LawLLM: Intelligent Legal System with Legal Reasoning and Verifiable Retrieval

Shengbin Yue[1†], Shujun Liu[1†], Yuxuan Zhou[1†], Chenchen Shen[1†], Siyuan Wang[1], Yao Xiao[4], Bingxuan Li[1], Yun Song[5], Xiaoyu Shen[3], Wei Chen[2], Xuanjing Huang[1], and Zhongyu Wei[1⋆]

[1] Fudan University, Shanghai, China
[2] Huazhong University of Science and Technology, Wuhan, China
[3] Eastern Institute of Technology, Ningbo, China
[4] New York University Shanghai, Shanghai, China
[5] Rule of Law Institute, Northwest University of Political and Law, Xian, China
{sbyue23,ccshen22,yxzhou23,bxli16}@m.fudan.edu.cn, 1171991@s.hlju.edu.cn, lemuria_chen@hust.edu.cn, yx2436@nyu.edu, xyshen@eitech.edu.cn
{shujunliu20,wangsy18,xjhuang,zywei}@fudan.edu.cn

Abstract. We propose LawLLM, an LLM-powered intelligent legal system featuring on (1) *Versatile Services*: LawLLM provides a versatile diverse range of services through its multi-task capabilities; (2) *Legal Reasoning*: It is fine-tuned on supervised instruction data curated with legal syllogism prompting, enabling LawLLM to develop stronger legal reasoning capabilities based on clear judicial logics; (3) *Verifiable Retrieval*: with verifiable labels, LawLLM can first distinguish relevant external knowledge, then incorporate and finally validate it, enhancing the quality and actuality of model output. A comprehensive legal benchmark, Law-Eval, is further constructed to evaluate intelligent legal systems from both objective and subjective dimensions. Experiments demonstrate the effectiveness of our system in serving various users across diverse legal scenarios. The detailed resources are available at *https://github.com/FudanDISC/DISC-LawLLM*.

Keywords: Intelligent Legal System · Legal Reasoning · Verifiable Retrieval · Large Language Model.

1 Introduction

With the rise of legal artificial intelligence (LegalAI) [1, 40, 12], the legal domain is undergoing significant transformation. Through automating legal tasks including legal information extraction [4, 35], interactive argument pair extraction [15, 37], case retrieval [19, 36], judgment prediction [14, 27], and legal question answering [17, 41], intelligent legal systems can offer broad applications that benefit different groups of the population.

Previous endeavors have primarily concentrated on individual tasks, involving the creation of task-specific datasets and algorithms, which has restricted the wider accessibility of intelligent legal systems. In the legal domain, tasks such as legal information

⋆ Corresponding Author
† These authors contributed equally to this work.

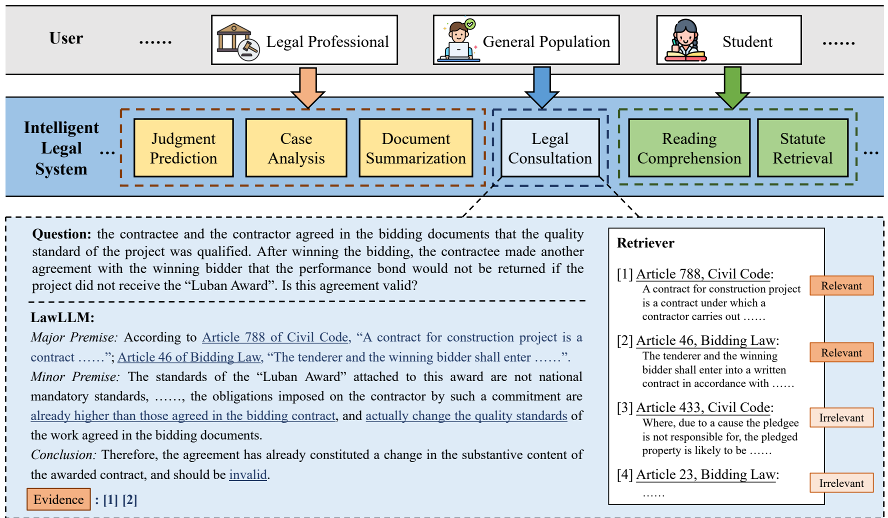


Fig. 1: Overview of LawLLM serving different legal needs. For an example of legal consultation, LawLLM first selects relevant segments from the retrieved statutes, then performs legal reasoning (*i.e.*, the statute is the major premise, the fact in the question is the minor premise, and the final inference is the conclusion), and generates validated identifiers of the selected statutes.

extraction, argument pair extraction, case retrieval, judgment prediction, and legal question answering are often intertwined and necessitate a seamless flow of information and analysis. By integrating multi-task functionality, an intelligent legal system can efficiently handle these interconnected tasks without requiring users to switch between disparate applications, thereby streamlining workflows and enhancing overall productivity. Recent advancements in large language models (LLMs) [5, 21, 22, 29] have shown remarkable instruction-following ability across varied domains, offering tremendous potential in developing multi-task legal assistants. While initial progress has been made [13, 8] by fine-tuning general LLMs to utilize legal knowledge for basic question answering, the scope of legal services is considerably more complex and extends beyond simple dialogue.

An example of legal consultation is illustrated in Fig. 1, where the intelligent legal system can leverage relevant information from retrieved knowledge to mine facts from the inquiry, and deduce the conclusion to provide legal service. This highlights two primary challenges for building reliable legal system: (1) The need for advanced legal reasoning capabilities, which involves nuanced and context-dependent interpretations in applying complex legal principles, statutes, and case law to specific scenarios in a manner that aligns with established legal frameworks. For example, all legal response logics in jurisprudence should follow the structure of legal syllogism [25, 16], which consists of a major premise representing the legal proposition, a minor premise symbolizing the factual proposition, and a conclusion representing the judgment. (2) The requirement to robustly leverage retrieved external legal knowledge. LLMs struggle to capture long-tail knowledge and require expensive costs to be kept updated [28], which can be mitigated

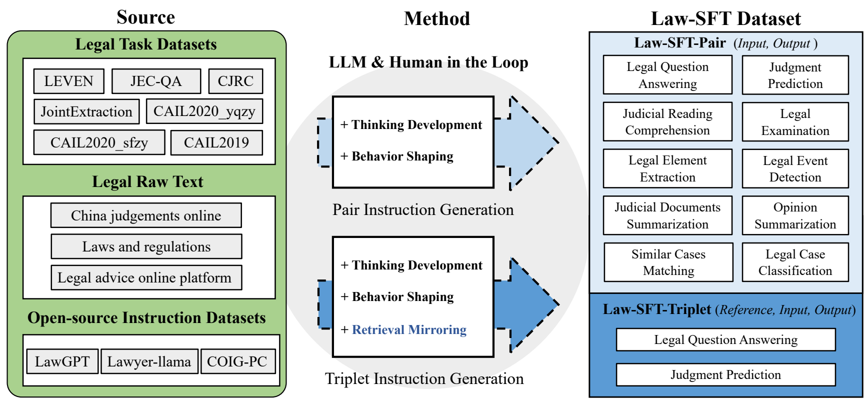


Fig. 2: Construction of Law-SFT Datasets. By distinct construction strategies, general LLMs and human labeller are involved to generate instructions in two forms subsets.

by explicitly decoupling knowledge retrieval from LLMs. Nevertheless, the retrieved content may include distractors, *i.e.*, irrelevant chunks. Even with ideally retrieved external knowledge, LLMs may not consistently utilize this information to reason. This exacerbates the model's hallucinations in legal scenarios, leading to unreliable outputs. Given the complexity and specificity of legal information, the ability to retrieve and integrate legal knowledge (first distinguish, then incorporate, and finally verify) from diverse sources is essential for ensuring that legal outputs are accurate, reliable, and aligned with established legal frameworks.

To this end, we present LawLLM, an LLM-powered intelligent legal system with legal reasoning and verifiable knowledge retrieval capability. In contrast to existing legal LLMs that primarily focus on basic dialogue, our model integrates three key advantages: (1) *Versatile services*: LawLLM provides a versatile diverse range of services via fine-tuning on an ensemble of over 10 types of legal tasks, encompassing a wide array of legal scenarios; (2) *Legal reasoning*: When constructing the supervised data, we guide LawLLM to think jurisprudentially through behavior shaping and thinking development, enabling it to enhance its legal reasoning capabilities based on clear judicial logic. (3) *Verifiable retrieval*: LawLLM has the ability to discern relevant fragments from the retrieved knowledge, faithfully incorporate this information for reasoning, and ultimately generate identifiers for the used chunks to further enhance verifiability. [6].

In addition, we design a legal benchmark, Law-Eval, to provide a comprehensive assessment of intelligent legal systems from both objective and subjective dimensions. The objective assessment measures models' grasp of legal knowledge and reasoning capabilities through single and multiple-choice questions across three difficulty levels (*Easy*, *Normal* and *Hard*). For the subjective assessment, we utilize GPT-3.5 as an arbitrator to assess the quality of long-form legal generations in terms of *Accuracy*, *Completeness* and *Clarity*. Further, we set up retrieved external references for each case

[6] We curate the training data to guide LawLLM to generate identifiers (*e,g.* [1] [2]) in the verifiable label ("[Evidence]") by evaluating each reference, and to generate "[No]" when references are all irrelevant, which is important for the legal scenario that requires solid evidence support.

Table 1: Statistics of Law-SFT Dataset.

Dataset	Task	Size	Topic
Law-SFT-Pair	Legal Element Extraction	32K	Legal Tools
	Legal Event Detection	27K	
	Legal Case Classification	20K	
	Similar Cases Matching	8K	
	Documents Summarization	9K	
	Public Opinion Summarization	6K	
	Legal Question Answering	93K	Legal Reasoning
	Judgement Prediction	11K	
	Document Reading Comprehension	38K	
	Judicial Examination	12K	
Law-SFT-Triplet	Judgement Prediction	16K	Legal Reasoning
	Legal Question Answering	23K	
General	Alpaca-GPT4	48K	General
	Firefly	60K	
Total		403K	

to assess the ability to utilize external knowledge. For each evaluation sample, a ground truth is provided to the arbitrator to reduce potential biases during the assessment phase.

Experimental results on Law-Eval and LawBench [11] reveal that LawLLM significantly outperforms existing legal and Multilingual LLMs with the same parameters. Even compared to GPT-3.5-turbo [20], LawLLM demonstrates superior performance across the majority of objectively evaluated subjects on Law-Eval, and beat it on the average performance (zero-shot) of 20 tasks on LawBench. Experiments demonstrate that equipped with legal reasoning and verifiable retrieval capabilities, our LawLLM consistently produces more reliable responses.

2 Method

2.1 Overview

As illustrated in Fig. 2, to train arbitrary LLM to become an intelligent legal system, we create a high-quality supervised fine-tuning dataset, Law-SFT, consisting of two subsets: *Law-SFT-Pair* and *Law-SFT-Triplet*. The Law-SFT-Pair subset aims to introduce multi-legal task and reasoning abilities to the LLM, while the Law-SFT-Triplet subset is intended to enhance the model's capacity to utilize external knowledge by mimicking the process of retrieval, incorporation, and verification.

2.2 Data Sources

We obtain original samples from the following three sources: 1) Public NLP legal task datasets. Public datasets cover a range of legal NLP tasks and provide human annotations which can be utilized to generate high-quality instructions, including legal information extraction (LEVEN [35] and JointExtraction [34]), legal text summarization

(CAIL2020-sfzy and -yqzy [7]), legal question answering (JEC-QA [41]), legal reading comprehension (CJRC [31]), similar cases matching [33] and judgment prediction (CAIL2018 [32]). 2) Legal raw text. To include more legal service scenarios, we crawl up an expansive collection of real-world legal text to construct instruction data, including consultation data from judicial advisory websites, Chinese laws and regulations, typical cases, judicial verdicts and law-related examinations. 3) Open-source instruction datasets. we also borrow some samples from recently opened instruction datasets and reshape them according to our expectations, including Lawyer-LLaMa [13], LawGPT-zh [8] and COIG-PC [38].

2.3 Law-SFT Generation

Legal intelligent applications in different scenarios require combinations of multiple fundamental capabilities of legal text understanding and generating. Therefore, we construct instruction samples converging more than 10 tasks, ensuring coverage of diverse scenarios. Detailed statistics of our datasets are provided in Table 1. General LLMs and human labeller are involved to re-construct original samples to generate instructions in two forms of the pair (<input, output>) and the triplet (<reference, input, output>).

Pair Instruction Generation To construct Law-SFT-Pair, we initially employ rule-based methods to clean the collected source data, and transform it into "input-output" pairs (x, y) according to the task. As shown in Table 1, we categorize these pairs into two topics, legal tools and legal reasoning. Legal tools emphasize legal text handling capability (*e.g.*, summarization, extraction, classification), so we manually set various prompts for every task in this topic according to its attributes. For legal reasoning (*e.g.*, question answering, judgement prediction, reading comprehension), these pairs (x, y) usually are rigid and noisy in linguistic patterns, and the expression styles can differ across sources. Therefore, we reconstruct these topic pairs with legal language and logic by using the following two methods with the assistance of general LLM.

Behavior Shaping. In the legal syllogism framework, the major premise is the applicable law, while the minor premise is pertinent facts, and the conclusion is the final judgment. This constitutes a foundational legal reasoning process for judges. Every case can culminate in a conclusion articulated through a syllogism, as outlined below:

- Major premise: laws
- Minor premise: pertinent facts
- Conclusion: judgment

Inspired by legal syllogism prompting [16] and self-construct [30], we utilize LLMs to refine output responses y for consistency with legal syllogism. We prompt GPT-3.5-turbo to make (x, y) transform to (x, y_b) with devised instructions, to ensure that each conclusion should be drawn from laws and pertinent facts. Appendix A shows the examples of instructions.

Thinking Development. Chain of Thought (CoT) in training data mixture has been shown to significantly enhance the reasoning ability of models [6]. To further cultivate

[7] https://github.com/china-ai-law-challenge/CAIL2020
[8] https://github.com/LiuHC0428/LAW-GPT

Table 2: Statistics of Law-SFT-Triplet subset with distractors.

Task	Distractors (d)				
	0	1	2	3	All
Judgement Prediction	1970	5635	2842	328	1219
Legal Question Answering	1769	14626	9158	1672	1280
Total	3739	20261	12000	2000	2499

legal reasoning, we devise law-specific chains of thought, termed LCoT, and then we combine it with input to enforce the model conduct legal syllogism to derive the answer. Taking the judgement prediction task as an example, LCoT incorporates input x that transform (x, y) into (x_t, y) as follows:

> In the legal syllogism, the major premise is articles of law, the minor premise is the facts of the case, and the conclusion is the judgment of case.
> Case: x
> Let us use legal syllogism to think and output the judgment:

Triplet Instruction Generation The verifiable retrieval process aims to allow the model to first select relevant fragments from the retrieved external candidates, then faithfully utilize these fragments, and finally generate the identifier of the used fragment, which is crucial for legal scenarios. To generate supervised instruction triplets for precisely mirroring this process, we create a subset called Law-SFT-Triplet by inserting a verifiable label "[Evidence]" e. When retrieved content is relevant, the label e outputs its identifier, otherwise, the label e outputs "[NO]".

As shown in Table 1, we use the source data with the referenced statute to build this triplet $(r, x, y \star e)$. Specifically, we first create the original triples <reference, input, output> (r, x, y) from source data. Then, each reference is extracted by heuristic rules to compose reference sets $r_{1,\ldots,k}$ that are all related to input x and output y. Every legal statute r_k is a retrieval chunk. Next, using methods described in Sec. 2.3, we process the original data to obtain the input and output. Subsequently, we tag each relevant reference r_k in the set and add distractors r_d, *i.e.* irrelevant reference by a certain percentage. In our setting, the maximum number of references r_{k+d} is 11. The complete set r_{k+d} aims to mirror the inference-time handling of retrieval and return top-$\widetilde{k}$ text chunks from external knowledge. The details of distractor number in the set are shown in Table 2. Finally, we enable the verifiable label e to output the identifier of each r_k in order.

2.4 Law-SFT Fine-Tuning

As shown in Fig. 3, Law-SFT enables LLM into an intelligent legal system with legal reasoning and verifiable retrieval ability. We form our training process using two steps, legal reasoning fine-tuning and retrieval augmentation fine-tuning.

Legal Reasoning Fine-Tuning Law-SFT-pair aims to cultivate the model's multi-task capability and legal thinking. Given the Law-SFT-pair data, $\mathcal{D}_{pair}$, we initialize a

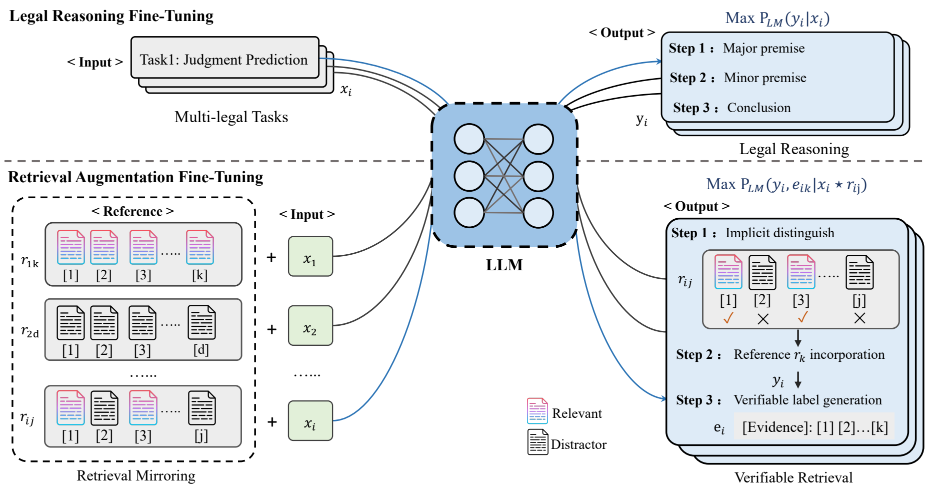


Fig. 3: Overview of the fine-tuning process, where Law-SFT-Pair encourages legal reasoning fine-tuning and Law-SFT-triplet encourages retrieval augmentation fine-tuning.

pre-trained LLM and train it on $\mathcal{D}_{pair}$. For each example $(x_i, y_i) \subset \mathcal{D}_{pair}$, we use a standard conditional language modeling objective, maximizing likelihood:

$$\mathcal{L}\left(\mathcal{D}_{pair}\right) = -\sum_{i} \log p_{LM}\left(y_i \mid x_i\right), \tag{1}$$

After legal reasoning fine-tuning, the model acquires proficiency not only in navigating the complexities of legal language across various legal contexts but also in executing the process of deductive reasoning within a framework reflective of legal thinking.

Retrieval Augmentation Fine-Tuning To further empower LLM with the ability to utilize retrieved information and generate verifiable labels, we use Law-SFT-triplet $\mathcal{D}_{triplet}$ to fine-tune the LLM with in-context retrieval augmentation. For each example $\{r_{ij}, x_i, (y_i, e_{ik}) | \, j = 1, ..., k+d\} \subset \mathcal{D}_{triplet}$, we create a separate fine-tuning example by sequentially prepending r_{ij} to the instructions x_i as a context. We train the LLM on the $\mathcal{D}_{triplet}$ with verifiable tokens e_k also using the standard next token objective:

$$\mathcal{L}\left(\mathcal{D}_{triplet}\right) = -\sum_{i}\sum_{j} \log p_{LM}\left(y_i, e_{1,\ldots,k} \mid x_i \star r_{ij}\right). \tag{2}$$

Where $\star$ denotes prepending operation. Unlike Legal reasoning fine-tuning (Eq. 1), LLM learns to predict the target output as well as from reference tokens $r_{1,\ldots,k+d}$ to verifiable tokens $e_{1,\ldots,k}$. That is, Law-SFT-triplet forces the model to first implicitly distinguish relevant chunks from external knowledge, then integrate and finally verify them, thus improving the consistency of the model from external knowledge to internal inference.

In the whole implementation, we combine two steps of legal reasoning fine-tuning and verifiable retrieval augmentation fine-tuning. Additionally, we incorporate general

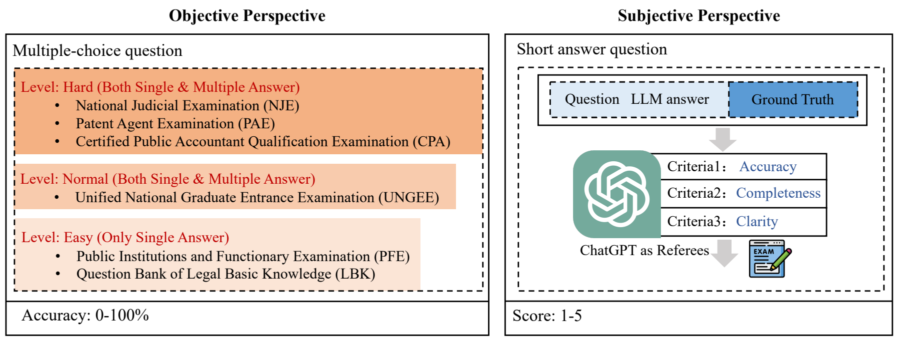


Fig. 4: Overview of Law-Eval Benchmark, assessing systems both objectively and subjectively.

instruction data, alpaca-gpt4-zh [24] and Firefly[9], to enrich the diversity of our training set, mitigating the risk that foundational capability diminishes during the training phase.

3 Law-Eval Benchmark

Drawing inspiration from the structure of the bar examination, as shown in Fig 4, we formulate a robust evaluation framework, Law-Eval Benchmark, providing a comprehensive assessment of systems from both the objective and subjective perspective.

3.1 Objective Evaluation

To objectively and quantitatively assess the legal knowledge and reasoning capabilities of intelligent legal systems, we design an objective evaluation dataset. It consists of multiple-choice questions, and each may have one or multiple correct answers.

As shown in Fig. 4, we collect multi-choice questions from a range of Chinese legal standardized examinations and knowledge contests, including the National Judicial Examination (NJE), Patent Agent Examination (PAE), Certified Public Accountant Qualification Examination (CPA), Unified National Graduate Entrance Examination (UNGEE), Public Institutions and Functionary Examination (PFE) and Question Bank of Legal Basic Knowledge (LBK). To ensure the utmost fairness, the dataset incorporates materials from the preceding five years, the majority of which are not accessible online.

According to content complexity and deduction difficulty, we categorize these questions into three levels: *Hard*, *Normal* and *Easy*. Higher level, *e.g. Hard*, encompasses a broader spectrum of long-tail legal knowledge and necessitate a higher degree of reasoning capabilities. The details of objective evaluation at different levels are shown in Table 3. Objective evaluation can provide a more challenging and reliable measure of whether the model can use its knowledge to reason toward correct answers. We calculate the accuracy to indicate the performance.

[9] https://github.com/yangjianxin1/Firefly

Table 3: Details of Objective Evaluation Dataset, where S and M are shorthand of single-answer and multiple answers, respectively.

Level	Subject	S	M	Total		Level	Subject	S	M	Total
Hard	CPA	197	120	317		Normal	UNGEE	320	87	407
	NJE	537	463	1000		Easy	LBK	275	-	275
	PAE	118	276	394			PFE	170	-	170

3.2 Subjective Evaluation

We further conduct a subjective evaluation to explicitly demonstrate the model's command over legal knowledge and reasoning ability. We adopt a question-answering paradigm, simulating the process of subjective examination questions. Considering data contamination, we manually construct a high-quality test set from legal consultations, justice-related publications, and legal documents, comprising 300 examples. These examples cover scenarios including legal tools, legal consultations, and judgment prediction. In addition, we equipped each example with external knowledge to explore the model's ability to utilize external knowledge, and 60 of them involve distractors.

The traditional metrics employ for generation tasks, such as ROUGE-L [18], [3] and BLUE [23], are not ideally aligned with the nuances of evaluating LLMs. So we design a method to evaluate this subjective quality by eliciting a referee model. Strong LLM judges like GPT-3.5, GPT-4 align well with controlled and crowdsourced human preferences [39]. In our evaluation, GPT-3.5 serves as a referee and performs the evaluation by providing a rating score from 1 (lowest) to 5 (highest) for each of the following three criteria: *accuracy*, *completeness* and *clarity*, where *accuracy* aims to gauge the semantic congruence between the model's outputs and the ground truth; *completeness* focuses on the presence of critical details, such as judicial facts, and compares them with the ground truth to determine if any vital elements are absent; *clarity* aims to assess the legal logic and sentence organization of the generated answers.

To reduce the self-bias of the referee model, we provide the ground truth to the referee as well, enabling them to score according to the ground truth. We repeat the scoring for each question and finally get the average score on different dimensions. Appendix A shows the full list of instructions.

4 Experiments

4.1 Implementation Details

We use 8*A800 GPUs with 80GB memory to train our model on top of the Baichuan-13B-Base model [2], which is an open-source LLM with over 13.2 billion parameters that was trained on 1.4 trillion tokens corpus. The model is trained for 2 epochs with a batch size of 64, the learning rate of 5e-5. We set the maximum token length to be 4096. We use Deepspeed [26] to conduct multi-GPU distributed training, with training precision Bfloat16 enabled.

Table 4: Results compared with LLMs on objective evaluation. Darker (best) to lighter green marks the best of the top three good results. S and M are shorthand of single-answer and multiple answers, respectively. "*" denotes the same base model as us.

Model	Hard						Normal		Easy		Average		
	NJE		PAE		CPA		UNGEE		PFE	LBK			
	S	M	S	M	S	M	S	M	S	S	S	M	Total
Multilingual LLMs													
ChatGLM $_{6B}$	31.66	1.08	27.97	2.90	37.06	13.33	39.69	20.69	37.65	42.91	36.18	4.97	24.66
Llama2 $_{13B}$	27.93	0.00	22.88	2.17	24.87	0.00	32.50	2.30	33.52	28.36	28.75	0.84	18.45
Llama2-chat $_{13B}$	28.49	0.22	26.27	0.36	29.44	0.83	37.50	1.15	45.29	39.27	33.83	0.42	21.50
Chinese-Llama2 $_{13B}$	15.27	1.94	21.18	3.62	28.42	17.50	33.43	19.54	33.52	38.54	26.77	6.02	19.11
Chinese-Alpaca2 $_{13B}$	25.70	10.15	30.51	11.59	32.99	19.17	40.94	21.84	44.12	43.27	34.88	12.79	26.73
Baichuan-chat* $_{13B}$	31.47	10.15	29.66	8.70	35.53	19.17	50.00	27.59	53.12	53.45	41.51	12.48	30.80
GPT-3.5-turbo	36.50	10.58	37.29	**17.03**	42.13	21.67	51.25	28.74	53.53	54.18	44.96	15.54	34.10
GPT-4	**44.32**	6.26	**44.32**	15.57	**57.36**	**35.00**	**66.85**	**45.98**	**70.00**	**61.45**	**55.98**	16.27	**41.33**
Legal LLMs													
LexiLaw $_{6B}$	20.11	7.56	23.73	10.14	24.87	19.17	31.56	16.09	31.76	40.36	27.89	10.57	21.50
LaWGPT $_{7B}$	22.91	6.26	31.36	7.61	25.38	16.67	30.31	13.79	34.71	29.09	27.58	8.67	20.60
Lawyer-LLaMa$_{13B}$	35.75	5.62	32.20	6.52	29.95	13.33	32.50	14.94	39.41	39.64	35.19	7.72	25.05
ChatLaw$_{13B}$	27.56	7.99	31.36	9.42	35.53	11.67	35.62	17.24	42.35	41.09	34.26	9.72	25.20
LawLLM$_{13B}$(OURS)	41.38	**14.90**	40.29	16.30	39.09	20.00	50.19	26.69	55.35	54.73	47.19	**19.87**	37.11

4.2 Baselines

To show comprehensiveness, we focus on comparisons with the following two types of LLMs: 1) The Multilingual LLMs includes Llama2-13B [29], Llama2-chat-13B [29], GPT-4 [21], GPT-3.5-Turbo [20], ChatGLM-6B [10], Chinese-Alpaca2-13B [9], Chinese-Llama2-13B [9] and Baichuan-13B-chat [2]. 2) The legal LLMs include LaWGPT[10], Lawyer-LLaMa [13], ChatLaw [8] and LexiLaw[11].

4.3 Experiments Results on Law-Eval

Comparison in Objective Evaluation. We conduct the objective evaluation in a few-shot setting (4-shot for single-answer questions and 5-shot for multi-answer questions). As shown in Table 4, We can see that 1) LawLLM surpasses almost all competing LLMs across all subjects with different difficulty levels. Even compared with GPT-3.5-turbo [20] and GPT-4 [21], which hold huge parameters, LawLLM shows excellent performance in some subjects, especially equaling GPT-3.5-Turbo and surpassing it in average performance. 2) Compared to the existing legal LLMs, LawLLM surpasses all legal LLMs by a large margin on all subjects. Note that existing legal LLMs are not as good as multilingual LLMs, the reason is likely to be the lack of few-shot instruction following ability during training and their weak base model. 3) For the same base model of Baichuan-chat [2], our lawLLM improved by 5.68%, 7.39% and 6.3% in weighted average correctness of single choice, multiple choice and total. Our experiments on objective evaluation demonstrate that LawLLM has strong legal knowledge and jurisprudential reasoning via the Law-SFT training.

[10] https://github.com/pengxiao-song/LaWGPT
[11] https://github.com/CSHaitao/LexiLaw

Table 5: Results compared with rivaled LLMs on Subjective Evaluation, where ACC, CPL and CLR are the shorthand of Accuracy, Completeness and Clarity, respectively.

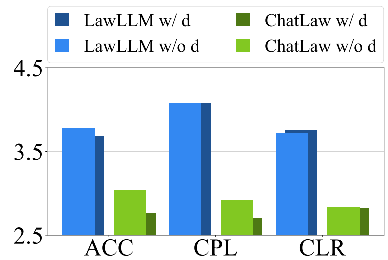
Model	ACC	CPL	CLR	Average
Multilingual LLMs				
ChatGLM $_{6B}$	2.87	2.97	3.37	3.07
Llama2 $_{13B}$	2.00	1.77	1.84	1.87
Llama2-chat $_{13B}$	2.35	2.44	2.97	2.59
Chinese-Llama2 $_{13B}$	2.36	2.36	2.24	2.32
Chinese-Alpaca2 $_{13B}$	2.67	2.89	3.30	2.95
Baichuan-chat* $_{13B}$	2.71	2.86	**3.52**	3.03
Legal LLMs				
LexiLaw $_{6B}$	2.95	2.79	3.01	2.92
LaWGPT $_{7B}$	2.11	1.70	1.66	1.82
Lawyer-LLaMa $_{13B}$	2.77	2.89	3.45	3.04
ChatLaw $_{13B}$	2.94	2.69	2.76	2.80
LawLLM $_{13B}$(OURS)	**3.09**	**2.95**	3.27	**3.10**

Comparison in Subjective Evaluation. In the subjective evaluation, GPT-3.5-turbo-0613 functions as an adjudicator. From Table 5, we can see that LawLLM achieves the best performance on most metrics. We can conclude that: 1) Leveraging the Law-SFT dataset enables LawLLM to generate more reliable responses, leading to the best results in both ACC and CPL. 2) Through the deliberate cultivation of the model's juridical thinking, the responses from LawLLM exhibit superior jurisprudential logic. We note that Baichuan-chat is slightly better than ours in CLR, and the reason is to introduce the Reinforcement Learning from Human Feedback (RLHF) strategy after fine-tuning to further enhance linguistic clarity. Experiments demonstrate that our approach effectively enhances the legal logic and factuality of the model.

Comparison in Retrieval Augmentation. To explore the model's ability to utilize external knowledge, we provide the law as external knowledge $\mathcal{R}$ for each evaluation case on the subjective evaluation. We compare our LawLLM with Baichuan-chat [2] with the same base model and advanced legal model, ChatLaw [7]. From Fig. 5 (a), we can observe that: 1) by external knowledge retrieval, the performance of LLMs can be effectively improved. 2) Our LawLLM can achieve better results compared to Baichuan-chat and ChatLaw in Retrieval Augmentation. 3) Compared to them, external knowledge improves our LawLLM the most. This indicates that by distinguishing, incorporating, and verifying capabilities, our LawLLM can better utilize external knowledge. As shown in Fig. 5 (b), we further explore the effect of distractor (irrelevant knowledge), where w/d denotes external knowledge with distractors, and $\mathrm{w/o}\ d$ denotes all are relevant. we can observe that ChatLaw does not perform well in the presence of distractors, while our LawLLM can still have good performance. When the references are all relevant, ChatLaw also performs poorly, while our performance is still excellent. This indicates that our LawLLM efficiently distinguishes relevant information from irrelevant knowledge and faithfully utilizes it. Overall, by verifiable retrieval capability, the model can better incorporate external knowledge, resulting in more reliable results. This allows our LawLLM to work with arbitrary retrieval modules.

Model	ACC	CPL	CLR	Average
Baichuan-chat*	2.71	2.86	3.52	3.03
Baichuan-chat* w/ $\mathcal{R}$	3.16	4.06	3.37	3.53(↑ 16.5%)
ChatLaw	2.94	2.69	2.76	2.80
ChatLaw w/ $\mathcal{R}$	2.96	2.73	2.84	2.84(↑ 1.4%)
LawLLM	3.09	2.95	3.27	3.10
LawLLM w/ $\mathcal{R}$	**3.70**	**4.07**	**3.72**	**3.83** (↑ 23.5%)

(a)



(b)

Fig. 5: Analysis in Retrieval Augmentation: (a) Comparison Results with LLMs with Retrieval knowledge $\mathcal{R}$. (b) Effects of distractors (d) in Retrieval knowledge $\mathcal{R}$.

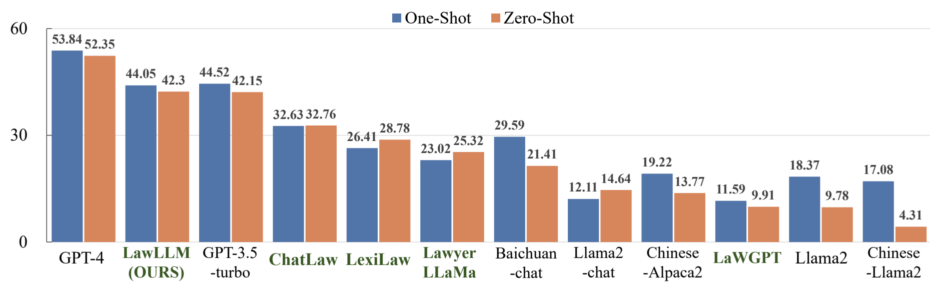


Fig. 6: Average performance (zero-shot and one-shot) of LLMs evaluated on LawBench.

4.4 Experiments Results on LawBench.

LawBench [11] is a contemporaneous benchmark that assesses the LLM's legal competence on 20 tasks covering 5 task types: single-label classification, multi-label classification, regression, extraction and generation. It use traditional metrics, such as F1 and ROUGE-L, to assess the quality of different tasks, we follow the official setup to evaluate our models. Fig. 6 illustrates the average performance of our model in 20 tasks with one-shot and zero-shot. We can observe that 1) Our LawLLM beats all the legal LLMs in both settings by a large margin. 2) Our performance is only inferior to GPT-4 and outperforms GPT-3.5 on zero-shot. 3) Our performance is significantly superior to the Baichuan-chat with the same base model as ours. The observation can conclude that Law-SFT makes the model capable of handling a wide range of legal tasks. Overall, the performance on Lawbench demonstrates that our model has strong zero-shot and few-shot capabilities on a variety of legal tasks, which proves the effectiveness and the advancement of our approach.

5 Related Work

Large Language Models (LLMs) have achieved astounding performance on different conventional linguistic tasks, demonstrating powerful generality. However, due to the lack of domain knowledge and special training paradigms, these generic LLMs have proven to be unsuitable for some domain-specific tasks, such as law. Currently, some initial progress has been made in legal LLMs. Specifically, LaWGPT enhances legal language understanding through pre-training and fine-tuning along with domain-specific terminologies. LexiLaw leverages various fine-tuning methods using datasets in the legal domain. Lawyer-LLaMa [13] and ChatLaw [8] inject domain knowledge during continuous training and incorporate retrieval modules to enhance factuality. Generally speaking, these methods focus on dialogue in legal scenarios, and the LLMs are not inherently trained to incorporate retrieved content. Different from the previous models, our proposed LLM aims to build an intelligent legal system that provides a wide range of legal services by integrating advanced capabilities. It is designed to excel in multitasking, legal reasoning, and knowledge-base-driven retrieval.

6 Conclusion

In this paper, we propose LawLLM, an intelligent legal system with legal reasoning and verifiable knowledge retrieval capability. With the constructed Law-SFT driven, our LawLLM can provide a wide range of legal services. Law-SFT adopts legal syllogism to construct supervised fine-tuning datasets with legal reasoning capability. Law-SFT enables the model with verification retrieval capability through the process of mirroring retrieval, distinguishing, incorporating and validating. A comprehensive legal benchmark, Law-Eval, is presented to evaluate intelligent legal systems from both objective and subjective dimensions. Experimental outcomes on Law-Eval and LawBench show that LawLLM markedly excels over current legal and multilingual LLMs of equal size. It surpasses GPT-3.5-turbo in most Law-Eval subjects and average performance across 20 tasks on LawBench. The results affirm LawLLM's enhanced reliability and superior legal reasoning and retrieval capabilities.

Acknowledge

We thank the EIT and IDT High Performance Computing Center for providing computational resources for this project.

References

1. Atkinson, K., Bench-Capon, T., Bollegala, D.: Explanation in ai and law: Past, present and future. Artificial Intelligence **289**, 103387 (2020)
2. Baichuan-inc: Baichuan-13b. https://github.com/baichuan-inc/Baichuan-13B (2023)
3. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
4. Bommarito, M., Katz, D.M., Detterman, E.: Lexnlp: Natural language processing and information extraction for legal and regulatory texts. Research Handbook on Big Data Law (2018)

5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
7. Cui, J., Li, Z., Yan, Y., Chen, B., Yuan, L.: Chatlaw. https://github.com/PKU-YuanGroup/ChatLaw (2023)
8. Cui, J., Li, Z., Yan, Y., Chen, B., Yuan, L.: Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092 (2023)
9. Cui, Y., Yang, Z., Yao, X.: Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177 (2023)
10. Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: Glm: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 320–335 (2022)
11. Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Shen, Z., Ge, J.: Lawbench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289 (2023)
12. Ge, J., Huang, Y., Shen, X., Li, C., Hu, W.: Learning fine-grained fact-article correspondence in legal cases. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3694–3706 (2021)
13. Huang, Q., Tao, M., An, Z., Zhang, C., Jiang, C., Chen, Z., Wu, Z., Feng, Y.: Lawyer llama technical report. ArXiv **abs/2305.15062** (2023)
14. Huang, Y., Shen, X., Li, C., Ge, J., Luo, B.: Dependency learning for legal judgment prediction with a unified text-to-text transformer. arXiv preprint arXiv:2112.06370 (2021)
15. Ji, L., Wei, Z., Li, J., Zhang, Q., Huang, X.J.: Discrete argument representation learning for interactive argument pair identification. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5467–5478 (2021)
16. Jiang, C., Yang, X.: Legal syllogism prompting: Teaching large language models for legal judgment prediction. arXiv preprint arXiv:2307.08321 (2023)
17. Kien, P.M., Nguyen, H.T., Bach, N.X., Tran, V., Le Nguyen, M., Phuong, T.M.: Answering legal questions by learning neural attentive text representation. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 988–998 (2020)
18. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
19. Ma, Y., Shao, Y., Wu, Y., Liu, Y., Zhang, R., Zhang, M., Ma, S.: Lecard: a legal case retrieval dataset for chinese law system. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. pp. 2342–2348 (2021)
20. OpenAI: Chatgpt: Optimizing language models for dialogue (2022), https://openai.com/blog/chatgpt/
21. OpenAI: Gpt-4 technical report (2023)
22. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
24. Peng, B., Li, C., He, P., Galley, M., Gao, J.: Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277 (2023)

25. Posner, R.A.: The problems of jurisprudence. Harvard University Press (1990)
26. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3505–3506 (2020)
27. Song, Y., Wei, Z.: Inferring association between alcohol addiction and defendant's emotion based on sound at court. Frontiers in Psychology **12**, 669780 (2021)
28. Tirumala, K., Markosyan, A., Zettlemoyer, L., Aghajanyan, A.: Memorization without overfitting: Analyzing the training dynamics of large language models. Advances in Neural Information Processing Systems **35**, 38274–38290 (2022)
29. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
30. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560 (2022)
31. Wu, D., Wang, S., Liu, T., Huo, T., Hu, Z., Wang, H., Liu, Z.: Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings. vol. 11856, p. 439. Springer Nature (2019)
32. Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., et al.: Cail2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018)
33. Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Zhang, T., Han, X., Hu, Z., Wang, H., et al.: Cail2019-scm: A dataset of similar case matching in legal domain. arXiv preprint arXiv:1911.08962 (2019)
34. Yang, B., Mitchell, T.: Joint extraction of events and entities within a document context. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 289–299 (2016)
35. Yao, F., Xiao, C., Wang, X., Liu, Z., Hou, L., Tu, C., Li, J., Liu, Y., Shen, W., Sun, M.: Leven: A large-scale chinese legal event detection dataset. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 183–201 (2022)
36. Yao, F., Zhang, J., Zhang, Y., Liu, X., Sun, C., Liu, Y., Shen, W.: Unsupervised legal evidence retrieval via contrastive learning with approximate aggregated positive. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 4783–4791 (2023)
37. Yuan, J., Wei, Z., Gao, Y., Chen, W., Song, Y., Zhao, D., Ma, J., Hu, Z., Zou, S., Li, D., et al.: Overview of smp-cail2020-argmine: The interactive argument-pair extraction in judgement document challenge. Data Intelligence **3**(2), 287–307 (2021)
38. Zhang, G., Shi, Y., Liu, R., Yuan, R., Li, Y., Dong, S., Shu, Y., Li, Z., Wang, Z., Lin, C., Huang, W., Fu, J.: Chinese open instruction generalist: A preliminary release (2023)
39. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023)
40. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: How does nlp benefit legal system: A summary of legal artificial intelligence. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5218–5230 (2020)
41. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: Jec-qa: a legal-domain question answering dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 9701–9708 (2020)

A Appendix: Supplementaries

Prompt used in GPT-3.5-turbo for Behavior Shaping.

Instruction 1: You are the legal expert.
Here is a judicial case: {input}, In this case, {criminal} have committed {crimes}.
Here are some relevant law articles: {law articles}

Now, please use the law of syllogism to deduce a possible judgement and the penalty. In the law of syllogism, the major premise is law articles, the minor premise is case facts, and each conclusion should be drawn from law articles and case facts. Let's start the deduction. Note that the deductions are directly exported and answered in Chinese.

Instruction 2: You are the legal expert.
Here is a question: {input} and reference answer: {output}.
Here are some relevant law articles: {law articles}

Please answer this question based on the referenced answer and the relevant law articles. You need to use Chinese legal language to answer, you should strictly follow the law of syllogism. Specifically, each conclusion you make should be drawn from a law article and a fact. Be formal and concise. Note that you should reply in Chinese.

Prompt used in GPT-3.5-turbo as a judge for Subjective Evaluation.

Instruction: You are a professional, impartial and strict scorer. Given a question, pending scored answer and reference answer in the legal domain. Please rate the pending scored answer on a scale of 1 (lowest)-5 (highest) for each of the 3 criteria according to reference answer. The detailed criterion is as follows:
Accuracy: The content of pending scored answer conforms to reference answer in semantic, especially note the content of the law, the facts in question and the conclusion.
Completeness: Compared to reference answer, pending scored answer does not miss any details in reference answer. Do not let the length of the answer influence your judgment.
Clarity: The logic of pending scored answer is rigorous and clear, and the sentences are well-organized. If Accuracy and Completeness are bad, Clarity should also be bad.
Your rating should be strict enough, and do not easily give full scores. In your response, you should only include a JSON object, with keys being the aspects and values being the scores. Do not include any additional information or explanation.

question: {question}
reference answer: {ground truth}
pending scored answer: {answer}