LawArgueAgent: A Framework to Enhance Legal Judgment Prediction via Lawyer-Adversarial Self-Play and Case Generation

Anonymous ACL submission

Abstract

Given the fact of a legal case, Legal Judgment Prediction (LJP) aims to predict judicial outcomes, including relevant legal charges, prison terms, and fines. However, LJP faces two key challenges: (1)Long-Tail Distribution: Existing datasets, derived from authentic cases, suffer from high human annotation costs and imbalanced distributions, leading to model performance degradation. (2)Lack of Lawyer Augmentation: Current systems focus on enhancing judges' decision-making but neglect the critical role of lawyers in refining legal arguments, thereby limiting overall judicial accuracy. To address these issues, we propose LawArgueAgent, an adversarial self-play lawyer augmented legal judgment framework, which integrates a case generation module to tackle long-tailed data distributions and an adversarial self-play mechanism to enhance lawyers' argumentation skills. Our experiments on a Chinese legal dataset show that our framework enables a weak model, Qwen1.5-7B-Chat, to surpass powerful models like GPT-4 in legal judgment prediction. This demonstrates the effectiveness of our approach in improving LJP performance by simulating a courtroom adversarial process.

1 Introduction

Given the facts of a case, Legal Judgment Prediction (LJP) aims to predict the applicable legal articles, prison terms, fines, and other judgment outcomes (Cui et al., 2022). With the advancement of legal AI, an increasing number of studies have been conducted to improve models' judgment prediction capabilities (Xu et al., 2020; Semo et al., 2022). In recent years, the emergence and popularization of LLM agents have made further progress in this field (Li et al., 2025a; Wu et al., 2023). LLM agents can simulate a real legal courtroom and thereby help to improve the performance of LJP tasks (Feng et al., 2022; Huang et al., 2024).

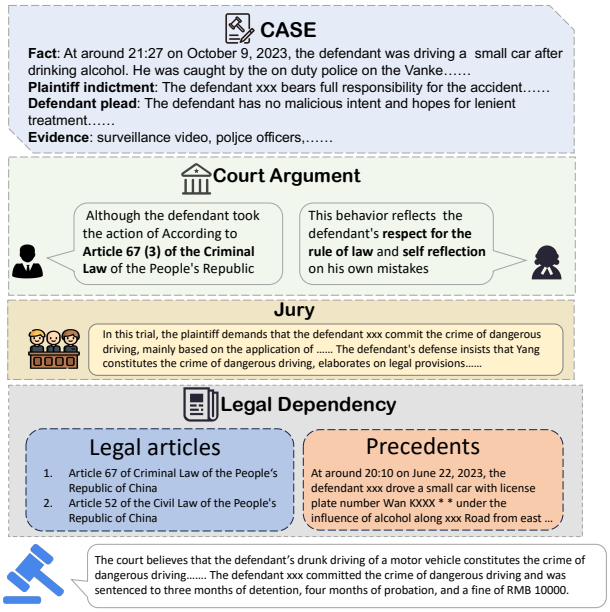

Figure 1: Based on a real case, the lawyers argue and the judge makes legal judgments according to legal dependency.

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(previous legal cases) to make a final judgment (Zhong et al., 2018; Ma et al., 2021). There is no doubt that legal precedents(prior cases) play a significant role as references to legal judgment in a judicial process. In the case law system (Mistica et al., 2020), precedents can serve as binding authorities for court decisions, whereas in the civil law system (Büttner and Habernal, 2024), precedents are often utilized for interpretative guidance or reference purposes.

However, the current distribution of cases is highly imbalanced, with many case types occurring at very low frequencies, resulting in the insufficiency of relevant legal cases; moreover, existing LJP tasks predominantly focus on optimizing judge agents, while overlooking the role of lawyers in the

process. Therefore, the current LJP task faces two major problems: cases' *long tail distribution* and *insufficient modeling of lawyer Agents*.

Long Tail Distribution: The distribution of the data adheres to the Pareto Principle (80/20 rule), and some rare cases are paid less attention by LLM (Li et al., 2025b), which results in the shortage of relevant knowledge. The current distribution of cases exhibits a long-tail characteristic (Li, Ting et al., 2025), meaning that certain types of cases represent a very small proportion of the overall cases, and even a human judge will be puzzled when he meets such cases. Empirical studies (Garrett, 2011; Gross et al., 2014; Završnik, 2021) demonstrate that specific cases have a higher probability of being overturned, such as contract disputes, sexual assault, murder, etc. For automated judicial systems, while (Wang et al., 2024) attempts to address this issue through LLM-generated cases, these methods still require manual annotation of judgment results, limiting their scalability and practical applicability. These limitations highlight the need for improved approaches to improve the capability of the automated judicial system to handle rare cases without human effort.

Insufficient Modeling of Lawyer Agents: Recent work uses simulated courts to improve judicial accuracy (He et al., 2024; Chen et al., 2024a), but mainly optimizes judges and leaves lawyers underdeveloped, which constrains the judgment quality. However, lawyers are also crucial. The experienced lawyers achieve higher litigation success rates than less experienced ones (Poppe and Rachlinski, 2015; Anderson and Heaton, 2012; Shiu-fan, 1983), and their legal arguments help judges understand cases more clearly and make decisions more fairly. Empirical studies show that the quality of legal argumentation significantly influences judicial outcomes (Habernal et al., 2024; Sheppard and Moshirnia, 2012). Yet current research often ignores legal arguments or is limited by scarce real-world data, making the improvement of lawyers' capabilities a key open challenge for judicial decision-making.

To address the challenges above, we propose a framework to Enhance Legal Judgment Prediction via Lawyer-Adversarial Self-Play and Case Generation, which enables the judge to reference the augmented lawyers' arguments and improve the judgment's objectivity, fairness, and rationality. In order to address the issue of real cases' long-tail distribution, we propose a case generation module. As the generated cases contain only case facts, plaintiffs' indictments, and defendants' pleas, excluding final judgments, our approach is independent of human judgment. Our framework 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 dataset called RareCases which encompasses rare cases in China, sampled from the China judgments Online. 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 across both datasets, particularly on RareCases.

Our key contributions are as follows:

- We propose a framework called LawArgueAgent, which is the first to incorporate a lawyer's perspective and utilize lawyers in a self-play optimized process for judgment. In our framework, base model Qwen1.5-7B-Chat gains large improvement by legal knowledge infusion.
- We construct RareCases, a legal dataset 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. The experimental results show that our framework outperforms the existing methods in various aspects. Impressively, in legal article generation, we get a 8% higher than GPT-4 in recall score, indicating the utility of the proposed framework in LJP tasks.

2 Related Work

Legal Artificial Intelligence is a rapidly growing field that has gathered significant interest among re-

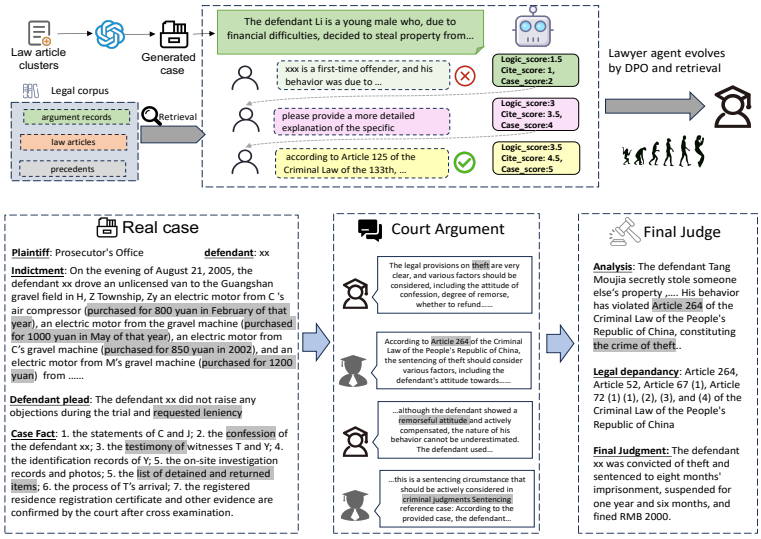

Figure 2: Lawyer Evolution: Before a lawyer's speech, he will retrieve some corpus; after that, some evaluator scores and explanations are given to improve his speech

searchers. Therefore, many methods are proposed to improve AI models' performance.

2.1 LLM+Law

Due to their strong reasoning capabilities, LLMs have become the preferred choice for developing AI assistants in the legal domain. For example, LawBench (Fei et al., 2023) comprises approximately 20 tasks focused on legal memory, understanding, and application, and GEAR (Qin et al., 2024) constructs a hierarchical structure of legal articles for augmenting the model's interpretative capabilities. However, LLMs are limited by insufficient legal knowledge, so fine-tuning and RAG are deemed as a valid knowledge supplement method. (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.

2.2 CourtRoom Simulation

The single model with external knowledge can not simulate a real world courtroom, which limits the models' legal AGI ability. Therefore, researchers utilize LLM agents to gather and incoroperate related information to improve downstream tasks' performance. For example, (Chen et al., 2024a) employs agents to engage in arguments 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. Recently, several works introduce knowledge augmentation to improve agents' ability. (Liu et al., 2025) can generate dialogue knowledge as retrieval corpus and employ RAG to retrieve semantic similar information to augment its capacity in downstream tasks. Besides, (Chen et al., 2025) focus on multi-agent debate to improve the judges' judgment precision by providing different views and lawyer agents' self-feedback.

However, these works overlook the critical role of lawyers which results in suboptimal performance, leaving room for further improvement.

3 Method

In this section, we introduce our self-play lawyer augmented LJP framework. As illustrated in Figure 2, we implement a multi-stage approach to simulate court proceedings with legal agents. Our pipeline is structured into four distinct phases: (1) legal case generation, (2) court argumentation, (3) lawyer evolution, and (4) judgment prediction, which aims to systematically improve the argumentation and reasoning skills of lawyer agents.

3.1 Legal Case Generation

Previous approaches rely on real cases as the foundation for legal arguments. However, these real cases' distribution follows a long-tail distribution, which limits LLMs' generalization. Lawyer agents also need diverse and adequate cases to enable their

retriever	recall@100	recall@200	recall@500	recall@1k
Lawformer	5.6	7.2	9.2	20.9
Bm25	5.5	10.6	22.1	29.6
BGE-M3	13.5	19.2	27.3	36.0
ours	23.6	51.7	65.3	75.8

Table 1: Different retrievers' recall@k in SimuCourt.

argumentation skills. To overcome these limitations, we propose a pipeline that utilizes LLM to generate simulated legal cases automatically.

We use a vast collection of Chinese legal articles (criminal, civil, administrative). For each case generation, several legal articles are randomly selected to serve as the legal foundation, and GPT-4o is instructed to generate cases consisting of case facts, plaintiff's indictment, and defendant's pleadings. To ensure quality and complexity, we employ rejection sampling, discarding overly simple or anomalous cases after GPT-4o evaluation. This guarantees sufficiently complex and debatable cases, providing a strong platform for lawyers to improve their skills. Additionally, three law students verify the cases' correctness and logic. Detailed information is provided in B.

This automated case generation process enables lawyer agents to engage in continuous argumentation grounded in a dynamic and expanding dataset. The increasing volume of generated cases provides lawyers with enhanced opportunities for expertise development, fostering progressive improvements in their professional capabilities. This approach not only mitigates the inherent limitations of relying exclusively on real-world case experience but also cultivates a dynamic environment for the further evolution of lawyer agents.

3.2 Court Argumentation

The simulated court argumentation process is similar to the procedural framework of real-world trials. Initially, the plaintiff is required to meticulously formulate the complaint, grounded in the case's factual description and relevant legal principles, clearly articulating the claims, factual basis, and legal reasoning. Conversely, the defendant's counsel must respond substantively to the complaint's content by addressing both factual determinations and legal applications, thereby constructing a comprehensive defense statement. Upon commencement of the formal argument phase, both parties engage in alternating presentations across three rounds, with each round comprising 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 refutations 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.

Subsequent to each round of argumentation, GPT-4o offers lawyer agents an opportunity for analysis and revision of their statements. The judge evaluates lawyer performance across multiple dimensions, encompassing legal citation, logical reasoning, and factual description, and provides constructive feedback. Lawyers then leverage this feedback to refine their arguments, thereby enhancing the overall 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 elevates the the lawyer arguments' caliber, but also facilitates the judge to render an equitable verdict.

3.3 Lawyer Agent Evolution

The judge's understanding of a case is partly shaped by the argumentative skills of the legal representatives. Clear, logically structured arguments that incorporate relevant legal citations, precedents, and accurate descriptions of the facts greatly facilitate judicial comprehension. Developing such advocacy skills, however, requires sustained practice and learning through simulated case scenarios.

Our objective is to improve the argumentative proficiency of lawyers, enabling them to present case information in a more organized and comprehensive manner. Inspired by (Fu et al., 2025), 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.

Model	Legal Articles			Civ. & Adm.			Criminal			Case Analysis		
	P	R	F	P	R	F	Cha.	Term	Fine	Cor.	Log.	Con.
First												
LawGPT	11.3	6.0	7.2	21.5	43.6	30.0	69.0	13.9	9.2	33.2	46.8	39.0
LexiLaw	7.8	7.4	7.4	17.8	17.6	17.6	87.9	18.1	34.5	77.4	78.5	81.2
ChatLaw	8.1	6.6	7.2	13.2	15.4	13.8	82.4	19.9	27.4	66.7	81.2	78.4
Qwen1.5-7B-Chat	9.7	5.8	6.3	26.5	34.2	29.3	86.2	25.0	27.5	75.8	79.5	76.4
GPT-3.5	13.0	11.5	11.8	33.5	41.8	33.2	85.0	48.7	31.2	67.7	71.4	71.0
GPT-4	**18.1**	13.3	14.1	45.8	51.3	**47.6**	88.1	**53.1**	35.5	83.5	78.0	83.5
AgentsCourt	16.7	12.9	13.6	**46.7**	45.1	45.3	87.5	48.8	32.5	74.2	77.8	74.8
ours	15.1	**21.3**	**15.6**	37.8	**52.9**	40.1	**89.5**	38.4	**46.3**	**83.5**	**88.5**	**86.0**
Second												
LawGPT	9.2	4.3	6.9	21.7	55.9	33.1	58.4	6.3	18.9	28.2	20.4	33.6
LexiLaw	5.6	5.7	5.6	27.8	27.8	27.8	78.2	27.3	32.6	41.7	39.6	45.2
ChatLaw	7.6	6.3	6.8	28.5	27.8	28.0	79.5	26.8	33.2	44.3	48.9	54.7
Qwen1.5-7B-Chat	20.4	5.8	8.5	29.0	65.0	38.0	86.0	9.8	22.8	39.0	42.5	42.5
GPT-3.5	11.6	9.6	10.1	45.4	55.0	48.3	85.0	28.0	31.2	40.0	42.1	43.5
GPT-4	17.1	16.1	16.1	**72.9**	**75.4**	**73.7**	88.6	**37.9**	39.2	**70.8**	**71.2**	**70.6**
AgentsCourt	22.6	18.5	18.9	54.9	61.7	57.1	84.9	32.0	**50.9**	46.3	52.6	47.6
ours	**27.1**	**22.3**	**23.1**	62.5	69.7	63.5	**91.0**	36.3	43.3	58.6	66.3	64.0

Table 2: Overall performance of SimuCourt and baselines in the first and second instance experimental settings.

Some works have demonstrated that judge's LJP capacity can improve by lawyers' arguments(Chen et al., 2024a; He et al., 2024; Chen et al., 2025). However, they ignore that the lawyer agents can improve their argument abilities through agents' self-play evolution, which leaves space for further improvement. To address this gap, we introduce a subjective evaluation metric tailored to assess the quality of lawyers' arguments, to identify 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. 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, and the highest-scoring argument is selected as the final submission. The detailed score criterion is illustrated in Appendix.

3.4 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 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. Therefore, we collect all the cases from Wenshu Web from 2018 to 2021, encompassing criminal, civil, and administrative cases. The detailed information is in Table 5. This retrieval mechanism ensures the precision and comprehensiveness of the LJP tasks. After completing the analysis of argument records and legal retrieval, the judge agent enters the legal judgment prediction phase. This phase mainly involves three core tasks: predicting the judgment results, determining the relevant legal articles, and analysing the whole case.

4 Experiment

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

4.1 Benchmark

We employ two benchmarks: SimuCourt and RareCases, which can evaluate the effectiveness of our framework. The elaborate information of dataset and metric are as below.

4.1.1 Dataset

SimuCourt is a Chinese benchmark consisting of 420 cases that encompasses objective evaluations and subjective analyses, including first-instance and second-instance cases.

Furthermore, as presented in Figure 3, in order to evaluate the capacity of current models to handle rare cases, we collected all Chinese cases from China's Online judgments and conducted a statistical analysis of the frequency of case causes from 2018 to 2021. The least frequently occurring causes are identified and categorized as rare case causes. We collect 266 cases involving rare causes from the WenShu web that occurred after 2022. Then we invite three undergraduate students

Model	Legal Articles			Civ. & Adm.			Criminal			Case Analysis		
	P	R	F	P	R	F	Cha.	Term	Fine	Cor.	Log.	Con.
LawGPT	4.4	2.3	2.9	9.7	11.3	10.1	56.0	9.8	2.2	3.0	6.3	14.3
LexiLaw	6.5	4.8	4.9	9.6	12.3	10.2	82.5	12.3	14.2	11.3	9.8	8.6
ChatLaw	3.3	5.9	3.7	15.4	18.0	16.2	80.0	37.2	14.3	12.3	26.8	17.2
Qwen1.5-7B-Chat	5.3	4.1	4.2	8.2	10.1	9.0	75.9	10.7	3.0	13.3	19.7	24.7
GPT-3.5	10.0	4.8	6.0	15.6	17.4	16.2	81.2	38.0	7.4	38.6	42.0	37.6
GPT-4	14.7	11.5	12.0	**18.2**	**19.8**	**18.6**	83.5	**43.9**	**15.6**	42.6	47.0	44.7
AgentsCourt	16.0	11.8	12.1	14.9	16.2	15.4	82.6	41.7	15.1	42.0	**47.6**	44.0
ours	**16.8**	**11.9**	**12.7**	16.0	18.3	16.7	**85.0**	16.1	10.7	**43.3**	44.7	**52.0**

Table 3: Overall performance of our RareCases.

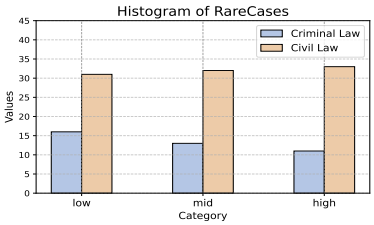

Figure 3: Detailed statistical histogram of RareCases.

Dataset	SimuCourt	RareCases
cases	420	136
average articles	3.71	6.90
if-rare	✗	✔
Average length per case	440.9	384.7

Table 4: Basic statistics of the datasets.

to help us check and filter some cases and instruct GPT-4o to transform the cases into specific structures. Finally, we get 136 cases, including 96 civil cases and 40 criminal cases.These legal cases are divided into "high", "mid", and "low" by their rarity, and "low" cases are rarer than "mid" and "high cases.

4.1.2 Metric

. We divide our task into three categories: legal articles, final judgment, and case analysis. For legal articles, we use regular expressions to extract law articles from generated law articles and check whether the extracted articles match our answers or not. Then we calculate the answers' precision, recall, and F1. For judgment results, to facilitate the evaluation process, we employ DeepSeek to standardize the answer format. For criminal results, we use a regular expression to check criminal results, and for civil and administrative cases, we use GPT-4o as an automatic evaluator. It compares the agent system's judgment results with the reference judgments and computes the number of matching and non-matching key points between them. Detailed illustration is presented in Appendix D.

4.2 Settings

Parameters. We adopt Qwen1.5-7B-Chat as base model and simultaneously compare it with lawGPT (Zhou et al., 2024), LexiLaw (Li et al., 2024), ChatLaw-13B (Cui et al., 2024), GPT-3.5-turbo-0613, and GPT-4-1106-preview. Qwen1.5-7B-Chat is used as the base model for subsequent retrieval and optimization, and GPT-4o is instructed to evaluate and score the lawyer agents' arguments.

Baselines. We compare our method with the following baselines:
(1) Vanilla. We choose Qwen1.5-7B-Chat, GPT3.5-turbo-0613, GPT-4-1106-preview as vanilla models. Besides, the base model of our framework is also Qwen1.5-7B-Chat.
(2) LawGPT (Zhou et al., 2024). LawGPT is Chinese-LLaMA-7B fine-tuned on a dataset of 300,000 legal question-answer pairs.
(3) LexiLaw (Li et al., 2024). LexiLaw is a fine-tuned Chinese legal large language model based on the ChatGLM-6B architecture. By undergoing fine-tuning on legal domain datasets, it demonstrates enhanced performance and professionalism in providing legal consultation and support.
(3) ChatLaw-13B (Cui et al., 2024). ChatLaw-13B is an innovative assistant that employs a Mixture-of-Experts (MoE) model and a multi-agent system to enhance reliability and accuracy in AI legal services.
(5) AgentsCourt (He et al., 2024). An LLM agent framework. They improve the judge's performance by introducing argument data and retrieving several law articles, precedents and law papers. We utilize GPT-3.5-turbo-1106 as its base model.

Finetune. As shown in Figure 2, each statement is scored against evaluation metrics, and each case undergoes 3 dialogue rounds with 3 evaluations

data	sum	criminal	civil	admin
cases	20.87M	2.45M	17.55M	0.87M
laws	13,117	5,425	7,692	

Table 5: Case Volume of our corpus

per round; top- and bottom-scoring arguments are selected for DPO training, with evolution scores in Figure 4. Additionally, we use GPT-4o to generate 10,000 legal-article-based cases, fine-tune BGE-M3, and retrieve 50 law articles per case via BM25 (gold articles as positives, others as negatives) to conduct DPO fine-tuning and boost retriever performance. As Figure 1 shows, our performance outperforms other methods, but the 75.8% peak Recall@1000 leaves significant room for improvement in Legal Judgment Prediction (LJP) tasks.

Retriever. BGE-M3 (Chen et al., 2024b) is an advanced retriever proposed by BAAI, which leads to superior performance in multilingual retrieval, cross-lingual retrieval, and multilingual long document retrieval tasks, while in legal tasks, the sparse retriever BM25 (Rosa et al., 2021) is in common use due to 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, we use BM25 for retrieval to obtain 100 candidate documents and then use BGE-M3 for reranking. Finally, due to the context limitations of Qwen1.5-7B-Chat, one document is selected as the retrieved document. For legal articles, we use our fine-tuned BGE-M3 to retrieve 200 legal articles as candidate articles.

Corpus. We collect all the Chinese legal cases in 2021, spanning criminal, civil and administrative cases, which are more than 27M, as our case corpus. Each case in our corpus consists of five factors: case name, action cause, stage, relevant articles, and full text. To protect the privacy of involved parties, all case records were anonymized by removing personally identifiable information, including names, geographic locations, and other sensitive details. Detailed data statistics are shown in Table 5. Besides, we collect 13,117 law articles as our law article corpus.

4.3 Main Results

Table 2 and Table 3 present the main experimental results on SimuCourt and our RareCourt.

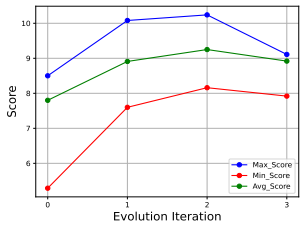

Figure 4: Score of lawyer agent by fine-tuning rounds

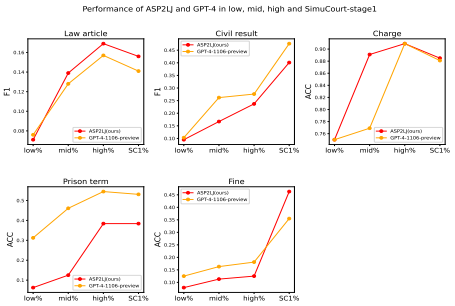

Figure 5: Models' performance between rare cases and common cases. Rare cases are divided into low, mid, and high, which represents their rarity. "Low" means the case is rare, while "high" indicates it is common.

4.3.1 Main Results of SimuCourt Dataset

As Table 2 shows, our method outperforms other methods overall. Compared with the vanilla models, 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 retrieval. Although our method brings large improvement, about 0.148 in criminal fine compared with vanilla Qwen-1.5-7B-Chat, we don't make obvious progress relative to GPT-4. 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 GPT-3.5. However, GPT-4 still surpasses us by 10% relatively. 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 15%, and exceeds that of GPT-3.5 and GPT-4 by 10% and 8% respectively. However, the highest score is just 23.1%

Model	Legal Articles		
	P	R	F
no argument	0.66	0.20	0.30
Qwen1.5-7B	0.76	0.23	0.34
GPT-4o	**0.76**	**0.26**	**0.37**

Table 6: Randomly sample 60 cases' legal articles generation performance between different arguments generated by different models

Model	Legal Articles			Civ. and Adm.			Criminal		
	P	R	F	P	R	F	Cha.	Term	Fine
iter-0	0.41	0.12	0.18	0.26	0.36	0.28	0.83	0.06	0.05
iter-1	0.41	0.12	0.17	**0.33**	0.38	**0.34**	0.90	**0.10**	0.06
iter-2	**0.41**	**0.13**	**0.20**	0.30	**0.40**	0.32	**0.95**	0.06	**0.06**

Table 7: Results of 120 Cases sampled from SimuCourt

Model	Articles(F1)	Judgement Results(F1)			
		Civil	Charge	term	Fine
LawArgueAgent	12.7	16.7	85.0	16.1	10.7
w/o Court argument	11.0	14.8	79.8	13.5	9.3
w/o Lawyer Evolution	11.4	15.6	82.3	14.6	9.8
w/o Retriever	9.4	12.5	77.9	13.0	5.5

Table 8: Ablation Experiment on RareCases

in F1, which indicates that Legal Articles Prediction is the most difficult during such tasks, and all the current methods, including ours, need further improvement.

4.3.2 Main Results of RareCases Dataset

As shown in Table 3, in prison term and fine tasks, LawArgueAgent's score are much lower than GPT-4, GPT-3.5, and AgentsCourt, which shows that maybe Qwen1.5-7B-Chat is not sensitive to numbers. Compared to SimuCourt, the performance totally declines in RareCases. For example, in civil and administrative tasks, the highest score is just 18.6%, which is much lower than 47.6%, which leaves large room to advance. It is obvious that all the models perform worse in our RareCases, and the long-tail distribution truly obstructs the development of legal AI.

4.4 Ablation and Analysis

Ablation As demonstrated in the table 8, retrieval plays a pivotal role in the task of legal article generation, enhancing the F1 score by 3%. In terms of judgment, the retrieval of precedents can also assist the model in adjudicating cases. Furthermore, the analysis of the original case arguments 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 5, the rarer the data, the worse the model performs, which indicates that the model's capability to handle rare cases is insufficient and there is large room for improvement. Basically, the performance of rare cases drops significantly compared to common cases, especially for cases with a low level of rarity.

Fine-tune Iteratively We generated 1,000 cases with GPT-4o for argument. Initially, Qwen1.5-7B-Chat generated arguments for each case, creating 1,000 records. In a subsequent round, the fine-tuned Qwen1.5-7B-Chat generated further arguments for these same cases. The evolved agents' performance was then evaluated on 50 true cases, where we recorded the highest, lowest, and average argument scores. As illustrated in Figure 5, as the tuning iterations progress, all three categories of scores have improved and gradually stabilized. However, in iteration 3, the assessment score declines, but is still better than the vanilla model. In Table 6, it is observed that the court arguments can enhance the Legal Judgment Prediction (LJP) ability of judge agents. Furthermore, GPT-4o demonstrates the capacity to generate more compelling arguments compared to smaller models such as Qwen1.5-7B-Chat. Table 7 illustrates the impact of iterative evolution. While Iteration 2 does not outperform Iteration 1 across all tasks, both Iteration 1 and Iteration 2 demonstrate substantial improvements over the no-iteration baseline.

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 in LJP. Besides, the proposed dataset, RareCases, also indicates that there is still an improvement room in the LJP tasks.

References

James M Anderson and Paul Heaton. 2012. How much difference does the lawyer make: The effect of defense counsel on murder case outcomes. *Yale LJ*, 122:154.

Marius Büttner and Ivan Habernal. 2024. Answering legal questions from laymen in German civil law system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2027, St. Julian's, Malta. 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.

Xi Chen, Mao Mao, Shuo Li, and Haotian Shangguan. 2025. Debate-feedback: A multi-agent framework for efficient legal judgment prediction.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model.

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.

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. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 648–664.

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.

Saiji Fu, Haonan Wen, Xiaoxiao Wang, and Yingjie Tian. 2025. Self-improved multi-view interactive knowledge transfer. *Information Fusion*, 114:102718.

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.

Brandon L Garrett. 2011. *Convicting the innocent: Where criminal prosecutions go wrong*. Harvard University Press.

Samuel R Gross, Barbara O'brien, Chen Hu, and Edward H Kennedy. 2014. Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences*, 111(20):7230–7235.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38.

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.

Wanhong Huang, Yi Feng, Chuanyi Li, Honghan Wu, Jidong Ge, and Vincent Ng. 2024. CMDL: A large-scale Chinese multi-defendant legal judgment prediction dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5895–5906, Bangkok, Thailand. Association for Computational Linguistics.

Ang Li, Yiquan Wu, Ming Cai, Adam Jatowt, Xiang Zhou, Weiming Lu, Changlong Sun, Fei Wu, and Kun Kuang. 2025a. Legal judgment prediction based on knowledge-enhanced multi-task and multi-label text classification. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6957–6970, Albuquerque, New Mexico. Association for Computational Linguistics.

Haitao Li, Qingyao Ai, Qian Dong, and Yiqun Liu. 2024. Lexilaw: A scalable legal language model for comprehensive legal understanding.

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

Ting Li, Lewen Mi, Xiangyu Meng, Yongju Jia, Lin Zhao, Qi Zhao, Zihao Wei, Guandong Gao, and Xiangxian Li. 2025b. Addressing long-tailed distribution in judicial text for criminal motive classification: a balanced contrastive learning approach. *EPJ Data Sci.*, 14:14.

Li, Ting, Mi, Lewen, Meng, Xiangyu, Jia, Yongju, Zhao, Lin, Zhao, Qi, Wei, Zihao, Gao, Guandong, and Li, Xiangxian. 2025. Addressing long-tailed distribution in judicial text for criminal motive classification: a balanced contrastive learning approach. *EPJ Data Sci.*, 14(1):14.

Rui Liu, Zhenqi Jia, Feilong Bao, and Haizhou Li. 2025. Retrieval-augmented dialogue knowledge aggregation for expressive conversational speech synthesis. *Inf. Fusion*, 118:102948.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal judgment prediction with multi-stage case representation learning in the real court setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 993–1002. ACM.

Meladel Mistica, Geordie Z Zhang, Hui Chia, Kabir Manandhar Shrestha, Rohit Kumar Gupta, Saket Khandelwal, Jeannie Paterson, Timothy Baldwin, and Daniel Beck. 2020. Information extraction from legal documents: A study in the context of common law court judgements. In *Proceedings of the 18th annual workshop of the australasian language technology association*, pages 98–103.

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

Emily S Taylor Poppe and Jeffrey J Rachlinski. 2015. Do lawyers matter? the effect of legal representation in civil disputes. *Pepp. L. Rev.*, 43:881.

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.

Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, bm25 is a strong baseline for legal case retrieval.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Brian Sheppard and Andrew Moshirnia. 2012. For the sake of argument: A behavioral analysis of whether and how legal argument matters in decisionmaking. *Fla. St. UL Rev*, 40:537.

Jenkin Chan Shiu-fan. 1983. The role of lawyers in the chinese legal system. *Hong Kong LJ*, 13:157.

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.

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020. Distinguish confusing law articles for legal judgment prediction. *arXiv preprint arXiv:2004.02557*.

Aleš Završnik. 2021. Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of criminology*, 18(5):623–642.

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

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.

A Evaluation Prompt

A scoring criterion encompassing citation, refutation, and comprehension is proposed for a comprehensive and objective evaluation of lawyer agents' discourse. Specifically, this criterion assesses:

Citation: The inclusion of relevant articles or precedents.

Refutation: The act of rebutting the opposing counsel's arguments.

Comprehension: The provision of a comprehensive and clarified case description.

As shown in Table 6, 7, and 8, our results demonstrate that the evolved lawyer agents can produce more effective arguments and improve lawyer agents' performance in LJP tasks. We present our scoring criterion in Table 10.

B Generated Case

To verify our generated cases' rationality, we invite three undergraduate students major in law to evaluate the cases' quality. Our criterion is in three dimensions:

Correctness: Make true if the case can happen in the real world.

Reality: Make true if there are relevant laws as a basis.

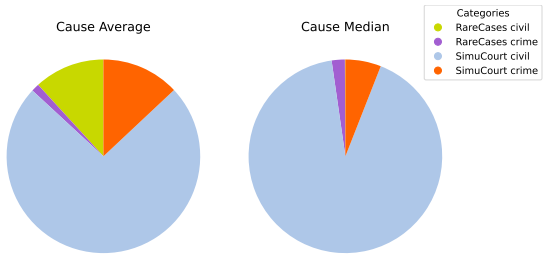

Figure 6: The mean and median number of case causes across SimuCourt and RareCases.

cases	Correctness	Reality	Rationality
Cases sampled 500	0.99	0.99	0.98

Table 9: Quality evaluation of our generated cases

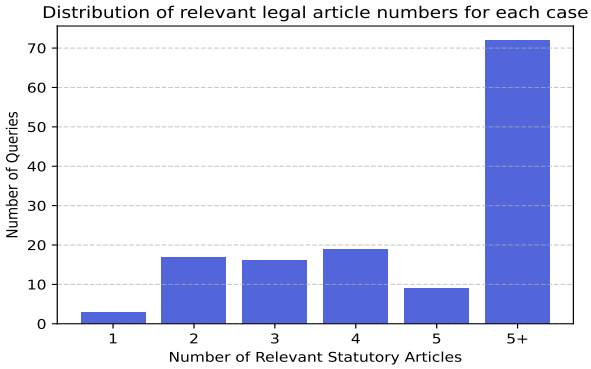

Figure 7: Distribution of relevant legal article numbers for each case.

Rationality: Make true if the plaintiff and defendant's claims are within a reasonable legal framework.
We randomly sample 500 generated cases, and the evaluation results are presented in Table 9.

C Case Distribution

In Figure 6, it is obvious that there is a significant difference in the distribution of case causes between the two datasets. Whether in terms of mean or median, the cases in the RareCases dataset are much rarer than those in the SimuCourt dataset. Correspondingly, the model's performance on RareCases is worse than its performance on SimuCourt.

Besides, as illustrated in Figure 8, the case distribution exhibits a long-tail characteristic. For instance, 'Legal inheritance disputes' accounts for 809,843 instances, whereas 'Disputes over cargo handling contracts' comprises only one. For a more in-depth analysis, we examined the distributions of criminal law, administrative law, and civil law cases from 2012 to 2021, as illustrated in Figure 9, 10, and 11. It is evident that all three domains follow a long-tailed distribution.

D Metric

We have three LJP tasks: law articles, final judgment, and case analysis.
legal articles. We use regular expressions to extract law articles from generated law articles and check whether the extracted articles match our answers or not. Then we calculate the answers' precision, recall, and F1.

$$\mathrm{F1} = \frac{2 \times precision \times recall}{precision + recall} \quad (1)$$

judgment results. For criminal cases, the results include charge, prison term, and fine. We utilize regular expressions to extract these items from generated answer, and calculate the accuracy.

$$\mathrm{ACC} = \frac{\text{Number of correct answers}}{\text{Total number of cases}} \quad (2)$$

For civil and administrative cases, the answers are flexible and hard to extract by regular expressions. So we employ GPT-4o to summarize the answers as several points and judge the number of correct answers. Detailed prompts are presented below.

Prompt 1

Please organize the given text into the required format.

Example 1: Current text: The judgment is as follows: Defendant should return the loan of 200000 yuan to Plaintiff; Defendant shall pay interest during the period of fund occupation at an annual rate of 6% from December 20, 2021 to October 19, 2023; The defendant shall bear all the litigation costs of this case. The above is the final judgment of this court. The defendant is requested to fulfill the repayment obligation within the time limit given in the judgment and pay interest and litigation costs in accordance with the law.

Output list: "Result 1": "The defendant should return the loan of 200000 yuan to the plaintiff", "Result 2": "The defendant should pay interest on the funds during the occupation period at an annual interest rate of 6% from December 20, 2021 to October 19, 2023"

Example 2: ...

Example 3: ...

Please organize the following content:

Current text: <RAW-RELUSTS>

Output List:

Prompt 2

Please compare the candidate's answer with the reference answer to determine if the answer is correct. No explanation is needed, and the result can be directly output in JSON structure

Example 1: Current text: Reference answers: "Result 1": "The defendant should return the loan of 200000 yuan to the plaintiff", "Result 2": "The defendant should pay interest on the capital occupation period at an annual interest rate of 6% from December 20, 2021 to October 19, 2023"

Candidate answers: "Result 1": "The defendant should pay interest on the capital occupation period at an annual interest rate of 6% from December 20, 2021 to October 19, 2023", "Result 2": "The defendant should return the loan of 10000 yuan to the plaintiff"

Output list: Result 1: 0, Result 2: 1

Example 2: ...

Example 3: ...

Please organize the following content and output it in JSON structure:

Current text: <RAW-RELUSTS>

Output list: "Result 1":"", "Result 2":"",...

case analysis To verify agents' comprehension of the legal cases, we instruct agents to generate case analysis and invite three law school undergraduate students to assess. The scoring criterion is the same as AgentsCourt (He et al., 2024): 1) Correctness: Mark true if and only if the analysis is satisfying and considers all parties involved. 2) Logicality: Mark false if the analysis contains any illogical or untrue reasoning. 3) Concision: Mark true if the analysis covers all necessary information without any extra information.

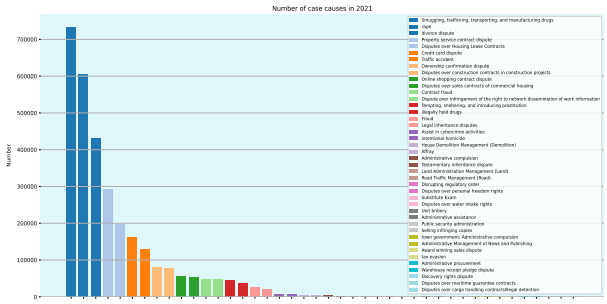

Figure 8: Number of sampled cases in 2021.

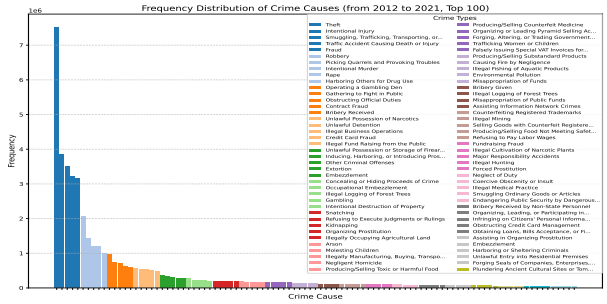

Figure 9: Frequency Distribution of Crime Causes (from 2012 to 2021, Top 100)

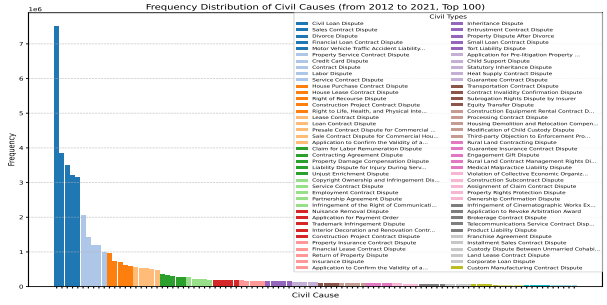

Figure 10: Frequency Distribution of Civil Causes (from 2012 to 2021, Top 100)

E Argument Evaluation Score

We present our evaluation score in Table 10.

Model
s1: Did the content of the argument cite legal provisions and cases, and is the relevance of legal articles and cases to this case significant? Can you provide effective assistance to the judge's judgment, etc., score 0-5 points s2: Does it include the explanation of our viewpoint and the rebuttal of the other party's viewpoint. Is our viewpoint reasonable? Is the rebuttal reasonable and well founded? Is there a correct judgment on the current situation of the case? Whether it is conducive to providing judges with a comprehensive and effective perspective, etc., score 0-5 points; s3: After a round of argument, can we give the outside world a comprehensive understanding of the case? Can summarize the plot of the case, give the judge room to make judgments, etc., score 0-5 points **Evaluation criteria**: s1: Did the content of the argument cite legal articles and cases, and is the relevance of legal articles and cases to this case significant? Can you provide effective assistance to the judge's judgment. 0.5-1 points: Refers to some relevant laws and cases, which can provide reference for the judge's judgment to a certain extent, but not comprehensive enough. 1-1.5 points: Some relevant laws and cases were cited, but it was not comprehensive or in-depth enough. 1.5-2.5 points: In addition to citing some legal articles and cases, the interpretation of applicability is also relatively in-depth, which is more helpful for judges to make judgments. 2.5-4 points: Many relevant laws and cases have been cited, which has a high relevance to this case and a clear interpretation, providing effective assistance for the judge's judgment. 4-5 points: Accurately cited a large number of highly relevant laws and cases, providing comprehensive, in-depth, and highly valuable references for judges' judgments. s2: Does it include the explanation of our viewpoint and the rebuttal of the other party's viewpoint? The viewpoint needs to be reasonable, and the rebuttal needs to be well founded in order to provide the judge with an effective perspective. 0-0.5 points: Explained our viewpoint and provided some rebuttal to the other party's viewpoint, but there are deficiencies in the strength, rationality, and basis of the viewpoint and rebuttal. 0.5-1.5: We have clearly stated our viewpoint and have a certain degree of counterattack against the other party, but the strength of the judge's decisive judgment is not strong enough. 1.5-3 points: Clearly and reasonably presented our viewpoint, and strongly refuted the other party's viewpoint with sufficient evidence. 3-4.5 points: The explanation of our viewpoint has basically gained accurate and strong persuasiveness, comprehensively elaborated our viewpoint, and provided most complete and reliable rebuttals to the other party's viewpoint. 4.5-5 points: Accurately, deeply, and persuasively presented our viewpoint, and provided a comprehensive, powerful, and well founded rebuttal to the other party's viewpoint. s3: After a round of argument, can we give the outside world a comprehensive understanding of the case? Being able to summarize the plot of the case and provide the judge with room for judgment. 0-0.5 points: Described some of the case plot, but there are some key information omissions that have certain limitations on the judge's judgment. 0.5-1.5 points: A comprehensive description of the case's plot can provide the outside world with a clear understanding of the case and provide sufficient judgment space for the judge. 1.5-2.5 points: It basically covers the details of the case, and the outside world has a basic understanding of the case. The judge can basically judge the outcome of the case. 2.5-3.5: Relatively complete, accurate, and comprehensive description of the case plot, providing the outside world with a thorough understanding of the case and offering direction for the judge's judgment. 3.5-5 points: A complete, accurate, and comprehensive description of the case plot provides the outside world with a very thorough understanding of the case, providing judges with broad and sufficient judgment space.

Table 10: The prompt for lawyer agent to evaluate its argument(translated from Chinese)

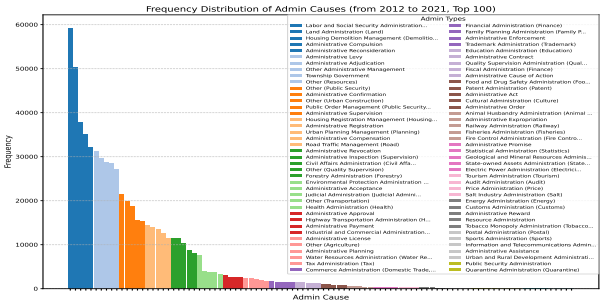

Figure 11: Frequency Distribution of Admin Causes (from 2012 to 2021, Top 100)

F Generated Case

A generated case is presented in Table 11, which consists of the case fact, the plaintiff's indictment, and the defendant's plead.

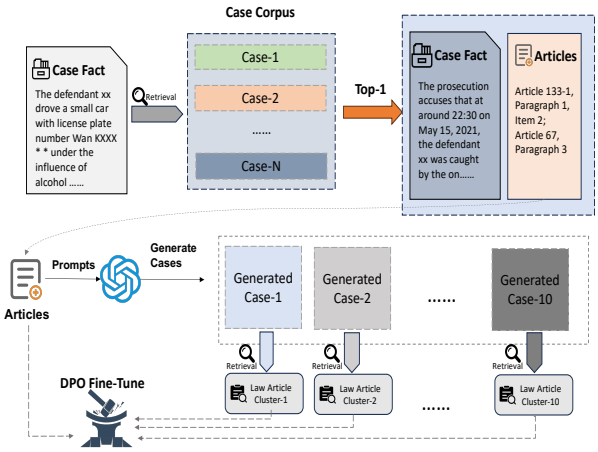

Figure 12: The pipeline of retriever's DPO fine-tuning.

Law Articles	**Item**	**Content**
Article 465	Case fact	On August 15, 2023, the plaintiff Li signed a housing lease contract with the defendant Zhang, agreeing that Li would lease a property located in Haidian District, Beijing to Zhang for a period of one year, with a monthly rent of 5000 yuan. According to the contract, Zhang was required to pay the monthly rent before the 5th of each month...
	plantiff's indictment	The plaintiff Li requests the court to order the termination of the lease contract between him and the defendant Zhang, and demands that Zhang pay a total of 15000 yuan in rent arrears (from October to December 2023), as well as bear the litigation costs of this case. The facts and reasons are: ...
	defendant's plead	The defendant Zhang argued that due to poor management of the company, the defendant is currently facing financial difficulties and is unable to pay rent temporarily. The defendant has communicated with the plaintiff and hopes to delay payment. They are also willing to make up for the overdue rent in one go after the Spring Festival. The defendant requests the court to consider the defendant's actual difficulties, and to give the defendant a lenient treatment and a certain grace period.

Table 11: An example of a generated case(translated from Chinese).