Qzhou-Law: An Open Source Series of Chinese Legal Large Language Models

Yuchen Xie[1], Yixin Zhou[1], Huaitong Gao[1], Wensen Jiang[1], Jingxiang Fan[1], Chuhan Yan[1], Zhiwei Fei[3], Yazhou Wang[1], Haibiao Chen[1], Yinfei Xu[1], Wei Zhou[2]*

[1]Kingsoft AI, Kingsoft Corp. Ltd., Beijing, China
[2]School of Law, Wuhan University, Wuhan, China
[3]National Key Laboratory for Novel Software Technology, Nanjing University, China
{xieyuchen1, zhouyixin, gaohuaitong, jiangwensen, fanjingxiang, yanchuhan, wangyazhou, chenhaibiao, xuyinfei}@kingsoft.com, zhiweifei@smail.nju.edu.cn, fxyzw@whu.edu.cn

Abstract

Both general large language models (LLMs) and Chinese legal LLMs lack legal capability for practical application due to the limitations of available training datasets and effective legal training methods. To further enhance the legal capabilities of LLMs, we introduce a series of legal large language models (LLMs), Qzhou-Law 7B/14B/32B/72B. The core innovations of these model series are that (1) we construct a large-scale legal instruction-tuning dataset and (2) we explore a new training method to train a legal-specific LLM in three phases better. To build this dataset, we curated 853,000 legal instructions with appropriate data processing and augmented them with legal consultation data using an article-based IRAC (Issue, Rule, Application, Conclusion) technique. We demonstrate that our three-stage training approach yields better results than training only on the legal instruction dataset. Our trained models achieve a new series of state-of-the-art (SOTA) performances on both LawBench and LexEval, and we are the first to test the effect of scaling model size in this setting. Increasing the size of the models consistently yields stronger results. To evaluate the timely changes in laws, regulations, and relevant knowledge, we collected 1.4K questions from the National Unified Legal Professional Qualification Examination (NULPQE) between 2018 and 2024. Our models outperform those of our competitors on NULPQE. We make our models and the NULPQE dataset publicly available to facilitate future research in applying LLMs within the legal domain.

Introduction

The rapid development of open-source generalist large language models (LLMs) like the Qwen and DeepSeek series has significantly expanded the boundaries of Natural Language Processing (NLP) and improved an increasing number of domains, such as math and coding. While these models have shown impressive performance on general language tasks, their expertise in legal tasks is still limited. Generalist LLMs are not good enough at document proofreading and contract review, which are frequently needed in legal practice.

In the Chinese legal domain, open-source high-quality LLMs remain scarce. Previous works have released several legal LLMs, but their performance on legal tasks is worse than the latest general-purpose LLMs, limited by their model size and training methodology. InternLM-Law (7B) (Fei et al. 2025) is the most powerful legal LLM outperforming on LawBench (Fei et al. 2024) at that time, but still performs worse than DeepSeek and Qwen in legal memorization. Besides, it is necessary to continue updating legal LLMs to follow the changes in laws and regulations.

Some main obstacles remain in training better legal-specific LLMs. The first one is lacking a sufficient training dataset. Although there are lots of open-sourced legal datasets, they are not suitable for instruction tuning, and we need to convert these datasets into a suitable format. The other one is limited by the post-training methodology. A general way to obtain a legal LLM is to train a generalist model to improve its legal knowledge and skills. DISC-Law (Yue et al. 2023) shows that without pretraining, only performing instruction tuning on the legal datasets is enough for the legal-specific LLMs. Based on their findings, InternLM-Law (Fei et al. 2025) proposed a two-stage Supervised Fine-tuning (SFT) training strategy with corresponding dataset construction. The first stage is to train on InternLM2 with resampling its disclosed general-purpose training datasets to mix with a 1-million constructed legal dataset. The second stage is to train on the checkpoint generated before, with only approximately one percent of one million legal datasets, including synthetic high-quality consultation datasets, resampled datasets of Chinese laws and regulations, and a subset of legal NLP task datasets. However, it is still a challenge to train an open-source LLM with proper data to obtain an outstanding legal LLM, because it is almost impossible to utilize the proper data recipe for the replay strategy without its original training recipe.

Inspired by InternLM-Law and DISC-Law, we utilize instruction tuning to train a legal LLM based on the open-source base models. We collect and clean a comprehensive dataset sourced from various public legal datasets on the internet, including both question-and-answer and plain text datasets for SFT training. To convert data formats into SFT datasets, we have developed customized data-cleaning pipelines, filtering rules, and augmentation strategies for different datasets to improve data quality and diversity. Legal skills can only be learned effectively based on foundation capabilities, such as text understanding, instruction follow-

[*] Corresponding authors

Table 1: Overall Statistics of Legal Datasets

Dataset	Source	Size
Legal Knowledge	JEC-QA	25K
	Website	55K
Laws and Regulations	Website	100K
Legal Counseling	Hanfei	42K
	DISC-Law	23K
	LawGPT	35K
	Lawyer-llama	1K
	Website	255K
Legal Scenarios	Public Datasets	317K

ing, and even foundation knowledge. It is not sufficient to train models solely on legal datasets. To enable the model to own general skills and keep them in legal training, we implement foundation and conversational training with Infinity-Instruct (Li et al. 2025) datasets and then incorporate data sampled from Infinity-Instruct to train them in conjunction with legal datasets with a proper data replay strategy. In brief, we design a three-stage instruction tuning method to train a series of Chinese legal LLMs named Qzhou-Law based on Qwen2.5 (Qwen et al. 2025) base models. To evaluate the performance of models for the changes in laws and regulations, we collect the objective questions and corresponding answers from the National Unified Legal Professional Qualification Examination (NULPQE) between 2018 and 2024 to construct a benchmark with their original scoring rules. Our models are demonstrated as a series of new state-of-the-art (SOTA) performances on LawBench, LexEval (Li et al. 2024), and NULPQE benchmarks.

Our main contributions can be summarized as follows:

- We propose a new approach to efficiently fine-tune a legal-specific LLM from the general models with three stages, including foundation training for instruction following, conversational, and legal training with a proper data replay strategy. We demonstrate that one-stage legal training is better than a two-stage training strategy.
- We collect and build a large-scale dataset selected for the Chinese legal tasks, containing over 853K training samples for existing legal NLP scenarios, legal knowledge, laws, and regulations, and legal counseling. We implement effective filtering and processing strategies to improve the quality of legal counseling data with an article-based IRAC technique.
- We construct a NULPQE benchmark of 1.4K samples between 2018 and 2024 to evaluate the performance of models for the changes in laws and regulations.
- We open a series of legal LLMs based on the Qwen2.5 series, achieving a new state-of-the-art (SOTA) performance on LawBench, LexEval, and NULPQE benchmarks among all existing competitors.

Table 2: Legal Scenarios and Corresponding Tasks

Task	Task Size
Dispute Focus Identification	14K
Marital Disputes Identification	14K
Issue Topic Identification	21K
Comprehension	12K
Named-Entity Recognition	9K
Summarization	35K
Argument Mining	7K
Event Detection	27K
Trigger Word Extraction	36K
Article Prediction	40K
Accusation Prediction	20K
Imprisonment Prediction	41K
Criminal Damages Calculation	2K
Judgment Generation	20K
Proofreading	19K

Related Work

Legal Large Language Model

In the Chinese legal domain, researchers have already attempted to build Legal LLMs. But these models are pre-trained and fine-tuned on limited legal datasets. Lawyer-LLaMA (Huang et al. 2023) is pre-trained on legal datasets and fine-tuned to enhance counseling abilities on consultation-related datasets. LawGPT_zh (Hongcheng Liu 2023) only efficiently fine-tuned LLMs on consultation and knowledge Q&A. ChatLaw (Cui et al. 2023) explored the impact of model size on performance scaling with both dense and MOE models. Hanfei (He et al. 2023) firstly trained on multi-turn dialogue and document generation through SFT mixed with general-purpose instruction datasets. Fuzi-Mingcha (Wu et al. 2023) continued to fine-tune a base legal LLM after pre-training based on prior work. In addition, DISC-LawLLM (Yue et al. 2023) constructed a comprehensive dataset of 403K samples with general, pair-wise of law, and triplet-wise of law data. Specifically, it introduced the syllogism of legal judgment to divide answers into the major premise of laws, the minor premise of pertinent facts, and the conclusion of judgment for pair-wise data generation. However, almost all legal LLMs mentioned above stopped updating to continue to improve their legal capabilities, except DISC-LawLLM. Although these works make their datasets and models partially public, top general LLMs are popular in the legal domain since open-source legal LLMs mentioned above are not competitive enough.

InternLM-Law introduced a two-stage legal SFT with a comprehensive data recipe and resampling strategy to outperform top generalist LLMs. Then, LAWGPT (Zhou et al. 2025) introduced a knowledge-guided data generation framework focusing on legal reasoning. It works well only on a reasoning subset of LawBench. Overall, InternLM-Law performs well but is not transparent enough for further up-

Table 3: Performance Comparison Across Models on NULPQE Scores Between 2018 and 2024

Model	2024	2023	2022	2021	2020	2019	2018	Average
GPT-5.1	111	147	132	107	121	144	145	130
DeepSeek-V3	113	154	150	127	141	146	133	138
LawLLM-7B	121	165	108	93	132	141	130	127
Qzhou-Law (7B)	154	172	143	167	179	189	204	173
Qzhou-Law (72B)	**193**	**227**	**202**	**180**	**242**	**219**	**241**	**215**

Table 4: Lawbench Performance Comparison Among SOTA Models

Level	GPT-5.1	DeepSeek-V3	InternLM-Law	Qzhou-Law (7B)	Qzhou-Law (72B)
Memorization	41.19/41.18	66.28/65.7	63.72/64.95	64.33/60.21	**77.12/75.05**
Understanding	49.2/55.97	44.68/52.36	71.81/71.58	74.53/68.17	**78.21/74.09**
Application	61.18/63.02	62.74/65.36	63.57/63.46	72.75/70.55	**75.76/74.78**
Average	53.19/57.34	54.06/58.90	67.71/67.67	72.8/68.32	**77.12/74.46**

dating. LAWGPT is transparent, but limited in application.

Legal Benchmarks DISC-Law-Eval (Yue et al. 2023) is inspired by bar examinations to construct an objective benchmark, but lacks data updating. Its subjective evaluation relies on the legal capabilities of judge LLMs like GPT-4 to assess accuracy, completeness, and clarity. Due to the limitations of the legal capabilities of judge LLMs, it is not popular in the Chinese legal community. LawBench (Fei et al. 2024) is a well-established benchmark constructed for the Chinese legal domain with customized metrics designed for specific tasks. It assesses models in memorization, understanding, and application in 20 tasks with various evaluation metrics. In addition to direct evaluation (zero-shot), it reports the model's one-shot performance, to take into account cases where LLMs are not familiar with the input-output format of a given task. LAiW (Dai et al. 2025) owns 14 tasks at 3 legal levels, including basic information retrieval, legal foundation inference, and complex legal application. Both LawBench and LAiW test legal capabilities from a computer-centric perspective. LexEval (Li et al. 2024) is designed to test the practical use of LLMs in legal applications through a comprehensive standardized Chinese legal benchmark with 23 tasks and 14,150 questions to evaluate both fundamental legal knowledge and ethical issues.

Instruction Datasets

To train Qzhou-Law, we construct a comprehensive dataset with foundation datasets, conversational datasets, and legal datasets. Both foundation and conversational datasets are sourced from Infinity-Instruct (Li et al. 2025), with over 8 million samples. Legal knowledge, legal scenarios, legal counseling, laws, and regulations make up the legal datasets. The legal knowledge datasets contain two main types of legal education and examination, such as exercises. The legal scenarios include a variety of tasks related to existing legal NLP tasks generated by intelligent legal applications. Legal counseling includes different types of conversations generated by frequent legal advice. The final category comprises up-to-date legal regulations and laws to support citation as a major premise in legal practice. Detailed statistics of legal datasets are provided in Table 1.

Foundation Datasets

The foundation dataset mainly includes world knowledge, math, and code categories to make LLMs learn fundamental capabilities. Knowledge integrates hundreds of high-quality templates, diverse formatting patterns, and extensive data augmentation, including a mix of zero-shot, few-shot, and chain-of-thought prompt formats. Math and code possess millions of open-source and synthetic data prompted from GSM8K, MATH, and HumanEval. The different subsets of the foundation data are filtered for various model sizes to achieve the best performance.

Conversational Datasets

Conversational datasets are utilized to improve the alignment of chatting. To make models chat like human beings, the conversational data is constructed by instruction labeling, high-quality seed instruction selection, instruction evolution, synthesis, and final diagnosis. The instruction seeds of high quality are filtered by long-tail distribution, multi-dimensional capabilities, high modeling loss of responses, and high convergence loss after model fine-tuning. Different combinations of conversational datasets are customized for different model sizes to implement SFT.

Legal Datasets

Legal Knowledge To make LLMs memorize and understand legal concepts, we collected more than 60K examination questions from the website, excluding the questions from the National Unified Legal Professional Qualification Examination (NULPQE) after 2017 for decontamination. We also sourced more than 26K samples from JEC-QA (Zhong et al. 2020). After de-duplication and filtering, we select a total of 80K up-to-date legal knowledge samples

Table 5: Performance Comparison Across Models on LexEval

Level	GPT-5.1	DeepSeek-V3	DISC-LawLLM (7B)	Qzhou-Law (7B)	Qzhou-Law (72B)
Memorization	43.14	50.98	49.07	55.44	**71.30**
Understanding	83.09	85.53	77.81	77.88	**87.92**
Logic Inference	66.13	70.03	61.79	72.26	**83.08**
Discrimination	30.74	30.07	32.67	22.97	**36.44**
Generation	21.56	24.76	26.46	41.28	**45.42**
Ethic	56.30	58.10	59.67	61.83	**70.73**
Average	54.71	58.53	54.66	60.25	**70.38**

Table 6: Consultation Performance in LawBench Comparison Among SOTA Models

Setting	InternLM-Law	DISC-LawLLM	DeepSeek-V3	GPT-5.1	Qzhou-Law	Qzhou-Law
	7B	7B	685B	N/A	7B	72B
zero-shot	23.17	17.91	16.32	9.27	**27.07**	26.58
one-shot	22.37	18.64	19.82	13.48	**24.36**	24.15

related to international law, judicial system and legal professional ethics, criminal law, criminal procedure law, administrative law, administrative procedure law, civil law, intellectual property law, commercial law, economic law, environmental resources law, labor and social security law, private international law, international economic law, civil procedure law (including arbitration system).

Laws and Regulations The foundational premise of legal LLMs is their ability to incorporate accurate laws and regulations properly. To this end, we selected around 100K frequent articles from Chinese laws and regulations according to their effectual level and distribution of frequency from the Chinese National Legal Database, including laws, constitutional law, administrative supervision laws, judicial interpretation, and administrative regulations, along with an extensive range of regional regulations.

Legal Scenarios The legal scenario datasets comprise tasks within the legal domain that are well-defined, such as standard NLP tasks from previous research endeavors in the field, including legal event detection (Yao et al. 2022), legal element extraction (Zongyue et al. 2023), datasets from public legal competitions like CAIL (Xiao et al. 2018), and others. In total, we construct a comprehensive dataset of 317K samples in 15 distinct legal scenarios, including dispute focus identification, marital disputes identification issue, topic identification, comprehension, named-entity recognition, summarization, argument mining, event detection, trigger word extraction, article prediction, accusation prediction, imprisonment prediction, criminal damages calculation, judgment generation, and proofreading after deduplication and decontamination. Detailed statistics of legal datasets are provided in Table 2.

Legal Counseling We construct a legal counseling dataset of 353K conversations. Some of the datasets are sourced from LAWGPT_zh (Hongcheng Liu 2023), HanFei (He et al. 2023), Lawyer-LLaMA (Huang et al. 2023), and DISC-Law (Yue et al. 2023). To further improve the quality of these formatted conversations, IRAC prompting (Jiang and Yang 2023) is used to re-generate answers for existing questions with GPT-4.

The rest of the counseling datasets are collected on the website involving real-world legal issues, spanning civil disputes, policy interpretation, and criminal cases. A length filter is used to discard answers less than 100 characters to avoid a lack of details or citation to an article. In addition, an article filter based on a regular expression rule is utilized to increase authority. Finally, an automatic scoring filter based on legal syllogism is designed to ensure the highest quality.

Data Processing We use GPT-4 to generate diversified and semantically similar instructions after manually writing about 10 seed instructions per scenario, followed by self-instruct (Wang et al. 2023). One instruction is randomly selected from the set to construct a legal task dataset for diversity. Deduplication and decontamination are implemented by BGE (Chen et al. 2024) with a threshold of 0.3.

Three-Stage Legal Instruct Training

There are two ways to fine-tune a legal LLM from a generalist model. The first is to train on a base model with both general and legal skills together, like DISC-LawLLM. But it was found to have poor performance on general capabilities and even lower than the upper limit of legal performance. The second way refers to InternLM-Law, where we train on an instruction model after alignment with the proper data replay strategy of the training dataset utilized for instruction training. Regretfully, there is almost no training data public of powerful open-source LLMs. Missing a dataset for data replay also harms the legal training. Therefore, we introduced a three-stage supervised fine-tuning to train a legal LLM from a base model as follows: foundation training, conversational training, and legal training with a proper data

Table 7: Scaling Model Performances on Lawbench

Task ID	**Metric**	**Qzhou-Law (7B)**		**Qzhou-Law (14B)**		**Qzhou-Law (32B)**		**Qzhou-Law (72B)**	
		zero-shot	one-shot	zero-shot	one-shot	zero-shot	one-shot	zero-shot	one-shot
1-1	Rouge-L	57.85	53.62	63.99	61.4	64.46	63.11	**72.99**	**72.50**
1-2	Accuracy	70.80	66.80	73.80	69.40	81.20	**80.40**	**81.20**	77.60
2-1	F0.5	62.98	60.59	65.71	69.19	**69.10**	**73.57**	68.72	73.13
2-2	F1	47.60	37.80	43.20	32.20	37.60	**38.60**	**51.20**	36.00
2-3	F1	**86.02**	84.59	85.33	85.49	83.54	85.54	85.78	**86.21**
2-4	Accuracy	**52.40**	50.80	48.60	46.80	49.40	49.00	50.60	**51.4**
2-5	rc-F1	82.94	85.77	94.94	94.54	95.87	93.11	**98.94**	**97.64**
2-6	soft-F1	58.71	57.88	58.54	58.44	**58.84**	**59.03**	58.02	58.78
2-7	Rouge-L	77.68	51.40	78.89	62.63	65.97	58.22	**90.54**	**65.08**
2-8	Accuracy	89.40	67.20	**92.20**	86.40	90.00	**87.20**	90.40	85.60
2-9	F1	92.15	90.68	**93.35**	91.03	92.8	91.77	92.89	**92.54**
2-10	soft-F1	**95.44**	**95.00**	94.48	94.36	94.61	94.02	94.99	94.47
3-1	F1	88.13	88.29	89.34	89.02	89.97	90.00	**90.30**	**90.00**
3-2	Rouge-L	57.69	52.53	58.41	52.67	55.88	54.49	**61.33**	**59.17**
3-3	F1	75.88	73.80	**77.18**	**75.49**	76.93	74.45	76.19	74.41
3-4	nLog-distance	**90.95**	90.77	90.35	90.24	90.69	**90.96**	90.60	90.72
3-5	nLog-distance	90.49	90.62	90.72	90.44	91.03	90.12	**91.08**	**90.98**
3-6	Accuracy	74.20	69.40	73.80	67.00	78.80	77.40	**82.00**	**81.60**
3-7	Accuracy	77.60	74.60	79.80	70.20	**88.60**	**88.00**	88.00	87.20
3-8	Rouge-L	27.07	24.36	27.46	25.30	**28.04**	**26.56**	26.58	24.15
AVG		72.80	68.32	74.00	70.61	74.17	73.33	**77.12**	**74.46**

replay strategy.

Foundation Training We trained base models of Qwen2.5 on foundation datasets that we constructed to teach models instruction following. We trained around 3 million foundation data for a small-sized model of 7B and trained 7 million foundation data for a large-sized model of 72B to obtain instruct models for the next SFT phase.

Conversational Training We continued to train the instruct models generated by foundation training on conversational datasets that we constructed to enhance chat abilities. A data replay strategy of 5% of math and code-related data in the foundation dataset was implemented to retain foundational capabilities of instruct models.

Legal Training Finally, we trained chat models developed by conversational training on well-designed legal datasets. We randomly combine 20% of both the foundation training data and the conversational training data through stratified sampling for data replay to maintain general capabilities.

Training Details

We adopted both LLaMA-Factory (Zheng et al. 2024) and Megatron (Shoeybi et al. 2019) to fine-tune the series of Qwen-2.5 models, including 7B, 14B, 32B, and 72B versions. Our training process was conducted on a cluster with 256 NVIDIA A800 GPUs.

In both foundation and conversational training, the training epochs were set to 3, and the learning rate was set to 1e-5 for a small size and 5e-6 for a bigger size with a cosine learning rate scheduler. The training epochs are set to 2, and the learning rate is set to 1e-5 for the small size and 5e-6 for the larger size, with a cosine learning rate scheduler for legal training to optimize performance.

Experiments

Legal Benchmarks There are three benchmarks that focus on evaluating LLMs' capabilities in the Chinese legal domain: LawBench, LexEval, and LAiW. Across these three benchmarks, five datasets are commonly used. After a detailed analysis, we found that LAiW is not well constructed. The Name Entity Recognition (NER) tasks selected in LAiW are simplified to be asked to respond to only one of the name entities instead of all in LexEval. Similarly, Judicial Summarization tasks in LAiW are organized with several fixed parts, including ”plaintiff claims”, ”defendant claims”, and ”judgment”. It is confusing to be considered as a general summarization task in LAiW instead of an information extraction task. Dispute Focus Identification tasks from LAIC-2021 in LAiW are constructed as the less difficult multiple-choice questions with options of labels instead of a multi-classification task in LawBench. The Legal Article Prediction tasks are utilized in all three benchmarks as multi-classification of article index in LAiW, of

Table 8: Scaling Model Performances on Lexeval

Level	Task	Qzhou-Law (7B)	Qzhou-Law (14B)	Qzhou-Law (32B)	Qzhou-Law (72B)
Memorization	1-1	63.60	67.00	71.80	**75.00**
	1-2	70.40	88.50	95.90	**98.90**
	1-3	32.33	39.33	32.33	**40.00**
Understanding	2-1	91.80	94.00	**97.80**	94.60
	2-2	34.00	47.33	**58.00**	53.00
	2-3	89.00	92.00	94.00	**96.00**
	2-4	76.20	88.80	**97.20**	96.40
	2-5	98.40	99.80	**99.80**	99.60
Logic Inference	3-1	80.30	**81.50**	81.20	81.00
	3-2	90.70	95.70	**98.80**	97.90
	3-3	64.80	72.90	69.30	**80.70**
	3-4	61.80	65.80	69.20	**70.00**
	3-5	46.75	69.00	**80.00**	79.25
	3-6	89.20	**90.40**	87.00	89.60
Discrimination	4-1	18.60	26.40	23.20	**28.20**
	4-2	27.33	29.33	42.33	**44.67**
Generation	5-1	77.62	79.98	51.13	**80.78**
	5-2	23.44	27.32	28.73	**31.76**
	5-3	32.89	34.31	**39.47**	38.61
	5-4	31.16	**33.43**	32.68	30.54
Ethic	6-1	59.40	60.90	61.60	**66.30**
	6-2	53.70	57.80	**63.50**	62.70
	6-3	72.40	76.20	81.00	**83.20**
AVG		60.25	65.99	67.65	**70.38**

article abbreviation in LawBench, and multi-choice of article abbreviation. The article index is not frequently used by itself. Therefore, we determined not to evaluate legal LLMs with LAiW in our work. Additionally, we evaluate our models with NULPQE between 2018 and 2024 with their original scoring rules. It includes 100 questions of single choice scored for 1 point each, 70 questions of strict multiple choices for 2 points each, and 30 questions of multiple choices for 2 points each. Generally, 180 points are necessary to pass the examination.

Model Baselines We compare our models of 7B and 72B with SOTA LLM baselines of both legal-purpose and general-purpose, including GPT-5.1, DeepSeek-V3 (685B-0324), InternLM-Law (7B), and DISC-LawLLM (7B). Qzhou-Law (7B) was selected for fair competition because both legal LLMs of SOTA have 7 billion parameters.

Evaluation Settings We evaluate LLMs in legal contexts using both LawBench and LexEval through OpenCompass (Contributors 2023) with the vLLM (Kwon et al. 2023) inference engine. For decoding, we set a maximum generation length of 1024 and use greedy decoding. In addition, we also evaluate the questions of the National Judicial Examination of China between 2018 and 2024 to further estimate the level of knowledge.

Results and Analysis

Our models achieve 75.79 of 72B, and 70.61 of 7B on average for LawBench, where the previous SOTA model of InternLM-Law only achieves 67.69. Meanwhile, our models achieve 70.38 of 72B, and 60.25 of 7B on average for LexEval, where the previous SOTA model of DeepSeek-V3 only achieves 58.53. Our models achieve 215 of 72B, and 173 of 7B on average for NULPQE, where the previous SOTA model of DeepSeek-V3 only achieves 138. In addition, our 72B model passed all 7-year examinations.

IRAC Effectiveness Our 7B model proved the effectiveness of the IRAC technique to improve the consultation capability on LawBench. The detailed result comparison among competitors is shown in Table 6. Specifically, GPT-5.1 performs the worst due to its extremely long length of responses. **The IRAC approach is demonstrated to be effective for counseling evaluation on LawBench.**

Ablation Study

We implemented two experiments to further discover the effectiveness of the two-stage legal training strategy and model scaling. We compare the results obtained from the two-stage legal training by the 1-epoch training checkpoint using a similar data recipe to InternLM-Law with the results

Table 9: Two-Stage Abalation Study on LawBench

Level	**One Stage**	**Two Stage**		
		Stage One	**Stage Two**	
	2 epoch	**1 epoch**	**1 epoch**	**2 epoch**
Memorization	**64.18/60.97**	54.19/51.72	54.55/52.34	56.36/53.92
Understanding	**76.04/69.96**	70.72/67.22	70.65/68.16	69.53/66.55
Application	**72.51/69.94**	69.9/68.44	69.99/68.25	69.93/68.53
Average	**73.44/69.05**	68.74/66.16	68.78/66.61	68.37/66.08

Table 10: Two-Stage Abalation Study on LexEval

Level	**One Stage**	**Two Stage**		
		Stage One	**Stage Two**	
	2 epoch	**1 epoch**	**1 epoch**	**2 epoch**
Memorization	**53.59**	47.46	42.88	52.47
Understanding	**85.49**	83.81	83.58	84.17
Logic Inference	69.86	69.56	**70.78**	70.66
Discrimination	23.74	25.10	**31.70**	23.94
Generation	**38.65**	32.54	32.65	36.19
Ethic	53.40	**60.30**	59.03	53.37
Average	**59.55**	58.26	58.36	58.91

obtained from our 2-epoch version of one-stage legal training. In addition, we compare with all detailed results of 7B, 14B, 32B, and 72B models on both LawBench and LexEval.

Training Stages According to InternLM-Law, due to limited resources for cleaning all collected datasets, they propose a two-stage SFT training strategy that has a modest effect on improving response style and reducing hallucinations about laws and regulations, as evaluated by LawBench. The dataset for second-stage training includes synthetic counseling data generated by GPT-4 with 6K refined questions and relevant legal references, crucial legal articles of 10K resampled based on frequency, and 4K legal NLP tasks sampled to retain legal capabilities. The crucial legal articles include marriage law, labor law, criminal law, constitutional law, etc. Similarly, we sampled 4K of legal scenario data, 6K of augmented legal counseling data, and 10K of crucial legal articles to train a second-stage model based on our legal-instruct models. We demonstrated that one-stage legal training with 2 epochs outperforms both on LawBench and LexEval, including both memorization subsets of benchmarks shown in Table 9 and Table 10. However, we found that our method performs worse in discrimination and ethics than a two-stage trained version of LLM.**When sufficiently large, high-quality training datasets are available, a one-stage legal SFT is enough to achieve better performance on LawBench and LexEval, including on memorization. However, it doesn't yield better performance on the discrimination and ethics subsets of LexEval.**

Scaling Trends Our method has been proven to be scalable, shown in Table 7 of Lawbench and Table 8 of LexEval. Although overall performance keeps scalable on LawBench and LexEval, performance on some specific tasks still fluctuates when model size increases, including dispute focus identification, issue topic identification, summarization, trigger word extraction, and event detection. The performance on all memorization subsets in LawBench, and 2/3 of the memorization subsets in LexEval keep increasing when the model size scales. The rest of the tasks perform almost similarly due to the limitations of these capabilities. Especially, we consider the minor changes in consultation subsets on LawBench as a limitation of the evaluation metric. Rouge-L is limited to evaluating the responses with answers and article recitations. **Performance scaling trends remain strong on LawBench and LexEval.**

Conclusion and Limitations

In this paper, we introduce a series of Chinese legal large language models named Qzhou-Law through a transparent and feasible method of three stages, including foundation, conversational, and legal training. We refined outdated data on laws, regulations, and legal knowledge, and augmented counseling data with regeneration by the IRAC technique. We demonstrate our models trained on Qwen2.5 series SOTA and scalable on both LawBench, NULPQE, and LexEval. We construct the NULPQE dataset to evaluate whether LLMs perform well on changes in laws, regulations, and related knowledge. We make our models and

NULPQE dataset public to facilitate further work in applying legal LLMs for the community.

Although our Qzhou-Law models demonstrate strong overall scaling trends on Lawbench and LexEval, there are still fluctuating performances for some tasks on both LawBench and LexEval, such as consultation in LawBench. We consider that Rouge-L is not an effective metric to measure consultation with reference citations. Lacking legal scenario data leads the poor performance on legal evolution, similar case identification, judicial analysis generation, legal translation, and open-ended question answering in LexEval.

References

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Contributors, O. 2023. OpenCompass: A Universal Evaluation Platform for Foundation Models. https://github.com/open-compass/opencompass.

Cui, J.; Li, Z.; Yan, Y.; Chen, B.; and Yuan, L. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*.

Dai, Y.; Feng, D.; Huang, J.; Jia, H.; Xie, Q.; Zhang, Y.; Han, W.; Tian, W.; and Wang, H. 2025. LAiW: A Chinese legal large language models benchmark. In *Proceedings of the 31st International conference on computational linguistics*, 10738–10766.

Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Huang, A.; Zhang, S.; Chen, K.; Yin, Z.; Shen, Z.; et al. 2024. Lawbench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, 7933–7962.

Fei, Z.; Zhang, S.; Shen, X.; Zhu, D.; Wang, X.; Ge, J.; and Ng, V. 2025. Internlm-law: An open-sourced chinese legal large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, 9376–9392.

He, W.; Wen, J.; Zhang, L.; Cheng, H.; Qin, B.; Li, Y.; Jiang, F.; Chen, J.; Wang, B.; and Yang, M. 2023. HanFei-1.0. https://github.com/siat-nlp/HanFei.

Hongcheng Liu, Y. M. Y. W., Yusheng Liao. 2023. XieZhi Chinese Law Large Language Model. https://github.com/LiuHC0428/LAW_GPT.

Huang, Q.; Tao, M.; Zhang, C.; An, Z.; Jiang, C.; Chen, Z.; Wu, Z.; and Feng, Y. 2023. Lawyer LLaMA Technical Report. *arXiv e-prints*, arXiv:2305.15062.

Jiang, C.; and Yang, X. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the nineteenth international conference on artificial intelligence and law*, 417–421.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Li, H.; Chen, Y.; Ai, Q.; Wu, Y.; Zhang, R.; and Liu, Y. 2024. Lexeval: A comprehensive chinese legal benchmark for evaluating large language models. *Advances in Neural Information Processing Systems*, 37: 25061–25094.

Li, J.; Du, L.; Zhao, H.; Zhang, B.-w.; Wang, L.; Gao, B.; Liu, G.; and Lin, Y. 2025. Infinity Instruct: Scaling Instruction Selection and Synthesis to Enhance Language Models. *arXiv preprint arXiv:2506.11116*.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; and Catanzaro, B. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, 13484–13508.

Wu, S.; Liu, Z.; Zhang, Z.; Chen, Z.; Deng, W.; Zhang, W.; Yang, J.; Yao, Z.; Lyu, Y.; Xin, X.; Gao, S.; Ren, P.; Ren, Z.; and Chen, Z. 2023. fuzi.mingcha.

Xiao, C.; Zhong, H.; Guo, Z.; Tu, C.; Liu, Z.; Sun, M.; Feng, Y.; Han, X.; Hu, Z.; Wang, H.; et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

Yao, F.; Xiao, C.; Wang, X.; Liu, Z.; Hou, L.; Tu, C.; Li, J.; Liu, Y.; Shen, W.; and Sun, M. 2022. LEVEN: A large-scale Chinese legal event detection dataset. *arXiv preprint arXiv:2203.08556*.

Yue, S.; Chen, W.; Wang, S.; Li, B.; Shen, C.; Liu, S.; Zhou, Y.; Xiao, Y.; Yun, S.; Huang, X.; et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Zhong, H.; Xiao, C.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020. JEC-QA: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 9701–9708.

Zhou, Z.; Yu, K.-Y.; Tian, S.-Y.; Yang, X.-W.; Shi, J.-X.; Song, P.; Jin, Y.-X.; Guo, L.-Z.; and Li, Y.-F. 2025. Lawgpt: Knowledge-guided data generation and its application to legal llm. *arXiv preprint arXiv:2502.06572*.

Zongyue, X.; Huanghai, L.; Yiran, H.; Kangle, K.; Chenlu, W.; Yun, L.; and Weixing, S. 2023. Leec: A legal element extraction dataset with an extensive domain-specific label system. *arXiv preprint arXiv:2310.01271*.