Legal-LM: Knowledge Graph Enhanced Large Language Models for Law Consulting

Juanming Shi[1], Qinglang Guo[2], Yong Liao[2(✉)], Yuxing Wang[1], Shijia Chen[1], and Shenglin Liang[3]

[1] University of Science and Technology of China, Heifei, China
sjm2022@ustc.edu.cn
[2] China Academic of Electronics and Information Technology, Beijing, China
gql1993@mail.ustc.edu.cn, yliao@ustc.edu.cn
[3] Xidian University, Xian, China

Abstract. This paper introduces Legal-LM, an advanced Large Language Model (LLM) enhanced with a Knowledge Graph, specifically designed for legal consulting in the Chinese legal domain. Addressing the challenges of domain-specific adaptation, data veracity, and consultations with non-professional users in legal AI, Legal-LM incorporates extensive legal corpora and a knowledge graph for effective legal knowledge acquisition. The model utilizes techniques such as external legal knowledge basis, soft prompts, and Direct Preference Optimization (DPO) to ensure accurate and diverse legal advice. Our experimental results demonstrate that Legal-LM exhibits superior performance over existing models in legal question answering, case analysis, and legal recommendations, these show its potential to facilitate legal consulting and education.

Keywords: Legal-LM · Knowledge Graph · Law Consulting · Large Language Models · Legal AI Applications

1 Introduction

Over the past decades, the rise of Legal Artificial Intelligence (LegalAI) has begun to reshape the future of the legal field [12]. Changes in this field are not only reflected in the automation of legal tasks such as legal information extraction [2], interactive argument pair extraction [16, 17, 31], case retrieval [23], judgment prediction [3, 28, 30], and legal question answering [18], but also in how they make legal services more inclusive. Intelligent legal systems benefit reducing the paperwork burden of legal professionals, and providing the general public with convenient access to legal services and remote counseling. In addition, these technologies provide valuable resources for legal learners for their studies and practices.

The early development of LegalAI focused on building specialized datasets and designing machine learning algorithms to perform specific legal tasks. Although these initial attempts yielded some good results, their applications were relatively limited in scope. Recently, rap-id-developing technologies in large-scale language models (LLMs)

D.-S. Huang et al. (Eds.): ICIC 2024, LNAI 14878, pp. 175–186, 2024.
https://doi.org/10.1007/978-981-97-5672-8_15

[23] have a great impact on the legal field. These models have demonstrated superior cross-domain adaptability and achieved remarkable success in the accurate understanding and execution of legal instructions. Especially when these models are tuned with legal domain-specific tuning method such as Vicunna [33] and ChatGLM [9] etc., they are good at handling various types of legal tasks which include but are not limited to instrumental analysis, case interpretation, and legal counseling [11, 20, 33] etc..

Adapting Large Language Models (LLMs) to specialized domains like law is challenging due to limited data and inadequate training approaches. We identify three main challenges: a lack of domain-specific knowledge since LLMs are primarily trained on general corpora like C4 [27] and Wikipedia; the issue of data illusions necessitating high accuracy and the ability to retrieve and reason with external legal knowledge; and non-professional user consultations that often result in less effective responses due to unspecialized questions.

To address these challenges, we have developed a new framework, Legal-LM, tailored for the Chinese legal domain, as shown in Fig. 1. This involves several approaches:

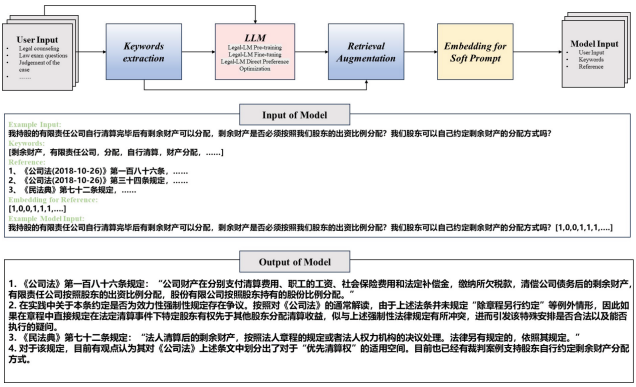

Fig. 1. Legal-LM's framework, first extract the user's input keywords, and then the user input into the Legal-LM model after pre-training, fine-tuning, Direct Preference Optimization, and at the same time based on the keywords and user input retrieval, retrieval of the content embedded in the form of the input, the bottom two frames show an example of input and output model.

1. Pre-training for the base model: We compiled a vast array of raw legal texts, such as legal articles and judicial interpretations, which were initially noisy and low in knowledge density. To enhance learning, we cleaned these texts and integrated structured data like legal knowledge graphs before pre-training the base model. This step ensures the model gains substantial legal knowledge.
2. Keywords extraction and Direct Preference Optimization: To tackle non-professional consultations, we implemented a method to distill legal knowledge from user questions using the GPT interface, which helps the model grasp the issues more precisely. Additionally, we applied the Direct Preference Optimization (DPO) [26] method into

our training process to cater to diverse legal response preferences, enhancing the variability and appropriateness of the advice given.

3. External legal knowledge base with soft prompt: We added an information retrieval module to address data illusions by filtering out irrelevant information and ensuring more reliable responses. This module retrieves legal statutes relevant to user queries, with an innovative prompt escape layer that uses soft prompts to mitigate issues typically associated with LLM responses.

2 Approach

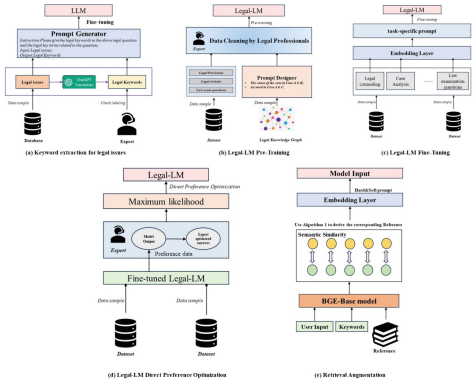

Fig. 2. The five strategies for the Legal-LM legal large language model are (a) Keyword extraction for legal issues, (b) Legal-LM Pre-Training, (c) Legal-LM Fine-Tuning, (d) Legal-LM Direct Preference Optimization and (e) Retrieval Augmentation.

2.1 Overview

As illustrated in Fig. 2, this section outlines five strategies to optimize legal question-answering using LLMs, including keyword extraction for legal issues (Sect. 2.2), LLM pre-training specific to legal aspects (Sect. 2.3), LLM fine-tuning specific to legal aspects (Sect. 2.4), LLMdpo specific to legal aspects (Sect. 2.5), and retrieval enhancement including legal aspects (Sect. 2.6).

2.2 Keyword Extraction for Legal Issues

Legal keyword extraction from user questions is crucial for minimizing the risk of unreliable and deceptive outputs, as shown in Fig. 2(a). This process involves two main components: (1) initializing the LLM with specific instructions and selecting the appropriate small batch parameter LLM, and (2) fine-tuning the LLM using a high-quality dataset.

To effectively guide the LLM, we crafted five prompts to direct the GPT in extracting keywords from legal problems. These prompts were processed by the GPT, which integrated and optimized them based on feedback from legal professionals. For fine-tuning, we utilized small batch-parameter LLMs including Qwen-1.8B [1], BART [5], and ChatGLM-6B [9]. Among these, Qwen-1.8B was selected for its exceptional performance, particularly in handling the Chinese language, due to its optimal parameter size.

For constructing the fine-tuned LLM high-quality dataset, we initially selected $X = \{x_1, x_2, \ldots, x_n\}$ from a dataset of local samples. The dataset capable of generating at least five or more keywords and legal terms was selected through screening by legal professionals such as lawyers. Then, using the GPT interface and the aforementioned prompts, each data's keywords and corresponding legal terms, $Y = \{y_1, y_2, \ldots, y_n\}$, were generated. Legal professionals tested and corrected these to produce the final output $Y = \{y'_1, y'_2, \ldots, y'_n\}$. . Finally, the dataset [prompt]- $X = \{x_1, x_2, \ldots, x_n\}$- $Y = \{y'_1, y'_2, \ldots, y'_n\}$ was fine-tuned for Qwen-1.8B to complete the keyword extraction of legal issues. With this approach, we aim to enhance the overall performance of the legal macromodel and provide users with reliable and trustworthy information. This system is also applicable to other large language models with robust context learning capabilities. Furthermore, the generated prompts lay a foundational basis for the development of subsequent algorithms.

2.3 Legal-LM Pre-training

Among natural language tasks, current LLMs in the legal domain struggle with complex legal tasks due to a lack of complete legal knowledge. They often rely on superficial legal descriptions instead of integrating external legal terms. As depicted in Fig. 2(b), the primary strategy for pre-training LLMs in legal fields involves two stages. First, a vast array of legal data including articles, clauses, exams, and knowledge graphs is compiled and organized to instruct LLMs on answering Chinese legal questions. Second, a novel model design is developed to improve the accuracy and clarity of LLM outputs.

Previous studies [19] indicate that models trained on domain-specific corpora perform better on complex legal tasks than those trained solely on general corpora. In addition to collecting general Chinese corpora such as C4, BBC, TH, and CCL, we also gathered texts from legal websites, which were then manually cleaned by professionals to reduce data noise. Moreover, structured data like legal knowledge graphs—less noisy and denser in knowledge—were incorporated. This specialized legal data enhances the pre-training process.

While LLMs cannot ensure 100% task accuracy, some legal tasks, such as quizzes and consultations, demand high accuracy. To ensure valid outputs, we adopted a hybrid approach. For legal advice questions, we created specialized ternary training sets (Reference, Legal Issue, Output) to optimize valid output generation. After pre-training, we conducted tests with 100 sample data points; any inaccuracies identified were corrected by legal professionals and the data reformatted (Legal Issue-Wrong Answer-Correct Answer) for further training. The pre-training phase is only deemed complete when tests achieve full accuracy, enhancing the precision and reliability of LLM outputs for legal tasks.

2.4 Legal-LM Fine-Tuning

Fine-tuning significantly enhances LLM performance, dependent on the quality and volume of labeled data. Figure 2(c) illustrates that this process involves training both non-text and natural language embedding, utilizing feedback from LLM applications. Innovative fine-tuning methods are also under exploration. In order to handle legal queries that exceed input length limits, we designed a compatible embedding layer for structured data, which is trained to integrate seamlessly with the LLM during fine-tuning. We utilize a diverse set of data sources for fine-tuning, including public NLP legal datasets like JEC-QA [34] and CJRC [10], open-source instruction sets such as Lawyer-llama [15] and LawGPT-zh [24], and real lawyer-user communications, supplemented by general datasets like gpt4_data_zh and Firefly [29] to prevent over-specialization to legal contexts.

The huge scale of LLMs makes traditional fine-tuning resource-intensive. An efficient alternative involves tuning a small fraction of the parameters, such as adding adapter modules (small MLPs) to each transformer layer [10], adjusting only these during fine-tuning while keeping the original LLM parameters intact. This approach not only cuts down on storage needs but also allows for the storage of adapter module parameters only, keeping a single copy of the original LLM. Fine-tuning optimizes parameters from θ to θ', with the change $\Delta\theta = \theta' - \theta$ represented in a lower-dimensional form, $W = BA$, where W represents the weight matrix of $\Delta\theta$, and B and A are the weight parameters that are actually tuned [11]. This method reduces the number of parameters tuned and enhances the efficiency of the fine-tuning process, addressing the challenges of scaling LLMs and facilitating more effective model optimization.

2.5 Legal-LM Direct Preference Optimization

To align with users' preferences in legal scenarios, we have integrated the Direct Preference Optimization (DPO) reinforcement learning method into our training process. This method not only diversifies the range of legal advice responses but also enhances their linguistic quality, as shown in Fig. 2(d). The key goal is to build a high-quality dataset that increases the LLM's sensitivity to legal users' preferences and improves response quality.

In legal consulting, DPO algorithms effectively capture and optimize the complex and varied preferences of clients. By directly optimizing decision-making preferences, DPO helps provide more accurate and efficient advice amidst various legal options. For the DPO dataset, we use a format of [Legal Issue]-[Reject-Answer]-[Chosen-Answer], where the Reject-Answer is derived from the fine-tuned model's output and the Chosen-Answer is a lawyer's correction. This format ensures the dataset closely matches the nuanced preferences of legal clients, enabling the LLM to offer more personalized and precise legal advice.

2.6 Retrieval Augmentation

In scenarios such as legal advice and judgment prediction, users expect responses from models to be robustly supported by legal precedents and statutes. Despite being fine-tuned with high-quality legal data, the LLM might still generate inaccurate responses

due to data illusions or outdated knowledge. To tackle this, as shown in Fig. 2(e), we developed a legal terms retrieval algorithm to reinforce responses with solid legal backing.

We established a knowledge base with over 50 categories of Chinese law, including the Constitution, Criminal Law, and Patent Law, encoded as vectors and stored locally. Upon receiving a user input, our algorithm retrieves the Top-K most relevant documents from this base by measuring their similarity to the input. These documents are then processed through a prompt escaping layer, transforming them into a series of soft-prompts represented by consecutive numbers. These prompts are fed into the large legal model alongside the user's question. By referencing these documents, the model enhances its understanding of the input, leading to more accurate and trustworthy answers.

Algorithm 1 Legal retrieval based on Large Langu Model keyword extraction

1. Initialize the BGE-Base model for embedding model.
2. Initialize the Legal provisions database as N, where $l_i \in N$ and i represents the i-th Legal provisions. Let M be the number of Legal provisions in the Legal provisions database.
3. Initialize the Legal provisions scores as S, where $s_i \in S$ represents the score corresponding to the i-th Legal provision, all initialized to 0. The number of elements in S is also M.
4. Input the user's question q into the BGE model and obtain a vector q 'for question q
5. Extracting keywords from user queries using Fine-tuned Qwen-1.8B model, and then inputting each keyword into a BGE-Base model to obtain a collection of K keyword vectors, where k_i represent the vector for the i[th] keyword.
6. $q'' = \frac{q'}{||q'||}$
7. for a to M do
8. $\quad s_a \leftarrow s_a + cossim(q'', l_i)$
9. end for
10. return Top2K(S)
11. Reinitialize the legal provisions for the Top2k score and set them to S', where $s_i' \in S'$ represents the score corresponding to the i-th Legal provision, all initialized to 0. The number of elements in S is $2K$.
12. Initialize α as the weight of each keyword in question q for k_i.
13. for x to K do
14. $\quad u_x = \frac{k_x}{||k_x||} + \alpha \cdot q''$
15. $\quad$ for y to 2K do
16. $\quad\quad s_y \leftarrow s_y + cossim(u_x, l_y)$
17. $\quad$ end for
18. end for
19. return TopK(S)
20. These K legal provisions will be converted into consecutive numbers through a prompt escape layer, which will be used as soft prompt combined with user questions as input for the legal large language model.

3 Experiments

To assess the capabilities of Legal-LM, we conducted baseline experiments focusing on the model's overall performance and its efficacy in answering subjective legal advice questions. Additionally, we investigated the importance of each model component, yielding relevant experimental outcomes that were subsequently analyzed.

3.1 Data Setup

For our assessment, we utilized two types of datasets to ensure diversity and representativeness: an objectively judged dataset and a subjective dataset. The objective dataset comprised three sets of legal questions categorized by difficulty: LBK [54] for easy questions from the Legal Basic Knowledge Question Bank, UNGEE [54] for medium-difficulty questions from the National Unified Examination for the Specialized Master's Degree of Laws, and NJE [54] for the most challenging questions from the National Unified Legal Professional Qualification Examination. These datasets included both single-choice and multiple-choice questions to evaluate the model's handling of diverse question types, as detailed in Table 1. For the subjective dataset, our attorneys manually constructed a high-quality test set comprising 1,000 examples from legal advice, online forums, justice-related publications, and legal documents, covering scenarios such as legal quizzes, legal counseling, and verdict prediction.

3.2 Training Setup

To validate the effectiveness of our proposed Le-galGPT design, we conducted a benchmark evaluation of legal large language models and various baseline models across different research areas. For the benchmark evaluation, we selected two types of models: the Chinese Large Language model and the Chinese law large language model. The Chinese base large language model lineup includes Ziya-LLaMA [21], ChatGLM [9], Baichuan [29] and Qwen [1]; the Chinese legal large language model comprises LexiLaw [7], HanFei, LawGPT [24], Lawyer llama [15] and ChatLaw [6].

3.3 Evaluation Protocols

For a comprehensive assessment of our Legal Large Model, we employed an accuracy metric on an objective rubric dataset for precise performance measurement. Subjectively, we simulated legal examination environments using a high-quality test set constructed from diverse legal sources, including online forums and legal documents. This test set was evaluated by three legal professionals who scored the model's outputs based on accuracy, completeness, clarity, and language quality, with scores averaged for the final assessment.

In this context, accuracy measures the correctness of the responses in adhering strictly to the question's requirements. Completeness evaluates whether the response thoroughly covers all aspects of the question, demonstrating the candidate's understanding. Clarity assesses the logical structure and comprehensibility of the answer, important for

showcasing logical thinking. Language quality checks grammatical precision and the natural flow of language, reflecting the candidate's linguistic proficiency. These criteria are crucial as slight errors in law can lead to significant consequences. The Law Large Model, implemented in PyTorch, was benchmarked against various models to verify its effectiveness.

3.4 Result

3.4.1 Comparison of Overall Performance

Table 1. Results of the overall performance evaluation of Legal-LM and comparative baseline models on three experimental datasets

Model	LBK	UNGEE	NJE
Ziya-LlaMA-13B	43.27	40.94	25.70
ChatGLM-6B	42.91	39.69	31.66
Baichuan-13B-Chat	53.45	50.00	31.47
Qwen-14B-Chat	50.12	44.87	33.89
LexiLaw	40.36	31.56	20.11
HanFei	31.08	33.23	20.17
LawGPT-7B	29.09	30.31	22.91
Lawyer LlaMA-13B	39.64	32.50	35.75
ChatLaw-13B	41.09	35.62	27.56
Legal-LM	62.12	59.11	55.09

For our assessment, we utilized two types of datasets to ensure diversity and representativeness: an objectively judged dataset and a subjective dataset. The objective dataset comprised three sets of legal questions categorized by difficulty: LBK [32] for easy questions from the Legal Basic Knowledge Question Bank, UNGEE [32] for medium-difficulty questions from the National Unified Examination for the Specialized Master's Degree of Laws, and NJE [32] for the most challenging questions from the National Unified Legal Professional Qualification Examination. These datasets included both single-choice and multiple-choice questions to evaluate the model's handling of diverse question types, as detailed in Table 1.

3.4.2 Comparison of Advice on Legal Issues

We tested ChatGLM, Baichuan, Chatlaw, Disclaw, GPT-3.5-Turbo, and Legal-LM using a 1000-item subjective review set, and the performance of each model is depicted in Fig. 3. Legal-LM demonstrates superior performance in answering subjective questions compared to other models. In Accuracy, Legal-LM slightly surpasses other models, scoring one point higher than the baseline performance. In Completeness, Legal-LM

exceeds other baseline models by at least ten points. In Clarity, Legal-LM outperforms other baseline models by over six points. In Language Quality, Legal-LM performs slightly better than other baseline models.

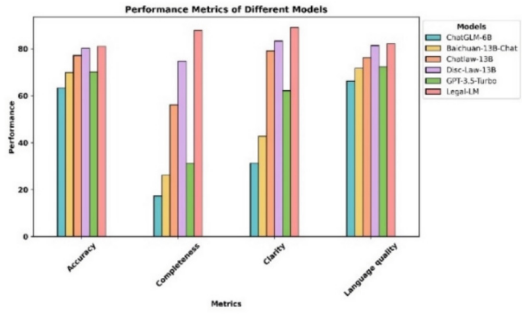

Fig. 3. Results for Legal-LM compared to ChatGLM, Baichuan, Chatlaw, Disclaw, GPT-3.5-Turbo in Counseling Subjective Legal Issues.

3.4.3 Ablation Study

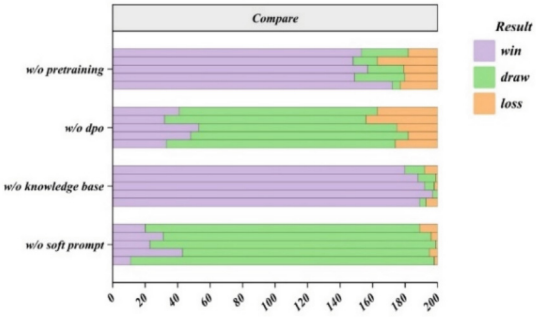

Fig. 4. The overall Legal-LM is compared with other variants of the model for results.

In this study, we examine the effectiveness of the key modules in our proposed legal macromodel from the perspective of the training approach we designed. To compare with the original approach, we constructed four model variants:

w/o Pre-training: Eliminating the pre-training process, allowing the base model to fine-tune its legal knowledge directly without prior learning.
w/o DPO: Eliminating the DPO process by not implementing the reinforcement learning process of DPO human preferences for the fine-tuned legal model.
w/o Knowledge Base: Eliminating the external knowledge base, user input questions are not retrieved from it, and the user's questions are directly used as inputs to the model.
w/o Soft Prompt: Eliminating the soft prompt, where the user's question is directly used as input to the model instead of escaping the content at the escape layer when retrieving content and laws based on the user's question.

The 1,000-item subjective review set was divided into five equal groups to evaluate the performance of the original Legal-LM against its variants, with each group being assessed by legal professionals. The variants lacking pre-training and a knowledge base showed significant negative impacts on performance, while those missing DPO and soft prompts had less noticeable effects, as documented in Fig. 4.

3.5 Analysis

Legal-LM's enhanced performance is attributed to extensive pre-training with datasets like LBK, UNGEE, and NJE, and its ability to accurately retrieve and answer related questions by incorporating these into an external knowledge base. Other models lag behind due to data contamination from mixing extensive legal data during pre-training and fine-tuning. Legal-LM surpasses in subjective legal counseling by integrating specific real-case data during both pre-training and fine-tuning phases, and refining real legal advice datasets to reduce bias, thus improving response quality significantly. Comparatively, the legal big model excels over baseline models in subjective advice due to its greater proportion of specialized legal data enhancing its question answering capabilities. The ablation study shows removing pre-training or the external knowledge base drastically affects Legal-LM's performance, introducing inaccuracies due to missing context or reliance on unnecessary legal provisions. Meanwhile, reducing the DPO process and soft prompts shows minimal impact, as these elements primarily influence preference bias and content retrieval with limited effect on the overall model performance.

4 Conclusion

In this paper, we elaborate on Legal-LM, a sophisticated large language model framework designed for the legal sector. Combining extensive legal knowledge with advanced LLM techniques, vector databases, knowledge graphs, and prompt strategies, Legal-LM overcomes the common issue of LLMs producing illogical or unrealistic outputs. Legal-LM is enhanced to effectively manage the complexity of legal data, improving its ability to interpret legal texts and arguments and provide actionable advice. This model excels in applications like legal Q&A, case analysis, and advisement. The development of Legal-LM marks a significant advancement in AI for the legal industry, broadening opportunities in the global market and accelerating the digital transformation of legal services. Its innovative approach promises substantial impacts in legal information services, education, and counseling.

Acknowledgment. This work is supported by the National Key Research and Development Program of China (2021YFC3300500); This work are also supported in part by the Natural Science Foundation of China under Grant U20B2047, U20B2053, U19B2044.

References

1. Bai, J., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

2. Bommarito, M., Katz, D.M., Detterman, E.: LexNLP: natural language processing and information extraction for legal and regulatory texts. In: Research Handbook on Big Data Law (2018)
3. Burstein, J., Doran, C., Solorio, T.: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019)
4. Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T., Yu, X.: Recall and learn: fine-tuning deep pretrained language models with less forgetting. arXiv preprint arXiv:2004.12651 (2020)
5. Chipman, H.A., George, E.I., McCulloch, R.E.: BART: Bayesian additive regression trees (2010)
6. Cui, J., Li, Z., Yan, Y., Chen, B., Yuan, L.: ChatLaw: open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092 (2023)
7. Dai, Y., et al.: LAiW: a Chinese legal large language models benchmark (a technical report). arXiv preprint arXiv:2310.05620 (2023)
8. Ding, N., et al.: Delta tuning: a comprehensive study of parameter efficient methods for pre-trained language models arXiv preprint arXiv:2203.06904 (2022)
9. Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: GLM: general language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360 (2021)
10. Duan, X., et al.: CJRC: a reliable human-annotated benchmark dataset for Chinese judicial reading comprehension. In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS, vol. 11856, pp. 439–451. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_36
11. Fu, Y., Ou, L., Chen, M., Wan, Y., Peng, H., Khot, T.: Chain-of-thought hub: a continuous effort to measure large language models' reasoning performance. arXiv preprint arXiv:2305.17306 (2023)
12. von der Lieth Gardner, A.: An artificial intelligence approach to legal reasoning (1987)
13. Houlsby, N., et al.: Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning, pp. 2790–2799 (2019)
14. Hu, E.J., et al.: LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
15. Huang, Q., et al.: Lawyer LLaMA technical report. arXiv preprint arXiv:2305.15062 (2023)
16. Ji, L., Wei, Z., Hu, X., Liu, Y., Zhang, Q., Huang, X.-J.: Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3703–3714 (2018)
17. Ji, L., Wei, Z., Li, J., Zhang, Q., Huang, X.: Discrete argument representation learning for interactive argument pair identification. arXiv preprint arXiv:1911.01621 (2019)
18. Kien, P.M., Nguyen, H.-T., Bach, N.X., Tran, V., Le Nguyen, M., Phuong, T.M.: Answering legal questions by learning neural attentive text representation. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 988–998 (2020)
19. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**, 1234–1240 (2020)
20. Li, X., et al.: AlpacaEval: an automatic evaluator of instruction-following models. GitHub repository (2023)
21. Lu, J., et al.: Ziya-VL: bilingual large vision-language model via multi-task instruction tuning. arXiv preprint arXiv:2310.08166 (2023)
22. Ma, Y., et al.: LeCaRD: a legal case retrieval dataset for Chinese law system. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2342–2348 (2021)
23. Muennighoff, N., et al.: Crosslingual generalization through multitask finetuning. arXiv preprint arXiv:2211.01786 (2022)

24. Nguyen, H.-T.: A brief report on LawGPT 1.0: a virtual legal assistant based on GPT-3. arXiv preprint arXiv:2302.05729 (2023)
25. Rabelo, J., Goebel, R., Kim, M.-Y., Kano, Y., Yoshioka, M., Satoh, K.: Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. In: Review of Socionetwork Strategies, vol. 16, pp. 111–133 (2022)
26. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: your language model is secretly a reward model. arXiv preprint arXiv:2305.18290 (2023)
27. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 5485–5551 (2020)
28. Song, Y., Wei, Z.: Inferring association between alcohol addiction and defendant's emotion based on sound at court. Front. Psychol. **12**, 669780 (2021)
29. Yang, A., et al.: Baichuan 2: open large-scale language models. arXiv preprint arXiv:2309.10305 (2023)
30. Yang, W., Jia, W., Zhou, X., Luo, Y.: Legal judgment prediction via multi-perspective bi-feedback network. arXiv preprint arXiv:1905.03969 (2019)
31. Yuan, J., et al.: Overview of SMP-CAIL2020-argmine: the interactive argument-pair extraction in judgement document challenge. Data Intell. **3**, 287–307 (2021)
32. Yue, S., et al.: DISC-LawLLM: fine-tuning large language models for intelligent legal services. arXiv preprint arXiv:2309.11325 (2023)
33. Zheng, L., et al.: Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv preprint arXiv:2306.05685 (2023)
34. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: JEC-QA: a legal-domain question answering dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9701–9708 (2020)