# InternLM-Law: An Open-Sourced Chinese Legal Large Language Model

**Zhiwei Fei[1], Songyang Zhang[2], Xiaoyu Shen[3], Dawei Zhu[4], Xiao Wang[4], Jidong Ge[1]*, Vincent Ng[5]**
[1]Nanjing University [2]Shanghai AI Laboratory [3]Eastern Institute of Technology
[4]Saarland University [5]University of Texas at Dallas
internlm@pjlab.org.cn

## Abstract

We introduce InternLM-Law, a large language model (LLM) tailored for addressing diverse legal tasks related to Chinese laws. These tasks range from responding to standard legal questions (e.g., legal exercises in textbooks) to analyzing complex real-world legal situations. Our work contributes to Chinese Legal NLP research by (1) conducting one of the most extensive evaluations of state-of-the-art general-purpose and legal-specific LLMs to date that involves an automatic evaluation on the 20 legal NLP tasks in LawBench (Fei et al., 2024), a human evaluation on a challenging version of the Legal Consultation task, and an automatic evaluation of a model's ability to handle very long legal texts; (2) presenting a methodology for training a Chinese legal LLM that offers superior performance to all of its counterparts in our extensive evaluation; and (3) facilitating future research in this area by making all of our code and model publicly available at https://github.com/InternLM/InternLM-Law.

## 1 Introduction

Legal Artificial Intelligence (LegalAI) (Zhong et al., 2020a) is an expanding area of Natural Language Processing (NLP) that concerns the computational study of a wide range of legal-related tasks. Examples of such tasks include Legal Judgment Prediction (LJP), which aims to predict the charge given the facts of a case, and Legal Consultation (LC), which aims to provide professional legal advice in the form of a response to a legal-related question posted by a user. Traditional work in this area has focused primarily on building *task-specific* models such as LJP models (Ge et al., 2021; Huang et al., 2021; Feng et al., 2022; Cui et al., 2023b). If a legal system needs to provide multiple services, one must build and coordinate several independent task-specific models, thereby increasing the system's complexity.

As a result, some researchers have begun work on building legal large language models (LLMs). For instance, SaulLM-7B (Colombo et al., 2024), which is a legal text model based on the Mistral 7B, is trained on a 30 billion-token English legal corpus that is designed explicitly for legal text comprehension and generation. In the Chinese legal domain, which is what we are focusing on in this paper, the state of research is rather unsatisfactory. Many Chinese legal LLMs, such as Lawyer llama (Huang et al., 2023), ChatLaw (Cui et al., 2023a) and DISC-Law (Yue et al., 2023), have only been evaluated on a limited set of tasks. This is less than desirable, as legal LLMs are typically constructed to handle a wide range of tasks. Nevertheless, some legal LLMs, such as Fuzi-mingcha (Wu et al., 2023) and Law-GPT (Zhou et al., 2024) have been evaluated more extensively on Chinese benchmark datasets such as LawBench (Fei et al., 2024). However, they offer inference performance to some general-purpose LLMs. Overall, existing Chinese legal LLMs have limited application scenarios, and have generally offered subpar performance compared to state-of-the-art LLMs, owing in part to the fact that they are built upon early models.

In light of the aforementioned status quo on legal LLM-based research for Chinese, our goal in this paper is two-fold. First, we introduce **InternLM-Law**, a LLM tailored for the Chinese legal domain. To train InternLM-Law, we assemble a large dataset composed of legal-related data as well as general-purpose data, and design a two-stage training strategy that aims to offer better results on a challenging version of the LC task. Second, we conduct an extensive evaluation and analysis of InternLM-Law and several state-of-the-art general-purpose and legal-specific LLMs. In an automatic evaluation on the 20 tasks in LawBench, InternLM-Law offers the strongest results overall by outperforming its counterparts on 14 and 15 tasks in zero- and one-shot scenarios respectively. Moreover, we

*Corresponding author.



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 9376–9392
January 19–24, 2025. 

conduct a human evaluation of InternLM-Law and state-of-the-art legal-specific LLMs on a challenging version of the LC task, which again demonstrates InternLM-Law's superiority. To our knowledge, this is the first human evaluation of legal-specific LLMs. Further, since each input sample in LawBench is limited to contain no more than 2000 tokens, we conduct a "long-text" evaluation that aims to evaluate a model's ability to handle input that contains up to 20,000 tokens. Once again, InternLM-Law offers the strongest performance.

In sum, our work contributes to Chinese LegalAI research by (1) conducting an extensive evaluation of state-of-the-art general-purpose and legal-specific LLMs; (2) presenting a methodology for training a legal-specific LLM that is arguably superior to its state-of-the-art counterparts; and (3) stimulating research on Chinese legal LLMs by making our code and model publicly available.

## 2 Related Work

In this section, we provide an overview of recently-developed Chinese legal LLMs.

Several Chinese legal models have only demonstrated their capabilities on a limited number of legal tasks, having a primary focus on judicial examinations. For instance, Lawyer llama (Huang et al., 2023) is a legal LLM for Chinese, which is created by continued pre-training on the Chinese-llama-13B model on legal datasets. The resulting model was evaluated only on two tasks, Charge Prediction and the national judicial examination on Marriage. ChatLaw (Cui et al., 2023a) is another legal LLM through which its authors explore the impact of model size on model performance. Specifically, they trained two large legal LLMs, one on Ziya-LLaMA-13B-v1 (Zhang et al., 2022) and the other on Anima-33B[1], and found that a larger-scale base model possesses stronger analytical capabilities. They evaluated their model on a self-compiled test dataset containing 2000 national judicial examination questions. DISC-Law (Yue et al., 2023) seeks to expand its application scenarios by enabling the model to provide a wider range of services such as completing legal tasks, legal consultations, and legal exam assistance. Given this goal, it is surprising that the model was only evaluated on legal exams and a few other scenarios.

In contrast, some legal LLMs have been evaluated on a wider range of tasks. For instance, Fuzimingcha (Wu et al., 2023) and Law-GPT (Zhou et al., 2024) were evaluated on LawBench (Fei et al., 2024). However, their performance was rather subpar when compared to general-purpose LLMs such as InternLM (Cai et al., 2024).

Finally, LexiLaw[2], which is based on the ChatGLM-6B (GLM et al., 2024), is fine-tuned on legal datasets to enhance its performance and expertise in providing legal consultation and support. It claims to offer accurate and reliable legal advice to practitioners, students, and the general public, but fails to provide any empirical support.

## 3 Model Training

To train our legal LLM, InternLM-Law, we employ InternLM2-Chat (Cai et al., 2024) as our foundation model and perform supervised fine-tuning (SFT) composed of *two* stages to specialize it for the legal domain. InternLM2-Chat is a multilingual LLM based on LLaMA that (1) is pre-trained and fine-tuned on vast, quality multilingual texts, (2) covers general, programming, and long contexts, and (3) is enhanced by SFT on 10M instructional instances and Reinforcement Learning from Human Feedback (RLHF) for human preference alignment. It has a 200K context window, allowing it to handle very long contexts. In the rest of this section, we present details of our two-stage training procedure and compare it with existing training procedures.

### 3.1 Stage 1 Training

In the first stage, we train the model on a mixture of legal and general-purpose data sources. Note that when addressing legal issues, merely relying on legal datasets to imbue the model with legal capabilities is inadequate. as it needs to rely on some general abilities such as text understanding and analytical skills, and employing general data sources could enable our model to transfer general skills to solving legal tasks. Below we describe the data sources (Section 3.1.1) and how we process the data (Section 3.1.2).

#### 3.1.1 Data Sources

As mentioned above, we employ two sources of data, legal data sources (Section 3.1.1.1) and general data sources (Section 3.1.1.2).

##### 3.1.1.1 Legal Data Sources

Using the legal data sources, we aim to construct a dataset encompassing a broad spectrum of legal

[1] https://huggingface.co/lyogavin/Anima33B

[2] https://github.com/CSHaitao/LexiLaw

| Task | Size | Source |
|---|---|---|
| **Legal Task Datasets** | | |
| Article Recitation | 20K | FLK |
| Knowledge QA | 20K | JEC_QA |
| Document Proofreading | 20K | CAIL2022 |
| Dispute Focus Identification | 20K | LAIC2021 |
| Marital Disputes Identification | 20K | AIStudio |
| Issue Topic Identification | 20K | CrimeKgAssitant |
| Reading Comprehension | 20K | CAIL2019 |
| Named-Entity Recognition | 20K | CAIL2021 |
| Opinion Summarization | 20K | CAIL2022 |
| Argument Mining | 20K | CAIL2022 |
| Event Detection | 20K | LEVEN |
| Trigger Word Extraction | 20K | LEVEN |
| Fact-based Article Prediction | 20K | CAIL2018 |
| Scene-based Article Prediction | 20K | LawGPT |
| Charge Prediction | 20K | CAIL2018 |
| Term Prediction w/o Article | 20K | CAIL2018 |
| Term Prediction w/ Article | 20K | CAIL2018 |
| Case Analysis | 20K | JEC_QA |
| Criminal Damages Calculation | 20K | LAIC2021 |
| Consultation | 20K | hualv.com |
| Judgement Generation | 20K | AC_NLG |
| Legal Element Extraction | 20K | LEEC |

Table 1: The 22 Legal NLP tasks.

knowledge from which our model can learn. The dataset is organized into three main categories: legal NLP, legal consultation, and legal regulations.

**Legal NLP data.** Our legal NLP data comprises tasks within the legal domain that are well-defined and exhibit consistent input and output formats, akin to standard NLP tasks. This data is sourced primarily from prior research endeavors in the field, including datasets from public legal competitions like CAIL[3], training data of tasks in LawBench (Fei et al., 2024), legal element extraction (Xue et al., 2024), and judgment generation (Wu et al., 2020). From these sources, we curated 22 distinct legal-related tasks, which are listed in Table 1.[4]

**Legal consultation data.** We collected six million records of legal consultation data from various online platforms.[5] These records represent a broad spectrum of real-world legal issues, spanning civil disputes, policy interpretation, and criminal cases. The data primarily consists of queries posed by individuals, which are then responded to by experienced legal practitioners, thus creating a rich collection of question-and-answer pairs. We employed robust anonymization procedures to safeguard sensitive information, including personal information such as lawyer names and detailed phone numbers.

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

[4] The definition of these tasks and additional information about these data sources can be found in Appendix A.

[5] Detailed URLs of these platforms can be found in Table 6 in Appendix B.

**Chinese laws & regulations data.** To enable a LLM to accurately incorporate relevant laws and regulations to address legal issues, we have sourced legal regulations data from the Chinese National Legal Database[6], a comprehensive data source encompassing Chinese civil, criminal, and constitutional laws, along with an extensive range of regulations. This endeavor aims to infuse the LLM with with precise and authoritative legal information. In total, we collected 100K entries from this database.

#### 3.1.1.2 General Data Sources

As our general data source, we sampled the SFT training dataset of InternLM2-Chat. This dataset consists of one million instruction data instances that have been screened for helpfulness and harmlessness and covers a comprehensive and diverse range of topics, such as everyday conversations, NLP tasks, mathematical problems, code generation, and function calls.

### 3.1.2 Processing Legal Data

Next, we describe our method for processing the data described above. Note that the data from the general data source is already pre-processed and ready to use for training SFT models, so no further processing is needed. Hence, below we will only describe our method for processing the legal data.

#### 3.1.2.1 Legal NLP Data Processing

To enable our model to learn from the 22 legal NLP tasks shown in Table 1 via instruction tuning, we need to write the instructions for each task. Fortunately, the data for 20 of these 22 tasks was derived from LawBench, which comes with manually written instructions that we can directly use for instruction tuning. For the remaining two tasks (judgment generation and legal element extraction), we need to manually write instructions. For each task, we then feed the manually written instructions to GPT-4 to generate diversified and semantically similar instructions, and randomly choose one of them to construct a legal task dataset, with the goal of enhancing the diversity of the legal task dataset.

Note that some of the 22 Legal NLP datasets mentioned in Table 1 originally contained fewer than 20K samples. To create a balanced distribution of samples over the 22 tasks (i.e., each task will have 20K samples), we upsampled those tasks that were under-represented and then paired each of the resulting samples with a different instruction

[6] https://flk.npc.gov.cn/

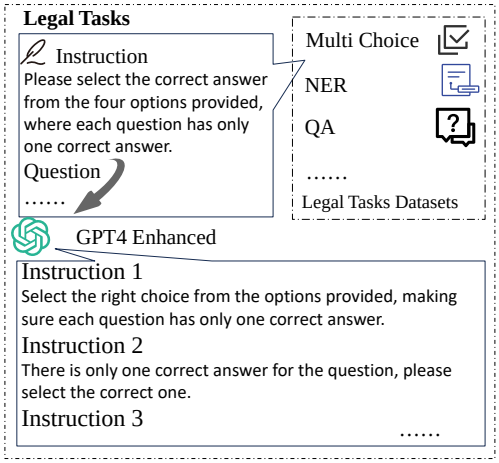


Figure 1: Procedure for compiling task instructions to be used for instruction tuning. Seed instructions are first manually written for each task, which are subsequently expanded using GPT-4 through paraphrasing This process results in a rich pool of diverse instructions. During training, each training instance is paired with an instruction randomly chosen from this pool.

generated according to the procedure shown in Figure 1. For the criminal damages calculation task, we further increased the number of samples by (1) using GPT-4 to generate the steps needed to calculate the answer, (2) manually verifying the steps, and (3) incorporating the steps into the answer.

#### 3.1.2.2 Legal Consultation Data Processing

Our legal consultation dataset, which is compiled from a wide range of online sources, contains extraneous information that can detrimentally affect data quality. To identify and remove such information, we employed the following filtering methods.

**Rule-based filtering.** Recognizing the complexity of legal consultations, which often involve complex legal matters requiring thorough analysis and discussion, comprehensive answers should exceed a certain length. Hence, we discarded answers with less than 20 characters as they likely lack the necessary details. Additionally, detailed answers typically reference legal provisions, as indicated by "《" and "法" ("law") in formal Chinese writing. Hence, responses lacking these markers were filtered out.

**Semantic filtering.** Next, we turned to instances characterized by poorly formulated questions. Notably, we noticed that questions may lack clarity and contain grammar errors. To identify such instances, we leveraged a LLM to assess instances based on the clarity and grammar correctness of the questions, discarding those that fall below a predetermined threshold. After that, we used a LLM to determine whether the question and the associated answer match, specifically by checking whether questions are comprehensively understood and appropriately addressed, and subsequently filtered out instances that receive low scores. Throughout this process, we used as the LLM Qwen-1.5-72B[7], because it is one of the strongest Chinese LLMs[8].

After filtering, 500K legal consultations remain.

#### 3.1.2.3 Chinese Laws & Regulations Processing

Since the legal regulation dataset is in pure text format, we transformed it into question-answering (QA) pairs for SFT. We formulate the dataset as QA pairs where in each pair, the question is the name of a law article and the answer is the content of the article. This would allow our model to recite the content of different laws or regulations, with the answer being the corresponding legal or regulatory text. This process yields 100K QA pairs.

### 3.2 Stage 2 Training

Among the data from the three sources we used in Stage 1 training, the LC data is of comparatively lower quality. The reason is that there are significant inconsistencies in the style of the human-written responses to the user queries in our LC dataset: not only is some content very brief, but there are also differences in expression and response structures. This variability in style makes a model's output less controllable, leading it to provide very detailed responses to some questions while offering only brief answers to other, possibly similar questions. In addition, the inconsistent response styles lead to a suboptimal user experience. This motivates us to perform Stage 2 training, where we refine the model by conducting additional training on high-quality legal consultation data that aims to enhance its response style and answer structure (Section 3.2.1) together with replay data sampled from the data used in Stage 1 SFT in an attempt to mitigate catastrophic forgetting (McCloskey and Cohen, 1989) of the knowledge acquired in Stage 1 (Section 3.2.2).

#### 3.2.1 Generating High-Quality LC Data

We generated expert-level LC data for Stage 2 training as follows. First, we selected 6K QA pairs from the LC dataset described in Section 3.1.2.2 that were tagged as "high-quality" by the website,

[7] https://huggingface.co/Qwen/Qwen1.5-72B
[8] https://qwenlm.github.io/blog/qwen1.5/

where each pair contains a question posed by individuals and a brief, manually written response with references to laws and regulations. Although these QA pairs were labeled as "high quality", the responses were provided by different lawyers, each with their own style, and some responses may lack thorough consideration. In other words, a closer inspection revealed that the quality of these pairs is not as good as what their tag suggests. To improve data quality, we employed three annotators with legal backgrounds to manually supplement the shorter responses with relevant legal references, and then fed this data into GPT-4 to generate more detailed and stylistically consistent replies.[9] Finally, the same three annotators manually checked the logic of the GPT-synthesized responses and the accuracy of the references in these responses, correcting any errors.

#### 3.2.2 Preventing Catastrophic Forgetting

Performing Stage 2 training solely on the 6k QA pairs mentioned above runs the risk of catastrophic forgetting of the knowledge acquired in Stage 1 from the other two data sources (i.e., legal NLP tasks and Chinese laws and regulations) (McCloskey and Cohen, 1989) during Stage 2 training. As a result, we additionally employed the data from these two sources during Stage 2 training. However, in order to ensure that the dataset used from Stage 2 training is relatively balanced over all three data sources, we needed to sample the data instances from the legal NLP tasks and the laws and regulations.

**Sampling laws and regulations.** We randomly sampled 10K QA pairs from the dataset described in Section 3.1.2.3. This ensures that the distribution of the sampled pairs over the type of laws (e.g., marriage law, labor law, criminal law, constitutional law) roughly follows that of the original dataset.

**Sampling legal NLP tasks.** We randomly sampled approximately 1% of the data from each of the 22 NLP tasks, which yields in 4K samples.

The resulting dataset used for Stage 2 training is composed of 20K samples: 6K samples from the LC dataset, 4K samples from legal NLP tasks, and 10K samples from the laws and regulations dataset.

[9]To get a better idea of the response style, we refer the reader to the example in Appendix C.

### 3.3 Implementation Details

We trained our model with 64 A100-80GB for 8 hours. To enable the model to process long legal texts, we set the training sequence length to 32k, ensuring that it can accommodate most legal text inputs. During training, we set the learning rate to 1e-5, and each stage was trained for one epoch.

### 3.4 Comparison with Other Training Schemes

To understand why our training procedure is novel, let us compare it with the existing procedures. Current legal LLMs are trained in one of two ways: (1) perform pre-training on legal corpora followed by supervised SFT on legal datasets (e.g., Lawyer LLaMA, LawGPT), or (2) perform only SFT (DISC-Law) or use LoRA (e.g., ChatLaw, LexiLaw). The second option is more popular due to its lower resource requirements. However, previous work fails to explore how to improve SFT, differing from us in three respects. First, they do not analyze the impact of general datasets on training legal LLMs. Second, they ignore the differences in response styles and quality in the LC datasets derived from various sources, fine-tuning their models on style-inconsistent and lower-quality data. Third, they fail to distinguish important, high-quality data (e.g., laws and regulations) from lesser-important, lower-quality data: mixing the two kinds of data causes the small amount of important data to be overshadowed by the large amount of lesser-important, lower-quality data.

## 4 Evaluation

In this section, we conduct a comparative evaluation of InternLM-Law and state-of-the-art general-purpose and legal-specific Chinese LLMs.[10]

### 4.1 Automatic Evaluation on LawBench

#### 4.1.1 Experimental Setup

**Dataset.** First, we conduct an automatic evaluation of the models using the test portion of LawBench (Fei et al., 2024), a well-established benchmark tailored for the Chinese legal domain.[11]

LawBench is composed of 20 evaluation tasks that can be divided into three categories: (1) *memorization* tasks, which examine the extent to which

[10]Due to space limitations, the application will be presented in the Appendix C.

[11]To prevent data leakage, we use the Levenshtein ratio (Levenshtein, 1965) to calculate the similarity between the training instances and the test instances and discarded training instances that are similar to the test instances.

LLMs encode legal knowledge within their parameters, (2) *understanding* tasks, which examine the extent to which LLMs can comprehend entities, events, and their relationships within legal texts, and (3) *application* tasks, which examine the ability of LLMs to not only understand legal knowledge but also simulate law professionals to apply the knowledge in solving realistic legal tasks.[12]

**Models.** We compare our model with seven existing LLMs, including (1) three general-purpose LLMs (Qwen-1.5-7B-Chat (Bai et al., 2023), InternLM2-7B-Chat (Cai et al., 2024), and Qwen-1.5-72B-Chat), which are the latest LLMs pretrained on massive amounts of Chinese data; (2) three legal-specific LLMs (DISC-Law-7B (Yue et al., 2023), Lawyer-LLaMA-13B (Huang et al., 2023), and ChatLaw-13B (Cui et al., 2023a)); and (3) GPT-4, a commercial LLM that has produced strong results on various tasks. For decoding, we set the maximum generation length to 2048 and use greedy decoding.

**Evaluation settings.** We evaluate the models using the zero-shot setting and the one-shot setting.[13]

#### 4.1.2 Results and Discussion

Zero-shot and one-short results are shown in the left half and the right half of Table 2, respectively. Each row shows the results of the models on a particular task in LawBench. On average, our model, InternLM-Law, outperforms all its counterparts, including GPT-4, on all three categories of tasks. At the task level, InternLM-Law achieves the best results in 14 of the 20 tasks in the zero-shot setting and 15 of these 20 tasks in the one-shot setting, demonstrating its superiority across different settings. Below we discuss the results of InternLM-Law on the three categories of tasks.

**Memorization tasks.** InternLM-Law achieves the best results in both tasks under the one-shot scenario and in Task 1-1 under the zero-shot scenario. For instance, for Task 1-1 (article recitation) InternLM-Law significantly outperforms all other models, scoring twice as much as the best Qwen-1.5-72B model. We found that our model is more accurate in citing legal statutes. Specifically, Qwen-1.5-72B tends to claim that certain laws or regulations do not exist and occasionally cites incorrect legal provisions, whereas our model can accurately reproduce legal texts verbatim, which contributes to its superior performance on this task, despite minor legal citation errors.

**Understanding tasks.** InternLM-Law outperforms other models in both the zero-shot and one-shot settings on all tasks except Task 2-6 (named entity recognition), where GPT-4 achieves the best performance. InternLM-Law struggles to understand the required output format, resulting in incorrect formatting for some responses and a lower score. As mentioned, InternLM-Law outperforms other models on the remaining understanding tasks. Below we take a closer inspection of the outputs to understand why.

For Task 2-5 (reading comprehension), InternLM-Law generates more concise content while other models tend to provide overly lengthy responses, which contributes to our higher score. For Task 2-7 (opinion summarization), our model's output is closer to the reference answer, whereas other models produce summaries that differ significantly in length and style from the reference, resulting in lower scores for them. For Task 2-9 (event detection), InternLM-Law generates more precise event labels, while DISC-Law produces too many labels and hence obtains a lower score. For Task 2-10 (trigger word extraction), our model predominantly generates words or phrases, whereas GPT-4 includes some longer expressions, which causes it to perform worse than our model. Finally, for Task 2-1 (document proofreading), despite our model's superior performance, it often assumes that the given sentence has no errors, leading to missed corrections.

**Application tasks.** InternLM-Law achieves the best results on Tasks 3-1 to 3-4 but underperforms its counterparts on Tasks 3-5 to 3-8. We took a closer inspection of the outputs to understand our model's inferior performance. For instance, for Task 3-5 (Prison Term Prediction w. Article), our model performs worse than GPT-4. In addition, we found that providing legal articles during crime prediction worsens our model's results, causing it to achieve a lower score in Task 3-5 than in 3-4 (Prison Term Prediction w.o. Article). This indicates that our model is unable to effectively leverage the legal articles, whereas larger models like GPT-4 are able to utilize this information to improve their performance. For Task 3-6 (legal case

[12] Information about LawBench, including the definition of the tasks and the evaluation metrics, is shown in Appendix D.

[13] For the one-shot setting, the demonstration is randomly chosen from the LawBench training set. Task instructions for the LLMs can be found in Fei et al. (2024).

| Setting | 0-Shot | | | | | | | | 1-Shot | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Legal-specific LLMs | | | | General LLMs | | | API | Legal-specific LLMs | | | | General LLMs | | | API |
| Model | Ours | DISC | Lawyer | ChatL | Qwen | ILM2 | Qwen | GPT4 | Ours | DISC | Lawyer | ChatL | Qwen | ILM2 | Qwen | GPT4 |
| Size | 7B | 7B | 13B | 13B | 7B | 7B | 72B | N/A | 7B | 7B | 13B | 13B | 7B | 7B | 72B | N/A |
| 1-1 | **52.84** | 21.29 | 12.33 | 14.85 | 18.80 | 13.03 | 29.13 | 15.38 | **57.50** | 21.84 | 13.04 | 15.98 | 18.15 | 17.04 | 25.71 | 17.21 |
| 1-2 | 74.60 | 54.80 | 23.20 | 28.40 | 51.00 | 50.20 | **76.40** | 55.20 | **72.40** | 52.20 | 10.60 | 29.40 | 46.20 | 47.00 | 74.40 | 54.80 |
| Mem. | **63.72** | 38.05 | 17.77 | 21.63 | 19.16 | 31.62 | 52.77 | 35.29 | **64.95** | 37.02 | 11.82 | 22.69 | 19.00 | 32.02 | 50.06 | 36.01 |
| 2-1 | **57.27** | 12.23 | 4.33 | 12.22 | 12.00 | 36.78 | 26.91 | 12.53 | **57.27** | 13.44 | 4.90 | 13.01 | 14.51 | 36.78 | 35.01 | 18.31 |
| 2-2 | **61.00** | 20.20 | 8.25 | 2.68 | 31.80 | 39.20 | 48.60 | 41.65 | **62.40** | 21.40 | 19.20 | 9.00 | 22.80 | 40.00 | 44.20 | 46.00 |
| 2-3 | **90.29** | 62.48 | 15.88 | 42.24 | 46.86 | 54.52 | 62.05 | 69.79 | **90.06** | 66.02 | 9.03 | 30.91 | 51.29 | 49.55 | 65.35 | 69.99 |
| 2-4 | **49.00** | 41.60 | 4.40 | 27.60 | 39.20 | 43.80 | 39.00 | 44.00 | **49.00** | 42.80 | 3.00 | 26.60 | 40.00 | 41.80 | 40.60 | 44.40 |
| 2-5 | **87.38** | 60.20 | 34.61 | 39.11 | 62.57 | 47.21 | 66.47 | 56.50 | **86.75** | 62.92 | 39.65 | 41.41 | 64.60 | 61.61 | 78.46 | 64.80 |
| 2-6 | 56.19 | 7.70 | 41.65 | 54.89 | 20.83 | 51.50 | 75.53 | **76.60** | 56.14 | 32.70 | 36.33 | 60.68 | 61.40 | 64.95 | 73.83 | **79.96** |
| 2-7 | **53.21** | 33.71 | 38.51 | 38.45 | 30.59 | 33.60 | 34.81 | 37.92 | **53.03** | 25.16 | 37.10 | 42.41 | 33.47 | 37.12 | 42.11 | 40.52 |
| 2-8 | **83.80** | 27.20 | 9.60 | 18.60 | 38.40 | 43.20 | 54.40 | 61.20 | **81.40** | 20.20 | 0.40 | 20.20 | 39.00 | 44.80 | 57.60 | 59.00 |
| 2-9 | **91.09** | 84.89 | 29.78 | 31.74 | 58.65 | 63.89 | 70.55 | 78.82 | **91.39** | 81.60 | 33.19 | 40.27 | 62.96 | 66.54 | 74.71 | 76.55 |
| 2-10 | **88.89** | 14.08 | 2.38 | 14.56 | 16.37 | 36.32 | 43.29 | 65.09 | **88.31** | 14.45 | 6.12 | 17.37 | 22.41 | 40.18 | 37.35 | 65.26 |
| Und. | **71.81** | 36.43 | 18.94 | 28.21 | 39.19 | 45.00 | 52.16 | 54.41 | **71.58** | 38.07 | 18.89 | 30.22 | 43.62 | 48.33 | 54.92 | 56.48 |
| 3-1 | **75.59** | 43.96 | 0.60 | 33.28 | 57.53 | 63.79 | 72.42 | 52.47 | **75.55** | 65.61 | 0.33 | 25.99 | 53.57 | 64.15 | 73.79 | 52.20 |
| 3-2 | **47.82** | 38.70 | 25.94 | 31.55 | 31.93 | 14.12 | 29.67 | 27.54 | **47.56** | 39.77 | 27.23 | 33.96 | 33.86 | 29.35 | 36.10 | 33.15 |
| 3-3 | **68.13** | 50.21 | 31.30 | 27.90 | 45.35 | 48.91 | 57.07 | 41.99 | **68.62** | 57.22 | 19.36 | 12.24 | 44.91 | 51.03 | 60.01 | 41.30 |
| 3-4 | **84.22** | 72.07 | 74.19 | 76.18 | 79.26 | 81.42 | 81.32 | 82.62 | **83.83** | 75.41 | 70.99 | 74.31 | 80.86 | 80.11 | 80.77 | 83.21 |
| 3-5 | 80.05 | 77.19 | 75.52 | 73.57 | 79.53 | 80.11 | 79.95 | **81.91** | 79.37 | 75.72 | 73.56 | 73.01 | 78.02 | 80.21 | 79.11 | **82.74** |
| 3-6 | 63.60 | 51.00 | 17.80 | 28.80 | 45.00 | 39.60 | **70.40** | 48.60 | 65.60 | 46.40 | 6.60 | 26.80 | 47.20 | 41.40 | **68.80** | 49.60 |
| 3-7 | 66.00 | 42.80 | 39.20 | 41.40 | 43.00 | 55.40 | 74.80 | **77.60** | 64.80 | 51.20 | 33.80 | 42.00 | 42.40 | 56.60 | 75.00 | **77.00** |
| 3-8 | 23.17 | 15.63 | 16.94 | 17.17 | 19.51 | 19.32 | **24.30** | 19.65 | 22.37 | 13.76 | 16.02 | 16.72 | 19.84 | 20.42 | **24.67** | 19.90 |
| App. | **63.57** | 48.94 | 35.19 | 41.23 | 49.75 | 50.33 | 61.24 | 54.05 | **63.46** | 53.14 | 30.99 | 38.13 | 50.40 | 52.91 | 62.28 | 55.01 |
| AVG | **67.71** | 41.60 | 25.32 | 32.76 | 41.41 | 45.80 | 55.85 | 52.35 | **67.67** | 43.99 | 23.02 | 32.63 | 43.87 | 48.53 | 57.38 | 53.85 |

Table 2: Per-task zero-shot (left) and one-shot (right) results of the LLMs. The strongest results are **boldfaced**.

analysis), InternLM-Law underperforms Qwen-1.5-72B, which excels at synthesizing legal knowledge and concepts and achieves the best results. For Task 3-7 (criminal damages calculation), we found that our model is less effective in numerical calculations than larger models like GPT-4.

### 4.2 Human Evaluation

Recall that in Legal Consultation (LC), a model is expected to generate a response given a user query on a legal-related issue. We consider a challenging version of the LC task where the user queries are *open-ended* (i.e., they do not have a *model answer*). Due to the lack of a model answer, it is no longer possible to conduct an automatic evaluation. Hence, we conduct human evaluation on this task instead. To our knowledge, we are the first to perform a human evaluation of legal-specific LLMs.

#### 4.2.1 Evaluation Setup

**Dataset.** To conduct an evaluation on the LC task, we assembled a dataset consisting of 1000 open-ended questions collected from the legal section of the online platform *zhihu*[14] and manually refined some of the questions. We took care to ensure that none of these questions appears in the SFT data used to fine-tune InternLM-Law.[15]

| Models | LC |
|---|---|
| Lawyer-LLaMA | 0.0 |
| ChatLaw | 0.0 |
| DISC-Law | 27.4 |
| InternLM-Law | 76.8 |
| InternLM-Law (Stage 1 only) | 26.2 |
| InternLM-Law (Two steps Merged) | 26.4 |

Table 3: Results of human evaluation on the Legal Consultation (LC) task. Each line shows the percentage of times a model's output is manually rated as better than the GPT-4 output among the test instances.

**Evaluation methodology.** Since these open-ended questions do not have a correct answer, traditional evaluation methods where a system-generated response is compared against a human-generated response is no longer applicable. As a result, in our evaluation, we compare the response of each model for each question with that provided by GPT-4. Specifically, we asked two law school students to select the better response between the two, and calculated the *win rate* of a model compared to GPT-4, which is defined as the percentage of times the model's output is preferred to GPT-4's.

To evaluate a response, the judges were asked to

[14] https://www.zhihu.com/

[15] Detailed question examples can be found in Appendix E.

take into account five dimensions of quality, namely (1) *Relevance*, which measures how closely the response aligns with the given question, ensuring the answer addresses the issue at hand; (2) *Correctness*, which assesses whether relevant legal provisions are correctly cited and applied, including accurate interpretation and applicability to the given problem; (3) *Clarity*, which evaluates the consistency, tone, and professional quality of the response, ensuring it aligns with the standard communication style expected from legal professionals; (4) *Fluency*, which assesses how naturally and clearly the response is articulated without awkward phrasing, grammatical errors, or disruptions in the flow of ideas; and (5) *Detail*, which measures the depth and comprehensiveness of the response, ensuring that it covers all relevant aspects of the legal issue with sufficient explanation and supporting information.

The two judges rarely disagreed, and when inconsistencies did arise, it was usually because both had provided responses that are of similar qualities. They resolved these differences through discussion.

#### 4.2.2 Results and Discussion

Results of the legal-specific LLMs used in the automatic evaluation in Section 4.1 are shown in the first four rows of Table 3.[16] As can be seen, our model (row 4) outperforms GPT-4 on the legal consultation task, achieving a win-rate of 76.8%. We conducted a detailed comparison of our model and GPT-4 across the five dimensions. First, in terms of Relevance, both models effectively identified legal issues and provided appropriate responses. Similarly, both models generated fluent replies. However, in terms of Clarity, our model outperformed GPT-4 by adhering more closely to the logical reasoning of judges and adopting a response style more aligned with that of professional lawyers. GPT-4, in some cases, provided more general answers without thoroughly addressing the legal aspects. For example, when asked "What should I do if someone flees the scene after a traffic accident?", our model not only offers practical advice but also analyzes the legal responsibilities of the person fleeing and how they could be held accountable. In contrast, GPT-4 provides only basic suggestions for seeking help. This lack of detail contributes to GPT-4's scoring lower w.r.t. Detail. GPT-4 did generate analyses that included legal references, but some of the referenced laws contained factual inaccuracies. In contrast, our model cited laws more accurately and incorporated them into its legal reasoning. We hypothesize that the second stage of training helped our model learn the legal logic from expert-annotated data, which in turn contributed to its superior performance.

[16]The last two rows of the table show results of variants of InternLM-Law and will be discussed in Section 4.4.

| General Data used | Mem. | Und. | App. |
|---|---|---|---|
| None | 54.2/54.6 | 72.1/72.5 | 63.8/63.9 |
| 10% | 54.6/55.7 | 72.6/72.9 | 63.9/63.9 |
| 100% | **55.7/56.8** | **73.6/73.8** | **64.5/64.1** |

Table 4: Zero-shot/one-shot results of Stage 1 training of InternLM-Law on the memorization (Mem.) tasks, Understanding (Und.) tasks, and Application (App.) tasks in LawBench when different amounts of general data are mixed with the legal datasets for model training.

### 4.3 Long Text Evaluation

#### 4.3.1 Evaluation Setup

Although LawBench can test a model's performance on input texts composed of around 2K tokens, there is a strong expectation for LLMs to handle longer texts. For instance, the *Needles in a Haystack*[17] framework was proposed to test a model's ability to retrieve information from texts composed of up to 128k tokens. Motivated by the desire to evaluate a LLM's ability to handle long legal texts, we extended LawBench's Task 2-5 (reading comprehension) to create an enhanced version of this evaluation task that involves long texts.

We developed a dataset focused on analyzing Chinese legal judgments. This dataset includes 20 legal judgments crawled from a website[18], each having more than 20K characters, with three questions per judgment. These cases involve complex, multi-defendant, multi-charge scenarios, where a single defendant may have committed multiple crimes. We manually annotated three questions per judgment, targeting precise information within the documents. For example, given the question "What did a suspect steal in Shanghai on August 15, 2019?", the suspect may have committed multiple thefts in different locations over several years.

We aim for the model to accurately extract the key information from these complex cases. The model must accurately locate the relevant content for each question through reading comprehension *without* multi-hop reasoning. In essence, this

[17]https://github.com/gkamradt/LLMTest_NeedleInAHaystack

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

dataset allows us to evaluate a model's ability to recall information from the legal judgments. LMDeploy (Contributors, 2023) was used as the inference backend, with the input length set to 25K.

#### 4.3.2 Results and Discussion

Early models like Lawyer-LLaMA and ChatLaw have a maximum input length of 2K tokens, which prevents them from handling very long texts, and therefore fail to produce any outputs. Other large legal models, such as DISC-Law, are based on newer models and can accept input length of 4K tokens, but still cannot handle the long texts in our evaluation. In contrast, our model can process long texts and retrieve the necessary information from legal documents, achieves a F1 score of 84.73%.

### 4.4 Ablation Studies

Next, we conduct two ablation experiments.

**Usefulness of general datasets.** To better understand the impact of employing general data sources in Stage 1 training, we conduct an experiment where we incrementally reduce the amount of general data used for Stage 1 training.

Results of our model that went through only Stage 1 training are shown in Table 4, where in different rows different amounts of general data are used in the training process. For example, the last row shows the results when all of the general data is used for training, whereas the first row shows the results when no general data is used. As can be seen, for both evaluation settings, model performance on all three categories of tasks in LawBench deteriorates as the amount of general data used for model training decreases. We speculate that this may be because the general dataset allows the model to generalize some problem-solving capabilities to legal tasks, but additional experiments are needed to determine the reason.

**Usefulness of two-stage SFT.** To determine whether there are benefits to be had for using two-stage SFT, we conduct two ablation experiments. In the first experiment, we discard Stage 2 SFT, effectively training our model using Stage 1 SFT only. Zero-shot and one-shot results on the LawBench tasks are shown in row 2 of Table 5. For comparison purposes, we show in row 1 the results of our model trained with two-stage SFT. Comparing the two rows, we see that for both evaluation settings, discarding Stage 2 training improves model performance on LawBench's understanding and application tasks, but substantially hurts performance on memorization tasks. In row 5 of Table 3, we show the human evaluation results of our model using Stage 1 SFT on the LC data. Comparing these results with those of the model using two-stage SFT (row 4), we see that two-stage SFT offers substantially better results in the human evaluation.

| Models | Mem. | Under. | App. |
|---|---|---|---|
| Both Stages | **63.7/65.0** | 71.8/71.5 | 63.5/63.4 |
| Stage 1 only | 55.7/56.8 | **73.6/73.8** | **64.5/64.1** |
| Merged | 55.6/56.3 | 73.1/73.5 | 65.0/64.4 |

Table 5: Zero-shot/one-shot results of of InternLM-Law on the Memorization (Mem.), Understanding (Und.) and Application (App.) tasks in LawBench.

One may argue that it is not fair to compare the Stage 1 only model with a two-stage model because the latter is trained with more data than the former. In other words, the performance differences between the two models could be attributed to the amount of data they are trained on, not whether one or two stages are used. To address this concern, we conduct our second ablation experiment where we train our model using only one stage, but the data used in this one stage is created by merging the data used in both stages. Results of this experiment are shown in row 3 of Table 5. As can be seen, there are only minor performance differences between this model and the one trained using the Stage 1 data only. This suggests that as far as automatic evaluation is concerned, increasing the amount of data for Stage 1 training does not yield noticeable improvements. Similar trends can be observed for the human evaluation results, where there is no difference between the Stage 1 only version of our model and the "Merged" version.

## 5 Conclusion

We introduced InternLM-Law, a state-of-the-art large language model (LLM) for the Chinese legal domain. We presented a two-stage supervised fine-tuning approach to train a legal-specific LLM for Chinese and examined the impact of both general and legal datasets on training. Extensive automatic and human evaluations demonstrated that (1) our two-stage training approach successfully balanced legal consultation capabilities with performance on other legal tasks, and (2) general data enhances the model's legal capabilities, offering valuable insights for training and improving legal models. We make InternLM-Law and our code publicly available to encourage further research in this area.

## Acknowledgments

We thank Yining Li, Wenwei Zhang, Dahua Lin, Kai Chen, Fengzhe Zhou, and Maosong Cao for their valuable assistance with general datasets and testing. We also thank the three anonymous reviewers for their useful feedback on an earlier draft of the paper.

## Limitations

Although we have made every effort to reduce model hallucinations, our model, like other large language models, still produces hallucinations and inevitably generates some inaccurate responses. In addition, due to the relatively small model size, there is room for improvement on more complex legal reasoning tasks.

## Ethical Statements

Our research utilizes publicly available legal consultation data. No proprietary or restricted data was used, and the collection process strictly followed the guidelines provided by the data sources. All personally identifiable information (PII) was either removed or anonymized prior to analysis to ensure privacy and confidentiality.

## References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianwei Zhang Jianhong Tu, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. InternLM2 technical report. *arXiv preprint arXiv:2403.17297*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024. SaulLM-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

LMDeploy Contributors. 2023. LMDeploy: A toolkit for compressing, deploying, and serving LLM. https://github.com/InternLM/lmdeploy.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023b. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Jidong Ge, Yunyun Huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. 2021. Learning fine-grained fact-article correspondence in legal cases. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3694–3706.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv preprint arXiv:2406.12793*.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer LLaMA technical report. *arXiv preprint arXiv:2305.15062*.

Yunyun Huang, Xiaoyu Shen, Chuanyi Li, Jidong Ge, and Bin Luo. 2021. Dependency learning for legal judgment prediction with a unified text-to-text transformer. *arXiv preprint arXiv:2112.06370*.

Vladimir Iosifovich Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. In *Doklady Akademii Nauk*, volume 163, pages 845–848. Russian Academy of Sciences.

Hongcheng Liu, Yusheng Liao, Yutong Meng, and Yuhao Wang. 2023. LawGPT: Chinese legal dialogue large language model. https://github.com/LiuHC0428/LAW_GPT.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Shiguang Wu, Zhongkun Liu, Zhen Zhang, Zheng Chen, Wentao Deng, Wenhao Zhang, Jiyuan Yang, Zhitao Yao, Yougang Lyu, Xin Xin, Shen Gao, Pengjie Ren, Zhaochun Ren, and Zhumin Chen. 2023. fuzi.mingcha. https://github.com/irlab-sdu/fuzi.mingcha.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

Zongyue Xue, Huanghai Liu, Yiran Hu, Yuliang Qian, Yajing Wang, Kangle Kong, Chenlu Wang, Yun Liu, and Weixing Shen. 2024. Leec for judicial fairness: A legal element extraction dataset with extensive extra-legal labels. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7527–7535. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 183–201.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. 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*, pages 5218–5230.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9701–9708.

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. *arXiv preprint arXiv:2406.04614*.

## A Tasks and Data Sources

Detailed information of the data sources in Table 1 is listed below:

- FLK: FLK is a national database[19] comprehensively collects Chinese laws and regulations, including the Constitution of the People's Republic of China, civil laws, local regulations, etc.
- JEC_QA: JEC_QA (Zhong et al., 2020b) is the largest question answering dataset collected from the National Judicial Examination of China. All data can be accessed from http://jecqa.thunlp.org/.
- CAIL: CAIL (Challenge of AI in Law)[20] is a competition website related to law, which aggregates many test tasks in the Chinese judicial field.
- LAIC: LAIC(Legal AI Challenge)[21] is another competition website about legal tasks that offer different competition tasks distinct from those on CAIL.
- AIStudio: AI Studio[22] is a learning platform for deep learning that offers extensive open datasets, including some relevant to legal. We constructed 2-3 tasks from the public dataset[23] on this platform.
- CrimeKgAssitant: CrimeKgAssistant[24] is an open-source crime assistant github project. This dataset consists of 856 pieces of crime knowledge graphs, a 2.8 million crime prediction training dataset, 200k legal Q&A pairs, and a 13-category topic classification for these 200k legal consultation questions.
- LawGPT: LawGPT (Liu et al., 2023) is an open-source Chinese legal large model github project. In this project, they public the training dataset but do not release the trained Chinese legal large language model. The training dataset includes scenario dialogues between lawyers and users, some of which are cleaned

[19] https://flk.npc.gov.cn/
[20] http://cail.cipsc.org.cn/
[21] https://laic.cjbdi.com/
[22] https://aistudio.baidu.com/index
[23] https://aistudio.baidu.com/datasetdetail/181754
[24] https://github.com/liuhuanyong/CrimeKgAssitant

from the publicly legal data CrimeKgAssistant, while others were generated by utilizing ChatGPT to conceive specific question-answering scenarios based on 9,000 key legal provisions thereby ensuring that the generated dataset has concrete legal grounds.

- hualv.com: hualv.com[25] is an online platform dedicated to providing legal consultation services, where numerous real users and lawyers engage in daily interactions by asking and answering questions. Topics of these conversation data range from marriage-related questions, labor disputes, contract controversies, etc. All data on this platform is public available and can be accessed through web scraping.

- LEVEN: LEVEN (Yao et al., 2022) is the largest Legal Event Detection(LED) dataset with 8, 116 legal documents and 150, 977 human annotated event mentions in 108 event types. Not only charge-related events, LEVEN also covers general events, which are critical for legal case understanding but neglected in existing LED datasets.

- AC_NLG: AC_NLG (Wu et al., 2020) is a paper containing a dataset based on raw civil legal documents, where each case is objectively split into three parts: plaintiff's claim, fact description, and court's view with human annotation on the judgment.

- LEEC: LEEC (Xue et al., 2024) is a comprehensive, large-scale criminal element extraction dataset, comprising 15754 judicial documents and 128 labels. It is the most extensive and domain-specific legal element extraction dataset in China.

The tasks in the first 20 rows of Table 1 align with the task definitions of LawBench. For more details, readers can refer to Table 7. Additionally, we introduce two extra tasks: Judgement Generation and Legal Element Extraction.

- Judgement Generation: This task involves generating the court's reasoning for the case based on the fact section of the judgment. We use the plaintiff's claim and fact description from the AC_NLG dataset as input, and we manually craft instructions to guide the model in generating the court's view.

- Legal Element Extraction: This task involves extracting predefined legal elements, such as the name, gender, and nationality of the defendant, from judicial documents. We designed the instructions by drawing on the work of (Xue et al., 2024). In this task, the large language model is instructed to extract legal elements from the given segments of judicial documents and produce a structured response.

## B Legal Website URLs

Our legal consultation dataset comes from the websites list in Table 6.

| Legal Website URLs |
|---|
| https://www.12348.gov.cn/ |
| https://www.66law.cn/ |
| https://china.findlaw.cn/ |
| https://www.dalvlaw.com/ |
| https://www.lawtime.cn/ |

Table 6: Website URLs used for collecting the legal consultation dataset.

## C Applications

Next, we provide examples of our model in various scenarios, including legal consultation, legal tasks, and the ability to write code to solve law-related computational problems. These examples demonstrate the model's capabilities under different instructions. We compare its performance with DISC-Law, showcasing our model's performance across different application scenarios.

**Legal Consultation.** Consumer rights are a common topic of legal consultation, and our model can be a legal advisor to answer various questions in wide range of domains. For example, model can help you claim compensation for personal injury, give advice in employment contract, and protect your intellectual property. We illustrate our model's performance on a consumer rights question in Figure 2. In legal consultations, the user's queries are often vague and do not clearly articulate the facts of the case. Our model responds to this query by first listing possible scenarios, then attempting to analyze the legal issues in each scenario and identifying potentially relevant statutes. It then concludes by summarizing the possible scenarios. We also present the performance of another legal

[25] www.66law.com

model, DISC-Law, on this issue. Although it also considers the potential liability of the restaurant and third parties, it does not provide the relevant laws and regulations in its responses, and its replies are not as well-structured as those of our model.

**Legal NLP Tasks.** Our model is capable of dealing with various legal NLP task, including memorization, understanding and also apply legal knowledge. The tasks cover different types of NLP tasks, classification, extraction, regression and also generation. We demonstrate our model's performance on a legal NLP task: predicting the charge of a crime based on the facts of the case. See Figure 3 for details. In this task, our model has a good understanding of the case and accurately analyzes the charges by utilizing corresponding legal knowledge. Furthermore, it will follow the instructions and provide answers that conform to the specified format. We also present the result of DISC-Law. Although it answers the question correctly, it does not follow the instructions to output the answer in the given format. This demonstrates that our model outperforms other legal models in instruction following.

**Tool Usage.** Our training strategy not only enables the model to retain its existing capabilities but also facilitates the integration of general competencies with newly acquired legal expertise, thereby enhancing the model's performance on legal tasks. We tested the model by writing code to solve amount calculation problems in LawBench. Detailed examples are shown in Figure 4. It is worth noting that none of the data points in our training dataset are examples of solving amount calculation problems using code. And by using programming methods, we find that model can improve its accuracy in the task compared to generating its final output directly. As our example demonstrates, our model successfully wrote a code snippet to solve the criminal damages calculation problem. Model accurately extracted the relevant amounts from the text and wrote executable Python code, demonstrating the model's ability to perform numeric computations. We demonstrate the performance of DISC-Law on this task. This legal model does not accurately understand the task. It analyzes the case based on its own logic, assigns charges, provides references to legal statutes, and finally gives what it considers a possible total amount. Although the final amount is correct, it neither writes code nor provides the calculation process, and most of the response is irrelevant to the task. This shows that other legal large language models not only lack tool usage capabilities but also perform poorly in terms of instruction following.

## D LawBench

There are three categories of tasks in LawBench, with each category corresponding to one cognitive level, and indicating the model's capacity to deal with legal knowledge tasks. We present the definition, types, and metrics for each task in Table 7.

*Knowledge Memorization* is the lowest cognitive level task in the benchmark. It examines the model's ability to accurately memorize laws and regulations as well as some legal knowledge. For detailed information, please refer to Table 7 ID 1-1 and 1-2.

*Knowledge Understanding* is the second level in the cognitive hierarchy, evaluating the extent to which LLMs can comprehend entities, events, and relationships within legal texts. This category includes 10 tasks, with detailed information available in Table 7, ID 2-1 to 2-10.

*Knowledge Application* is the most challenging category in the benchmark. It tests the ability of LLMs to not only understand legal knowledge but also to simulate legal professionals in applying this knowledge to solve real-world legal tasks. These tasks extensively evaluate the model's reasoning abilities, including three legal content reasoning tasks: legal judgment prediction (3-1 through 3-5), case analysis (3-6), one numerical reasoning task: criminal damages calculation (3-7), and consultation (3-8).

## E LC Questions for Human Evaluation

Below we present a sample open-ended legal consultation question:

*Does a person with a formal job violate labor laws by earning money through live streaming?*

This question requires the model to conduct a detailed analysis of various possible scenarios to explain under which circumstances it would be a violation of labor laws and in which cases it would not.

Example of Legal Consultation

**Question:**
Can I claim compensation from the restaurant if I break teeth while eating there?

**Model Answer:**
**DISC-LAW**
If a customer is injured by slipping and falling while dining in a restaurant, they can request the restaurant to take responsibility. According to legal regulations, the operator should take necessary safety precautions to prevent harm to others.
If they fail to fulfill their safety obligations, they should bear tort liability. If the injury to the customer is caused by a third party, then that third party should bear tort liability. However, if the restaurant as the operator has fulfilled its safety obligations, it does not need to bear responsibility.

**Ours**

When dining at a restaurant and breaking teeth, you can consider the following points to determine who should be compensated:

1. **Responsibility of the restaurant**: If you broke your teeth due to the presence of hard objects (such as bones, stones, etc.) in the food provided by the restaurant during normal dining, then the restaurant may be liable for compensation. According to Article 148, Paragraph 2 of the "Food Safety Law of the People's Republic of China", in cases where food products that do not meet food safety standards are produced, or knowingly sold by businesses, consumers can demand compensation for losses and may also require the producer or operator to pay ten times the price of the products or three times the compensation for losses.

2. **Providing evidence**: To protect your rights, you need to provide relevant evidence, such as retaining the food that caused your tooth damage, medical records, and expense receipts, to prove the causal relationship between your tooth damage and your dining experience at the restaurant.

3. **Responsibility of third parties**: If your tooth damage is caused by the negligence of other customers or restaurant staff, then the responsibility may lie with that third party. In this case, you can first negotiate with the third party for a resolution, and if negotiation fails, you can consider legal recourse to protect your rights.

If you believe the restaurant should be responsible for your tooth damage, you can first negotiate with the restaurant. If negotiation fails, you can seek legal assistance or file a complaint with the consumer association to protect your legal rights. Maintain calm and rationality during the problem-solving process, and retain relevant evidence.

Figure 2: An example of legal consultation.

Example of Legal NLP Tasks

**Question:**
Please simulate a judge and provide the charge based on the following fact. Only output the name of the charge, and place the answer between **<charge>** and **<eoa>**. For example, **<charge>** theft, scam **<eoa>**. Please follow the format strictly.
Fact: The People's Procuratorate of Da Ming County accuses the defendant, Zhang Mou, of a crime. On March 25, 2014, at around 9 a.m., the defendant Zhang Mou got into a dispute with Ximou, a villager from his own village, at his home in Zhuangmou, due to certain reasons, which escalated into a physical fight. During the fight, the defendant Zhang Mou injured Ximou and his sons, Ximou B and Ximou C, using a crowbar. According to the appraisal by the Physical Evidence Appraisal Room of the Da Ming County Public Security Bureau, Ximou's injuries were classified as second-degree minor injuries, while Ximou B and Ximou C sustained minor injuries. To substantiate the aforementioned accusations, the prosecuting authority presented evidence including the victim's statement, the defendant's confession, witness testimonies, forensic appraisal reports on the extent of bodily injuries, on-site investigation records, and photographs. The prosecuting authority believes that the defendant Zhang Mou intentionally caused bodily harm to others, resulting in minor injuries, thus violating the provisions of the Criminal Law of the People's Republic of China. The facts of the crime are clear, and the evidence is reliable and sufficient, warranting criminal liability under the relevant provisions. Additionally, the plaintiff and their legal representative in the accompanying civil litigation do not contest the characterization of this case and request the defendant to compensate for the economic losses incurred, totaling 120,960.6 yuan, due to their actions. They have submitted relevant evidence, including medical bills, appraisal fees, diagnosis certificates, and hospitalization records.

**Models Answers:**
**DISC-LAW**
intentional injury

**Ours**
**<charge>**intentional injury **<eoa>**

Figure 3: An example of legal NLP Tasks.

| ID | Definition | Type | Metric |
|---|---|---|---|
| 1-1 | Article Recitation: Given a law article number, recite the article content. | Generation | Rouge-L |
| 1-2 | Knowledge Question Answering: Given a question asking about basic legal knowledge, select the correct answer from 4 candidates. | SLC | Accuracy |
| 2-1 | Document Proofreading: Given a sentence extracted from legal documents, correct its spelling, grammar and ordering mistakes, return the corrected sentence | Generation | F0.5 |
| 2-2 | Dispute Focus Identification: Given the original claims and responses of the plaintiff and defendant, detect the points of dispute. | MLC | F1 |
| 2-3 | Marital Disputes Identification: Given a sentence describing marital disputes, classify it into one of the 20 pre-defined dispute types. | MLC | F1 |
| 2-4 | Issue Topic Identification: Given a user inquiry, assign it into one of pre-defined topics. | SLC | Accuracy |
| 2-5 | Reading Comprehension: Given a judgement document and a corresponding question, extract relevant content from it to answer the question. | Extraction | rc-F1 |
| 2-6 | Named-Entity Recognition: Given a sentence from a judgement document, extract entity information corresponding to a set of pre-defined entity types such as suspect, victim or evidence. | Extraction | soft-F1 |
| 2-7 | Opinion Summarization: Given a legal-related public news report, generate a concise summary. | Generation | Rouge-L |
| 2-8 | Argument Mining: Given a plaintiff's perspective and five candidate defendant's viewpoints, select one viewpoint that can form a point of dispute with the plaintiff's perspective. | SLC | Accuracy |
| 2-9 | Event Detection: Given a sentence from a legal judgement document, detect which events are mentioned in this sentence. | MLC | F1 |
| 2-10 | Trigger Word Extraction: Given a sentence from a legal judgment document and its corresponding events, predict which words in the sentence triggered these events. | Extraction | soft-F1 |
| 3-1 | Fact-based Article Prediction: Given a fact statement from the legal judgement document, predict which article items should be applied. | MLC | F1 |
| 3-2 | Scene-based Article Prediction: Given a described scenario and a related question, predict the corresponding article item. | Generation | Rouge-L |
| 3-3 | Charge Prediction: Given fact statement from the legal judgement document and the applied article number, predict the cause of action (charge). | MLC | F1 |
| 3-4 | Prison Term Prediction w.o. Article: Given fact statement from the legal judgement document, the applied article number and charge, predict the prison term. | Regression | nLog-distance |
| 3-5 | Prison Term Prediction w. Article: Given fact statement from the legal judgement document, the applied article content and charge, predict the prison term. | Regression | nLog-distance |
| 3-6 | Case Analysis: Given a case and a corresponding question, select the correct answer from 4 candidates. | SLC | Accuracy |
| 3-7 | Criminal Damages Calculation: Given a fact description about a criminal process, predict the amount of money involved in this case. | Regression | Accuracy |
| 3-8 | Consultation: Given a user consultation, generate a suitable answer. | Generation | Rouge-L |

Table 7: Details of the definition, type, and metric of each task. LawBench contains 5 task types: generation, single-label classification (SLC), multi-label classification (MLC), regression, and extraction.

Example of Tool Usage

**Question:**
Please write a code to calculate the amount of money involved in the following case: The public prosecution accuses the defendant Meng of stealing RMB 1400 and an Apple iPhone 4 mobile phone valued at RMB 1300 from the victim Xiao's pocket and computer desk, respectively, while Xiao was washing up in the bathroom of Room 205 at the "Xiangyun Hotel" at 114 Gongnong Road, Dadong District, Shenyang City at around 8 am on July 14, 2013. The stolen money has been spent. Meng was arrested by the public security organs on July 19, 2013. The above facts were not objected to by defendant Meng during the court trial. There is evidence such as proof of arrest, physical evidence photos, detention decision, detention list, return list, basic information of permanent residents, phone search records, and identification records, which are sufficient to establish the case.

**Model Answer:**
**DISC-LAW**
This case involves the crime of theft. According to Article 264 of the Criminal Law, for theft of public or private property, if the amount is relatively large, the offender shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance, and shall also be fined; if the amount is huge or there are other serious circumstances, the offender shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years, and shall also be fined; if the amount is especially huge or there are other especially serious circumstances, the offender shall be sentenced to fixed-term imprisonment of not less than ten years or life imprisonment, and shall also be fined or have their property confiscated.
Based on the facts of the case, the defendant Meng stole 1400 yuan in cash and a mobile phone worth 1300 yuan from the victim. Therefore, it can be concluded that the defendant Meng is guilty of theft.
According to Article 264 of the Criminal Law, since the specific amount is not mentioned in this case, it cannot be determined whether it falls under the category of "relatively large amount", "huge amount", or "especially huge amount". However, considering that the total value of the stolen items is 2700 yuan, it can be preliminarily judged that the theft in this case falls under the category of "relatively large amount".
In conclusion, based on the provisions of Article 264 of the Criminal Law, it is recommended that the defendant Meng be sentenced to fixed-term imprisonment of not more than three years and be fined.
Calculated crime amount: 2700 yuan

**Ours**
According to the information provided, the following Python code can be written to calculate the involved amount:

```
# Involved amount
amount_1 = 1400
amount_2 = 1300

# Total amount involved
total_amount = amount_1 + amount_2

# Print total amount involved
pirnt("Total amount involved: ", total_amount)
```

Figure 4: An example of Tool Usage.