a pragmatic benchmark for assessing korean legal language understanding in large language models.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023. Lecardv2: A large-scale chinese legal case retrieval dataset.

Nut Limsopatham. 2021. Effectively leveraging BERT for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 210–216, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain.

Weicong Qin, Zelin Cao, Weijie Yu, Zihua Si, Sirui Chen, and Jun Xu. 2024. Explicitly integrating judgment prediction with legal document retrieval: A law-guided generative approach. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 2210–2220. ACM.

Yen-Hsiang Wang, Feng-Dian Su, Tzu-Yu Yeh, and Yao-Chung Fan. 2024. A cross-lingual statutory article retrieval dataset for taiwan legal studies.

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models.

Zhi Zhou, Jiang-Xin Shi, Peng-Xiao Song, Xiao-Wen Yang, Yi-Xuan Jin, Lan-Zhe Guo, and Yu-Feng Li. 2024. Lawgpt: A chinese legal knowledge-enhanced large language model.