ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases

Jiaxi Cui*
Peking University
jiaxicui@chatlaw.cloud

Zongjian Li*
Peking University
chestnutlzj@chatlaw.cloud

Yang Yan
Peking University
yyang@stu.pku.edu.cn

Bohua Chen
Peking University
bohua@chatlaw.cloud

Li Yuan†
Peking University
yuanli-ece@pku.edu.cn

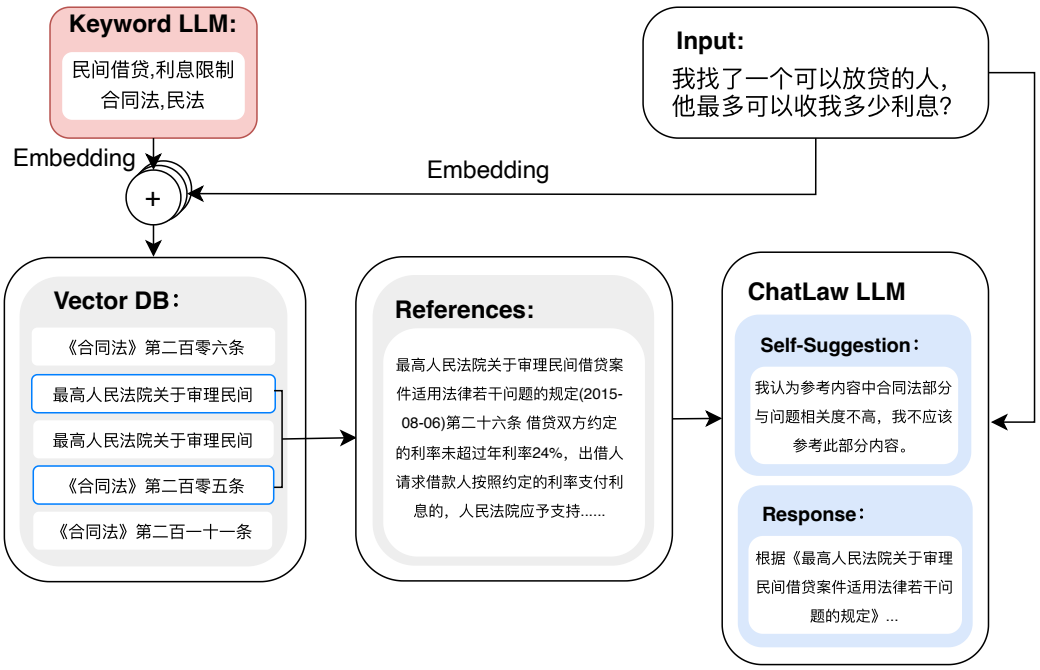

Figure 1: ChatLaw Framework

Abstract

Large Language Models (LLMs) have shown the potential to revolutionize natural language processing tasks in various domains, sparking great interest in vertical-specific large models. However, unlike proprietary models such as BloombergGPT and FinGPT, which have leveraged their unique data accumulations to make strides in the finance domain, there hasn't not many similar large language models in the Chinese legal domain to facilitate its digital transformation.

In this paper, we propose an open-source legal large language model named ChatLaw. Due to the importance of data quality, we carefully designed a legal domain

*Equal Contribution.
†Corresponding Author

Preprint. Under review.

fine-tuning dataset. Additionally, to overcome the problem of model hallucinations in legal data screening during reference data retrieval, we introduce a method that combines vector database retrieval with keyword retrieval to effectively reduce the inaccuracy of relying solely on vector database retrieval. Furthermore, we propose a self-attention method to enhance the ability of large models to overcome errors present in reference data, further optimizing the issue of model hallucinations at the model level and improving the problem-solving capabilities of large models. We also open-sourced our model and part of the data at `https://github.com/PKU-YuanGroup/ChatLaw`.

1 Introduction

The continuous expansion and development of artificial intelligence have provided a fertile ground for the proliferation of large-scale language models. Models such as ChatGPT, GPT4 [5], LLaMA [7], Falcon [1], Vicuna [2], and ChatGLM [12] have demonstrated remarkable performance in various conventional tasks, unleashing tremendous potential for the field of law. However, it is evident that acquiring high-quality, relevant, and up-to-date data is a crucial factor in the development of large language models. Therefore, the development of effective and efficient open-source legal language models has become of paramount importance.

In the realm of artificial intelligence, the development of large-scale models has permeated various domains such as healthcare, education, and finance: BloombergGPT [9], FinGPT [10], Huatuo [8], ChatMed [14], These models have demonstrated their utility and impact in tackling complex tasks and generating valuable insights. However, the field of law, with its inherent importance and demand for accuracy, stands as a domain that necessitates dedicated research and development of a specialized legal model.

Law plays a pivotal role in shaping societies, governing human interactions, and upholding justice. Legal professionals rely on accurate and up-to-date information to make informed decisions, interpret laws, and provide legal counsel. The complexities of legal language, nuanced interpretations, and the ever-evolving nature of legislation present unique challenges that require tailored solutions.

However, when it comes to legal issues, there is often a phenomenon of hallucination and nonsensical outputs, even with the most advanced model like GPT4. People tend to believe that fine-tuning a model with specific domain knowledge would yield satisfactory results. However, in reality, this is not the case with early legal LLM (LawGPT), as there are still many instances of hallucination and unreliable outputs.

We initially recognized the need for a Chinese legal LLM. However, at the time, there were no commercially available Chinese models surpassing the scale of 13 billion parameters. Therefore, we built upon the foundation of OpenLLAMA, a commercially viable model, by expanding the Chinese vocabulary and incorporating training data from sources like MOSS. This allowed us to create a foundational Chinese language model. Subsequently, we incorporated legal-specific data to train our legal model——ChatLaw.

The key contributions of this paper are as follows:

1. **Effective Approach to Mitigate Hallucination:** We propose an approach to address hallucination by enhancing the model's training process and incorporating four modules during inference: "consult," "reference", "self-suggestion" and "response." By integrating vertical models and knowledge bases through the reference module, we inject domain-specific knowledge into the model and leverage accurate information from the knowledge base, reducing the occurrence of hallucinations.
2. **Legal Feature Word Extraction Model based on LLM:** We train a model that extracts legal feature words from users' everyday language. This model identifies words with legal significance, enabling efficient identification and analysis of legal contexts within user input.
3. **Legal Text Similarity Calculation Model based on BERT:** We train a model to measure the similarity between users' everyday language and a dataset consisting of 930,000 relevant

legal case texts. This enables the creation of a vector database for efficient retrieval of similar legal texts, facilitating further analysis and reference.

4. **Construction of a Chinese Legal Exam Testing Dataset:** We curate a dataset specifically designed for testing legal domain knowledge in Chinese. Additionally, we design an ELO arena scoring mechanism to compare the performance of different models in legal multiple-choice questions.

Furthermore, we observed that a single general-purpose legal LLM may not perform optimally across all tasks in this domain. Therefore, we trained different models for various scenarios, such as multiple-choice questions, keyword extraction, and question-answering. To handle the selection and deployment of these models, we employed a big LLM as a controller using the methodology provided by HuggingGPT [6]. This controller model dynamically determines which specific model to invoke based on each user's request, ensuring the most suitable model is utilized for the given task.

2 Dataset

In constructing the dataset, we employed several approaches to ensure its comprehensiveness and diversity. The dataset composition methods are as follows:

Collection of a vast amount of original legal data: This includes gathering legal news, social media content, and discussions from legal industry forums. These sources provide a diverse range of real-world legal text, offering insights into various legal topics and discussions.

> **Question**: 在一名大学生的日记里写满"一定要干掉乙，可以采用的方法很多，最好是投毒"的话语，但是未见行动，某日辅导员看到日记后报案，请问这种行为是否属于故意杀人罪?
> **Answer**: 根据您提供的情况，一名大学生在日记中写满"一定要干掉乙，可以采用的方法很多，最好是投毒"的话语，但是一直未见行动，某日辅导员看到甲的日记，很担心便报案。根据我国《刑法》规定，这种行为属于犯意表示，无罪。

Construction based on legal regulations and judicial interpretations: To ensure comprehensive coverage of legal knowledge, we incorporate relevant legal regulations and judicial interpretations into the dataset. This ensures that the dataset reflects the legal framework and provides accurate and up-to-date information.

> **Question**: 如果某个工人每天工作10小时，那么每周工作时间超过了规定吗?
> **Answer**: 根据《劳动法》第三十六条，每日工作时间不得超过8小时，因此每周工作时间将超过44小时，超过了规定。
> **Question**: 如果一个外国人在我国战时向国外武装力量提供武器装备，是否构成资敌罪?
> **Answer**: 不构成资敌罪。根据《刑法》第一百一十二条规定，资敌罪主体只能是我国公民，外国人、无国籍人不能成为资敌罪主体。因此，外国人在我国战时向国外武装力量提供武器装备，不构成资敌罪。

Crawling real legal consultation data: We retrieve authentic legal consultation data, leveraging existing legal consultation datasets. This enables the inclusion of real-world legal scenarios and questions that users commonly encounter, enriching the dataset with practical legal examples.

> **Question**: 我想问一下，当事人如果无力支付律师费用，律师应该减收或免收费用的情况有哪些? 比如，因为被检察院错误逮捕而请求赔偿的案件，在作业期间发生伤残并请求单位支付医疗费和生活费的案件，以及因交通肇事罪而聘请律师作辩护的案件?
> **Answer**: 是的，如果当事人无力支付律师费用，律师应该减收或免收费用的情况包括：因为被检察院错误逮捕而请求赔偿的案件、在作业期间发生伤残并请求单位支付医疗费和生活费的案件，以及因交通肇事罪而聘请律师作辩护的案件。这些案件都属于特殊情况，律师应该及时帮助当事人解决相关问题，支持当事人实现正义，因此，律师应根据当事人的实际情况减收或免收费用。

Construction of multiple-choice questions for the bar exam: We create a set of multiple-choice questions specifically designed for the bar exam. These questions cover various legal topics and test users' understanding and application of legal principles.

> **Question**: 红星中学采用伪劣产品铺设足球场，致使刺激性气味四处散发，并严重污染了场地底下土壤。于是，甲环保协会向市中级人民法院提起诉讼，请求判令红星中学拆除新建的足球场，并对污染的土壤采取修复措施。法院在受理后第7日书面告知市环保局。此时，市人民检察院也就此向法院提起公益诉讼，法院将其列为共同原告。双方当事人经协商达成的和解协议，法院未予审查即发出公告。公告期满后，应双方当事人请求，法院未制作调解书。关于本案，市中级人民法院的下列哪些做法是不合法的？A. 受理后第7日书面告知市环保局B. 对和解协议未经审查即发出公告C. 将市人民检察院列为共同原告D. 应双方当事人请求未制作调解书
> **Answer**: 《民诉解释》第286条规定，人民法院受理公益诉讼案件后，应当在10日内书面告知相关行政主管部门。市中院受理后第7日告知，符合法律规定。A选项合法，不当选。《民诉解释》第289条第2款规定，当事人达成和解或者调解协议后，人民法院应当将和解或者调解协议进行公告。公告期间不得少于30日。据此，法院公告前不审查，并不违反法律规定。B选项合法，不当选。《民事诉讼法》第55条第2款规定，人民检察院在履行职责中发现破坏生态环境和资源保护、食品药品安全领域侵害众多消费者合法权益等损害社会公共利益的行为，在没有前款规定的机关和组织或者前款规定的机关和组织不提起诉讼的情况下，可以向人民法院提起诉讼；前款规定的机关或者组织提起诉讼的，人民检察院可以支持起诉。据此，本案已有环保协会提起公益诉讼，检察机关只能作为支持起诉人参与公益诉讼，而不能成为共同原告。C选项不合法，当选。《民诉解释》第289条第3款规定，公告期满后，人民法院经审查，和解或者调解协议不违反社会公共利益的，应当出具调解书。据此，公益诉讼案件法院必须制作调解书。D选项不合法，当选。

By incorporating data from these diverse sources and construction methods, our dataset encompasses a wide range of legal contexts, ensuring that the developed model is capable of effectively understanding and addressing various legal scenarios.

Once these data components are collected, the dataset undergoes a rigorous cleaning process. This involves filtering out short and incoherent responses, ensuring that only high-quality and meaningful text is included. Additionally, to enhance the dataset, we leverage the ChatGPT API for assisted construction, allowing us to generate supplementary data based on the existing dataset.

3 Training Process

The Keyword LLM is a language model that extracts keywords from abstract consulting problems raised by users. The Law LLM, on the other hand, extracts legal terminology that may be involved in user consultations. The ChatLaw LLM is the ultimate language model that outputs responses to users. It refers to relevant legal clauses and utilizes its own summarization and Q&A function to generate advice for users in their consultations.

3.1 ChatLaw LLM

To train ChatLAW, we fine-tuned it on the basis of Ziya-LLaMA-13B [11] using Low-Rank Adaptation (LoRA) [3]. Additionally, we introduced the self-suggestion role to further alleviate model hallucination issues. The training process was carried out on multiple A100 GPUs and the training costs were further reduced with the help of deepspeed.

3.2 Keyword LLM

Creating ChatLaw product by combining vertical-specific LLM with a knowledge base, it is crucial to retrieve relevant information from the knowledge base based on user queries. We initially tried traditional software development methods such as MySQL and Elasticsearch for retrieval, but the results were unsatisfactory. Therefore, we attempted to use a pre-trained BERT model for embedding,

Algorithm 1 Legal retrieval based on Large Langu Model keyword extraction

Initialize the BERT model for embedding and keyword extraction model.
Initialize the legal database as $\mathcal{L}$, where $\mathbf{l}_i \in \mathcal{L}$ and i represents the i-th law. Let M be the number of laws in the legal database.
Initialize the legal scores as $\mathcal{S}$, where $s_i \in \mathcal{S}$ represents the score corresponding to the i-th law, all initialized to 0. The number of elements in $\mathcal{S}$ is also M.
Extracting keywords from user queries using a keyword extraction model, and then inputting each keyword into a BERT model to obtain a collection of $\mathcal{K}$ keyword vectors, where $\mathbf{k}_i$ represents the vector for the ith keyword, with a total of N keywords. We obtain $\mathbf{s}$ by inputting the user's question into BERT.
Initialize α for assigning weight to $\mathbf{s}$.
for i *to* N **do**
    $\mathbf{v}_i = \frac{\mathbf{k}_i}{||\mathbf{k}_i||} + \alpha \frac{\mathbf{s}}{||\mathbf{s}||}$
    for j *to* M **do**
        $s_j \leftarrow s_j + cossim(\mathbf{v}_i, \mathbf{l}_j)$
    end for
end for
return $TopK(\mathcal{S})$

followed by methods such as Faiss [4] to calculate cosine similarity and extract the top k legal regulations related to user queries. However, this method often yields suboptimal results when the user's question is vague. Therefore, we aim to extract key information from user queries and use the vector embedding of this information to design an algorithm to improve matching accuracy.

Due to the significant advantages of large models in understanding user queries, we fine-tuned an LLM to extract the keywords from user queries. After obtaining multiple keywords, we adopted **Algorithm 1** to retrieve relevant legal provisions.

input	keywords	laws
公司无缘无故突然要把我辞退，我应该怎么办	公司、劳动纠纷、合同、辞退、赔偿、	《劳动合同法》(2012-12-28)第四十一条 有下列情形之一，需要裁减人员二十人以上或者裁减不足二十人但占企业职工总数... 《劳动合同法》(2012-12-28)第三十七条 劳动者提前三十日以书面形式通知用人单位，可以解除...
有人未经我同意用我的照片在网上传播非法言论，我该做什么	民事相关、名誉权侵犯隐私、网络、照片、非法言论	《侵权责任法》(2009-12-26)第三十六条 网络用户、网络服务提供者利用网络侵害他人民... 《最高人民法院关于审理利用信息网络侵害人身权益民事纠纷案件适用法律若干问题的规定》(2014-08-21)第十八条 被侵权人为制止侵权行为所支付的合理开支，可以认定为侵权责任法第二十条规定的财产损失。合理开支包括被侵权人... 《民法典》(2020-05-28)第一千零二十四条：民事主体享有名誉权，任何组织或者个人不得以侮辱、诽谤等方...
网上买到了假货，商家说是我自己用假的掉包了不给退货，我该怎么办	假货、退货、消费者维权、电子商务、民事诉讼	《最高人民法院关于审理网络消费纠纷案件适用法律若干问题的规定》(2022-03-01)第七条 消费者在二手商品网络交易平台购买商品受到损害，人民法院综合销售者出售商品的性质、来源... 《直销管理条例》(2017-03-01)第二十五条 直销企业应当建立并实行完善的换货和退货制度。\n\n消费者自购买直销产品之日起30日...

Figure 2: Result of Keyword LLM and Law LLM

3.3 Law LLM

We trained a BERT model using a dataset of 937k national case law examples to extract corresponding legal provisions and judicial interpretations from user queries. This Law LLM model forms an essential component of the ChatLaw product.

4 Experiment and Analysis

Evaluating the performance of the Large Language Model (LLM) has always been a challenge. For this purpose, we have collected national judicial examination questions over a decade and compiled a test dataset containing 2000 questions with their standard answers to measure the models' ability to handle legal multiple-choice questions.

However, we found that the accuracy rates of the models are generally quite low. Under these circumstances, simply comparing accuracy rates seems to hold little significance. Therefore, we have established an evaluation mechanism for model competition for Elo points, inspired by the matchmaking mechanism in e-sports and the design of Chatbot Arena [13], to more effectively assess the models' abilities to handle legal multiple-choice questions.

Model	Score
ChatLaw(13B)	1733.85
gpt-4	1712.03
lawyer-llama(13B)	1597.18
gpt-3.5-turbo	1573.35
OpenLLaMA(13B)	1475.55
LawGPT(7B)	1452.35

Figure 3: ELO Ranking up until June 25

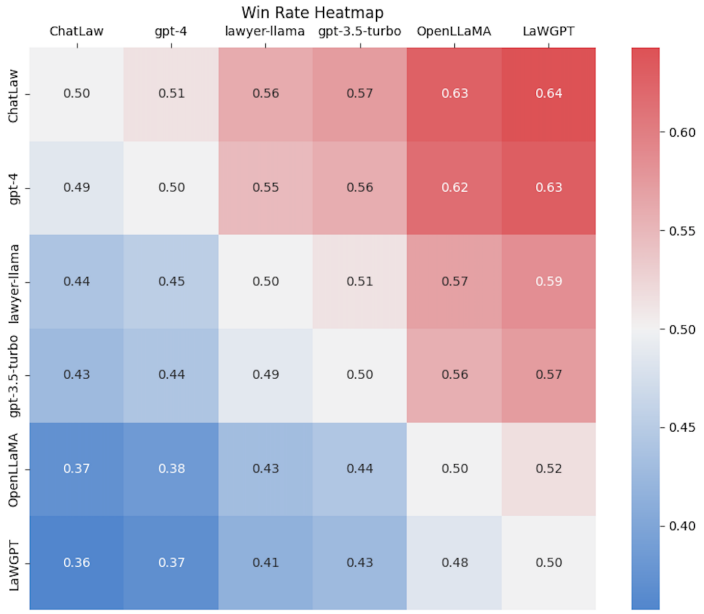

Figure 4: LLM Win Rate

Through the analysis of the above experimental results, we can make the following observations:

(1) The introduction of legal-related Q&A and statute data can to some extent improve the model's performance on multiple-choice questions;

(2) The addition of specific task types for training significantly improves the model's performance on such tasks. For example, the reason why the ChatLaw model outperforms GPT-4 is that we used a large number of multiple-choice questions as training data;

(3) Legal multiple-choice questions require complex logical reasoning, so models with a larger number of parameters usually perform better.

5 Conclusions

In this paper, we proposed ChatLaw, a legal large language model(LLM) developed using legal domain knowledge. We propose a novel approach that combines LLM with vector knowledge databases, which significantly alleviates the hallucination problem commonly seen in LLM. Our stable model handling strategies enable the resolution of various legal domain problems. Additionally, we release a dataset for legal multiple-choice questions and design an ELO model ranking mechanism.

However, our limitations arise due to the scale of the base model. Our performance in tasks such as logical reasoning and deduction is not optimal. Additionally, after incorporating a large amount of domain-specific data, further research is required to improve the generalization of ChatLaw for generic tasks. There are potential social risks on ChatLaw, and we advise users to make use of our method for proper purposes.

References

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[4] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[5] OpenAI. Gpt-4 technical report, 2023.

[6] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, 2023.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[8] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge, 2023.

[9] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.

[10] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models, 2023.

[11] Ping Yang, Junjie Wang, Ruyi Gan, Xinyu Zhu, Lin Zhang, Ziwei Wu, Xinyu Gao, Jiaxing Zhang, and Tetsuya Sakai. Zero-shot learners for natural language understanding via a unified multiple choice perspective, 2022.

[12] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[13] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[14] Wei Zhu and Xiaoling Wang. Chatmed: A chinese medical large language model. `https://github.com/michael-wzhu/ChatMed`, 2023.