Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain

Anonymous ACL submission

Abstract

While large language models (LLMs) like GPT-4 have recently demonstrated astonishing zero-shot capabilities in general domain tasks, they often generate content with hallucinations in specific domains such as Chinese law, hindering their application in these areas. This is typically due to the absence of training data that encompasses such a specific domain, preventing GPT-4 from acquiring in-domain knowledge. A pressing challenge is that it's not plausible to continue training LLMs of the GPT-4's scale on in-domain data.

This paper introduces a simple yet effective domain adaptation framework for GPT-4 by reformulating generation as an **adapt-retrieve-revise** process. The initial step is to **adapt** an affordable 7B LLM to the Chinese legal domain by continuing learning in-domain data. When solving an in-domain task, we leverage the adapted LLM to generate a draft answer given a task query. Then, the draft answer will be used to **retrieve** supporting evidence candidates from an external in-domain knowledge base. Finally, the draft answer and retrieved evidence are concatenated into a whole prompt to let GPT-4 assess the evidence and **revise** the draft answer to generate the final answer.

Our proposal combines the advantages of the efficiency of adapting a smaller 7B model with the evidence-assessing capability of GPT-4 and effectively prevents GPT-4 from generating hallucinatory content. In the zero-shot setting of four Chinese legal tasks, our method improves the average score by $+33.6$ points, compared to GPT-4 direct generation. When compared to two stronger retrieval-based baselines, our method outperforms them by $+17.0$ and $+23.5$. Our code will be released [1].

1 Introduction

Recent large language models like GPT-4 bring remarkable improvements in various general domain NLP tasks (Brown et al., 2020a; OpenAI, 2023a; Thoppilan et al., 2022; Chowdhery et al., 2022; Rae et al., 2022; Hoffmann et al., 2022). However, in specific domains such as the Chinese legal domain, the performance of such general LLMs still lags considerably behind. We show a real example of Chinese LegalQA (Chen et al., 2023) on the left of Figure 1, which requires the model to generate the corresponding legal provision (i.e., the law name and the clause index) and the rationale for the judgment, given a brief case description as the query.

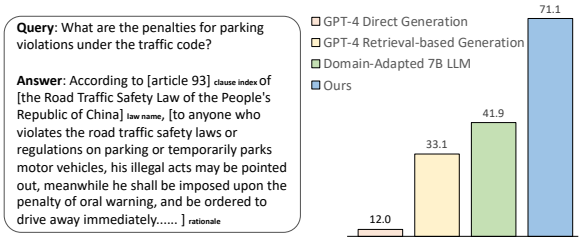

Figure 1: **Left**: A real translated example of Chinese LegalQA. The square brackets and subscripts are offered for the purpose of clear demonstration, not actually exist in the ground-truth answer or generation. **Right**: Models' F1 scores on the LegalQA dataset.

We initialize the research with a preliminary examination of utilizing GPT-4 to address the Chinese LegalQA task, which involves responding with a law clause relevant to a given query case. Figure 1 reveals the extremely low performance (F1 12.0) of directly prompting the query case to ask GPT-4 to generate the corresponding law clause. Though the generated answers are grammatically fluent, they often consist of non-logical content, factual mistakes, and fail to refer to the correct legal provision (also known as "hallucination"). For example, in the left answer of Figure 2, the direct generation of GPT-4 seems to retell the case description but fails to point out the corresponding clause. A

[1] The code and 7B legal model will be released

potential reason is the insufficient Chinese legal domain text during pretraining, leading to a lack of domain knowledge acquisition, and consequently generating hallucinatory content.

For the LLMs with the scale of GPT-4, it's generally not feasible for researchers to do learning-based adaptation. The enormous model size could make the cost of continual learning extremely high, and meanwhile, the access functions are often limited by APIs. Therefore, recent work (Lewis et al., 2020b; Yu et al., 2023; Shuster et al., 2021; Ma et al., 2023) introduces retrieval-based methods that first use the given query to retrieve relevant evidence candidates from the external domain-specific knowledge base or the internet and then concatenate the query and the evidence candidates into the prompt. GPT-4 could implicitly validate the relevance between the query and the evidence, as well as the correctness of the evidence, before producing a generation.

Our replicated retrieval-based method improves the LegalQA F1 from the 12.0 points of direct generation to 33.1 as shown in Figure 1. It indicates that even though GPT-4 may not generate domain content, it possesses sufficient evidence-assessing capacity to select the correct evidence from candidates. Nevertheless, the retrieval module is limited by the capability of representation mapping from query to evidence and is also influenced by the domain issue, leading to a decline in search quality. GPT-4 still produces hallucinations in responses as the middle answer in Figure 2.

On the other hand, with the rapid development of open LLMs led by LLaMA (Touvron et al., 2023), it becomes affordable to continually train an open LLM tailored to your demands on sufficient in-domain texts, resulting in a domain-adapted LLM. We therefore conduct the second examing of continually training Baichuan 7B (Baichuan-inc, 2023), a Chinese foundation model, on over 50B token Chinese legal data. Its performance (the green bar in Figure 1) even surpasses the retrieval-based GPT-4 generation on Chinese LegalQA. Hallucinations caused by the lack of domain knowledge are largely reduced but not completely solved. As shown in the right answer in Figure 2, adapted LLM generates generally correct responses but still makes errors in certain words. Although the law name is correct and the rationale part is reasonable, the clause index is a hallucinatory generation which raises the difficulty in anchoring target clauses. We argue that these fails are accordingly caused by the limited capability of a 7B size to memorize the knowledge accurately.

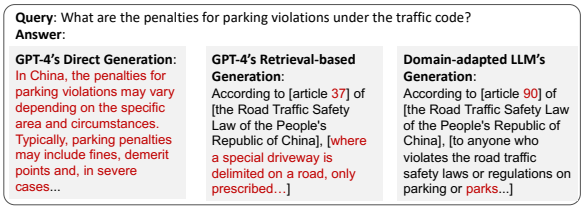

Figure 2: Examples of hallucinations of various models. Red denotes the content containing hallucinations. The ground-truth answer refers to the left case in Figure 1.

Building upon the observation of the evidence-assessing capability of GPT-4 and the high-quality domain content generated by the domain-adapted 7B model, this paper proposes a novel approach to reformulate GPT-4's domain content generation to an **adapt-retrieve-revise** process: (1) the **domain-adapted** model generates a draft answer given a query; (2) the **retrieval** module uses the draft answer as input for searching external evidence candidates because the answer is usually more informative and semantically similar to the evidence compared to the query as long as the answer quality is acceptable; (3) GPT-4 assesses retrieved evidence and **revises** the draft answer to generate the final answer.

The rest sections of the paper anchor the Chinese legal domain and comprehensively validate the effectiveness of our proposal. In Section 2, we explain each component of our adapt-retrieve-revise method and elaborate on the implementation details. In Section 3 and 5, we conduct the experiments and the result analysis on four Chinese legal domain tasks. The experimental results show substantial improvements against the direct generation and the retrieval-based generation baselines. In the final Section 7, we elicit the conclusion and future work. To the best of our knowledge, this is the first study to examine the zero-shot performance of LLMs on four Chinese legal benchmarks.

2 Methodology

Our adapt-retrieve-revise method consists of three steps. In the first step (Section 2.1), we continually train a Chinese pre-trained LLM on the Chinese legal domain corpora to derive a domain-adapted legal LLM and given the query, the legal LLM will generate the draft answer. In the second step (Section 2.2), we use a sentence embedding model to produce embeddings for both the draft answer

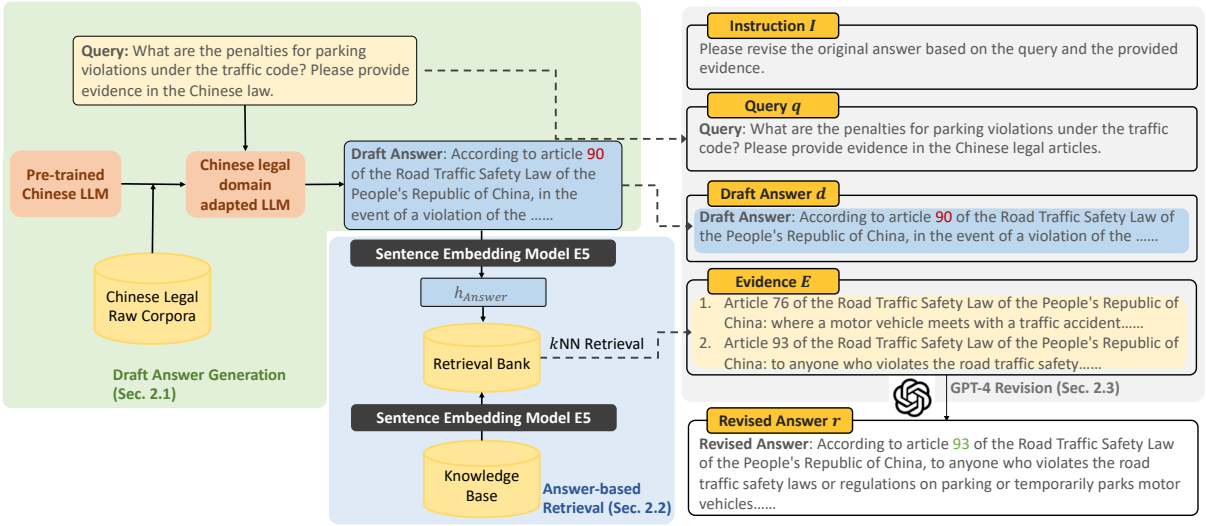

Figure 3: **Overview of our proposed method.** The example and prompt are translated from Chinese to English for the demonstration purpose.

and each paragraph in the corresponding knowledge base, then evidence retrieval will be computed by the similarities between the answer embedding and the paragraph embeddings. In the third step (Section 2.3), we concatenate the query, the draft answer, and the retrieved evidence in the prompt for GPT-4 to revise and produce the final response. Figure 3 shows the overview of our method. In the following sections, we will introduce details of each step.

2.1 Draft Answer Generation by the Domain-adapted LLM

This step could be actually flexible to save the effort of training a LLM yourself. For instance, an off-the-shelf Chinese legal domain LLM can be used for generating draft answer. However, most of such LLMs have been fine-tuned on various in-domain evaluation tasks. The potential data leakage could lead to unreliable evaluation, especially in a zero-shot setting. Therefore, in this paper, we adapt an open 7B LLM to the Chinese legal domain by continual learning on the domain data. We collect the training data from the following two open Chinese legal sources:

- **Chinese Law Clauses** (https://flk.npc.gov.cn/) form the foundation of the judicial system, containing a wealth of legal terms, provisions, and judicial practices. They are essential for the model to understand and generate relevant content.
- **Chinese Judgments Online** (https://wenshu.court.gov.cn/) is the largest online publication platform for legal documents in China. The platform contains judicial documents from courts at all levels, covering various legal fields such as civil, criminal, administrative, and enforcement. Such documents contain knowledge for LLMs to understand the usage of laws in various scenarios.

During the inference, given an input query, we will first prompt the trained 7B legal LLM to generate the draft answer, which will be used in the next step. For the prompt, we add the instruction "Please provide evidence in the Chinese law" at the end of the query to enforce the model to generate related law clauses, as in Figure 3.

2.2 Answer-based Evidence Retrieval

Since the draft answer of the 7B legal LLM is usually more informative and semantically similar to the evidence than the query. We further use the generated evidence to retrieve ground-truth evidence from the target knowledge base for the purpose of revision since it contains much more information than the query, even though the hallucinations can not be totally reduced. We implement this method with two subsequent steps: knowledge bank construction and retrieval.

Knowledge Bank Construction For the i-th paragraph p_i, we construct the *key-value* pair $(\mathbf{p_i}, p_i)$ where the *key* $\mathbf{p_i}$ is the representation obtained from the sentence embedding model and the *value* p_i denotes the paragraph. The memory

$(\mathcal{K}, \mathcal{V}) = \{(\mathbf{p_i}, p_i) | p_i \in \mathcal{KB}\}$ is thus the set of all *key-value* pairs constructed from all the paragraphs in the external knowledge base $\mathcal{KB}$.

Retrieval Given the generated draft answer d, the sentence embedding model outputs its representation $\mathbf{h_{Answer}}$. We then query the constructed knowledge bank with $\mathbf{h_{Answer}}$ to retrieve its k nearest neighbors E according to a distance function by L^2 distance.

2.3 GPT-4 Revision

To effectively combine the high-quality draft answers generated by the 7B domain adapted model with GPT-4's powerful evidence-assessing capability, we propose the following process. As shown in Figure 3, the whole prompt consists of the following components: (1) the instruction I to require GPT-4 to revise the draft answer given the query and the evidence candidates; (2) the query q itself; (3) the draft answer d for GPT-4 to revise; (4) and the retrieved evidence candidates E to provide related Chinese legal knowledge for GPT-4. Then, the final revised answer r will be outputted by $GPT4(I, q, d, E)$.

3 Experiment Settings

We conducted a series of experiments to compare our adapt-retrieve-revise method to the baselines of direct generation and retrieval-based generation on various Chinese legal benchmarks. We show the model details and the task settings in this section.

3.1 Model Settings

Details of training 7B legal LLM: We utilize the general domain Baichuan 7B model[2] for continual learning Chinese legal corpora. In total, we trained 50B tokens of Chinese Law Clauses and Chinese Judgments Online corpora with the input length limit of 16K and the batch size of 256 on 32 A100 GPUs, and the time-consuming is 167 hours. After continual learning, we subsequently supervised fine-tuning our model on $70K$ instruction examples, including $52K$ GPT-4 self-instruct Chinese data (Peng et al., 2023) and 18K legal instructions (See Appendix A.1) for the alignment.

Retriever setting: We utilize Multilingual-E5-large (Wang et al., 2022), a Roberta-based (Liu et al., 2019) sentence embedding model that achieves robust performance on various tasks. We also compare with other retrieval modules in Appendix A.2.

GPT-4 setting: For the utilization of GPT-4, we select "gpt-4-0613" with maximal $8K$ input tokens and use the original Chinese prompt as shown in Sec 2.3 and Figure 3 via OpenAI API.

3.2 Evaluation of Chinese Legal Tasks

We evaluated our Adapt-Retrieve-Revise method on a diversity of tasks with different knowledge base for retrieval in the zero-shot setting:

- **Law Clause Recommendation (LCR) and Criminal Prediction (CP)** (Xiao et al., 2018) are two tasks using the legal report as the input, and let the model generate the most related law clause and predict the criminal type based on the law clause. For these two tasks, we use the **Chinese law clauses** as the domain knowledge base for retrieval.

- **LegalQA** is a filtered set of EUQALS (Chen et al., 2023) that, given an input query, the model should generate an answer based on the most related legal clause. The filtering is based on the quality of the questions and we will release the filtered set. We also use the **Chinese law clauses** as the domain knowledge base for retrieval.

- **JEC-QA** (Zhong et al., 2020) is the official test for getting a lawyer's certificate in China. We chose the single-choice selection questions in our evaluations with the **Legal Textbooks** (`https://github.com/thunlp/jec-qa`) as the knowledge base for retrieval.

- **Similar Case Retrieval** (Ma et al., 2021) is the task, given a query legal scenario as the input, we aim at selecting similar **Legal Judgement Documents** from the 100 candidates. We conducted this experiment to assess the reliability of our proposed retrieval method in Section 5.1.2.

Due to the cost of GPT-4 API and the human evaluation, we randomly sampled a subset of 250 test examples for each task of LCR, CP, LegalQA, and JEC-QA. See Appendix A.3 for the statistics of the original datasets.

[2] `https://huggingface.co/baichuan-inc/Baichuan-7B`

Generator	Retriever	Revisor	Chinese law clauses			Textbooks	Avg.
			LCR F1 (Rec.)	CP F1 (Rec.)	LegalQA F1 (Rec.)	JEC-QA Acc.	
Direct Generation							
GPT-4	-	-	61.7 (67.6)	70.7 (71.2)	12.0 (14.4)	36.2	45.1
7B legal LLM	-	-	83.0 (88.4)	82.9 (84.0)	41.9 (48.8)	39.8	61.9
Retrieval-based Generation							
GPT-4	Query-based	-	72.0 (74.4)	74.0 (75.2)	33.1 (36.0)	41.6	55.2
7B legal LLM	Query-based	-	77.3 (87.6)	81.3 (82.4)	47.4 (50.2)	40.8	61.7
Adapt-Retrieve-Revise							
7B legal LLM	Answer-based	7B legal LLM	84.1 (88.4)	82.7 (83.6)	47.3 (49.0)	40.2	63.6
7B legal LLM (ours)	Answer-based	GPT-4	**90.6 (96.4)**	**86.9 (87.8)**	**71.1 (72.4)**	**66.2**	**78.7**

Table 1: **Main Results on four Chinese legal datasets**. "Rec." denotes recall, "Acc." denotes accuracy, and "Avg." is computed by the F1 and accuracy scores of all four tasks

3.3 Evaluation Metrics

Since generative models produce diverse formats in the output and the Chinese legal domain has its own features, checking the evaluation metrics in the experiments is crucial. For tasks LCR, CP, and LegalQA, our metric is the *Micro F1* and the *Recall* of whether the title of the ground-truth law clause is included in the generated answer. This is because, in real-world applications, with the correct title, the contents of the law clause can be easily revised by the rule-based system, indicating that the title is more important than the content.

For the JEC-QA task, we use accuracy as the metric, but controlling the output into an identical format for automatic evaluation is difficult, especially for the 7B LLM that has not been fine-tuned on the JEC-QA task. We select human evaluation to ensure the accuracy of our evaluation.

For the Similar Case Retrieval task, we chose the widely used *precision@k* and *MAP* as the evaluation metrics.

4 Main Experimental Results

We provide the main results as in Table 1. As we claimed before, this is the first work to targeting the zero-shot LLM performance on Chinese legal domain tasks. Generally, we compare our adapt-retrieve-revise proposal with baselines of direct generations and retrieval-based generations using the query, showing that our method outperforms all baselines by substantial margins. Our main results also provide some ablation results.

We first observe the effectiveness of domain adaption. Our 7B legal LLM significantly beats GPT-4, and even compared with the retrieval-based generation of GPT-4, the 7B legal LLM still outperforms on three tasks and has competitive results on the JEC-QA task, indicating that our continual learning on Chinese legal raw corpora shows a fast and effective domain adaptation on various legal tasks.

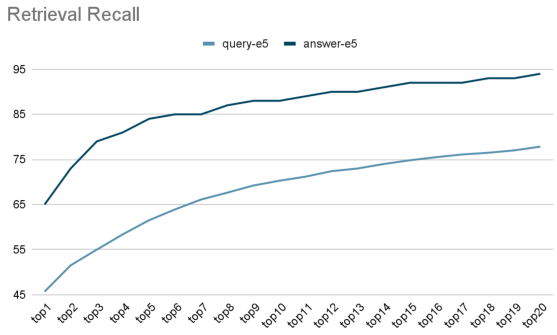

Figure 4: **Comparison of retrieval recalls on the LegalQA dataset.**

Then, considering the results of GPT-4 and the GPT-4 retrieval-based generation, we find that after providing evidence of related legal knowledge, GPT-4 can improve its responses significantly (+10.1 points). This indicates that the retrieval-based method is a proper way to reduce hallucinations caused by the lack of domain knowledge, and owing to the robust evidence-assessing capacity, GPT-4 can adapt to the Chinese legal domain well with convincing evidence available.

In our final experiment, using the draft answers generated by the 7B legal model for retrieval and revision, the performance significantly exceeded two query-based retrieval baselines by large margins of +17.0 and +23.5 points. It's worth noting that the improvement here comes from both the enhanced answer-based retrieval quality and the revision setup. In the subsequent ablation study,

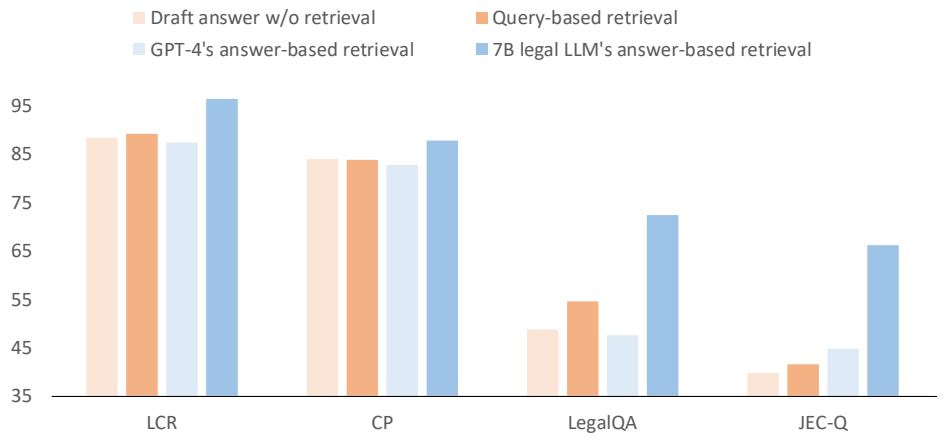

Figure 5: We compare performances of the draft answer of 7B legal LLM and our proposed adapt-retrieve-revise model using different contents in retrieval.

we will further examine the quantified improvement of the retrieval quality through an additional retrieving task.

An interesting observation is that, by comparing the direct generation of the 7B legal model and the adapt-retrieve-revise method with the revision model as the legal 7B model, we find that with retrieved evidence, the revised answers seem to be no obvious difference from the direct generation. This indicates that the 7B legal LLM shows almost zero evidence-assessing capacity.

5 Further Analysis

5.1 Analysis of Retrieval Methods

The previous section has demonstrated the stepwise effectiveness of our adapt-retrieve-revise proposal. Nevertheless, the retriever, as a key component that directly affects the quality of evidence, how its variations would impact the final performance is a crucial research question to investigate.

5.1.1 Retrieving A Query or Retrieving an Answer?

We believe the answer-based approach is more effective due to two reasons. (1) The query-based retrieval requires a query-to-evidence representation mapping. The answers are usually more semantically similar to the evidence, which avoids the mapping process. (2) A query is often very brief, while an answer containing the legal provision and rationale is more informative. In this sub-section, we analyze the retrieval component, including the apple-to-apple comparisons between the query-based and answer-based performance and the influence of answer quality for the answer-based retrieval.

Setup	Precision@5	Precision@10	MAP
Query-based	42.1	42.0	47.8
Answer-based	**45.2** (+3.1)	**42.1** (+0.1)	**49.5** (+1.7)

Table 2: **Results of two retrieval setups on the Similar Case Retrieval dataset**.

We ordered the top-similar law clauses in each retrieval and evaluated the recall in top-k, indicating whether the ground-truth law clause appears in the top-k retrieved law clauses. As shown in Figure 4, the top-1 retrieved law clause based on the answer competes with the top-5 law clauses based on the query, and the answer-based retrieval beats the query-based retrieval with a large margin for all k. This confirms our first reason that the draft answer contains much more information than the query for retrieval, indicating that LLMs can be intrinsic retrievers.

We further compare the query-based and answer-based retrieval on a public **Similar Case Retrieval** task. This task aims to select similar legal judgments based on the query from the candidates with a query the case brief given. As shown in Table 2, we compare two setups: 1) using the original query to retrieve, 2) using the legal 7B LLM to complete a whole legal judgment document given the brief query, and then retrieving. We follow the original task repository for the other settings[3]. The results show that on each metric, the answer-based retrieval works better, indicating that using the generated answer provides a more robust retrieval.

[3] https://github.com/myx666/LeCaRD

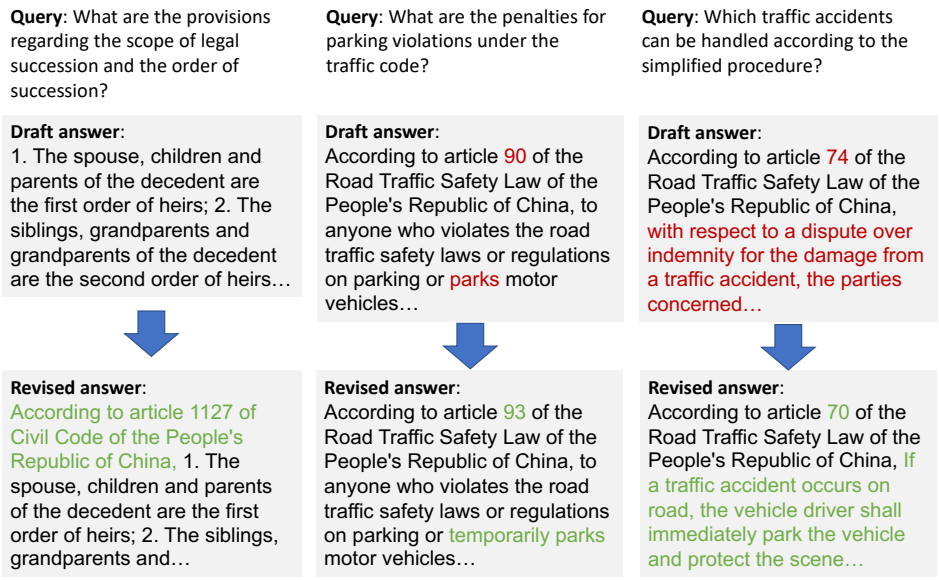

Figure 6: **Case studies on the GPT-4 revision**. The examples are translated from Chinese to English for the demonstration purpose.

5.1.2 Does the Quality of Answer Matter for Answer-Based Retrieval?

It's an intuitive thought that the quality of answers will significantly impact the outcome of answer-based retrieval. Therefore, we compare the retrieval using the answers of GPT-4 and the 7B legal LLM. We change the contents in retrieval for our proposed adapt-retrieve-revise method. As shown in Figure 5, by comparing query-based and GPT-4's answer-based retrievals, we find that the answer-based retrieval fails on three datasets (LCR, CP, LegalQA), indicating that the lack of domain knowledge in the GPT-4 responses leads to a more noisy retrieval, which even hurts the performance of the draft answer (LCR, CP, LegalQA). Meanwhile, after domain adapting, our 7B legal LLM provides robust answers in retrieval and leads to the best performances, indicating that the learned Chinese legal domain knowledge improves our answer-based retrieval.

5.2 Case Analysis of the Improvements after the GPT-4 Revision

We conclude the improvements by GPT-4 in three aspects as shown in Figure 6:

- **Adding law clauses for reference**: Sometimes, the 7B legal LLM only provides a fluent response without following the input instructions to provide the key information of the referred law name and clause index. In this case, the faithfulness of the answer remains unchecked for the users. However, after the revision, each answer is equipped with the referred law clause, which makes it easier to check the accuracy of the responses.
- **Revising hallucinations in the evidence**: even the domain-adapted LLM can provide evidence from its learned legal knowledge; the hallucination remains to some degree, such as the wrong clause index, even the law name and rationale are roughly correct. Since the rationale content is accurate, the answer-based retrieval will search for the correct evidence, and the revision by GPT-4 will solve the hallucination to produce a more robust response.
- **Choosing correct evidence**: In a more significant scenario, even though the 7B legal model's answers might contain partial hallucinatory content, the retrieval component can still possibly identify correct evidence through the partially correct descriptions in the rationale generation. During the revision stage, GPT-4 could assess the correct evidence, leading to the generation of correct answers.

5.3 Does the iteration make the generation better?

Since our method provides more accurate responses than the original response from the domain LLM, one question is whether this procedure can be iterated to improve the responses. We can use the

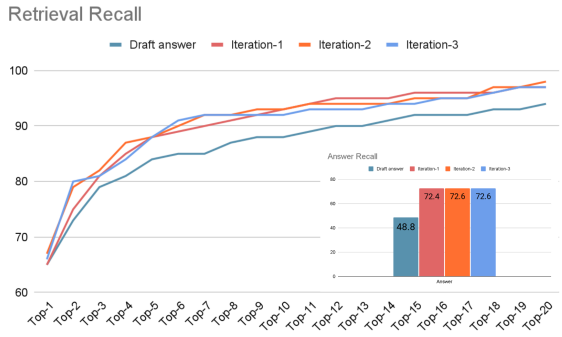

Figure 7: **Comparison of iterations on LegalQA dataset.**

revised response to retrieve related evidence and further improve the response. To verify this probability, we iteratively test on the LegalQA dataset. As the result is shown in Figure 7, during the iteration, the retrieval recall does not show consistent improvements compared with the first revision, leading to the performance nearly unchanged.

6 Related Work

6.1 Chinese Legal Domain Tasks

The rapid advancements in LLMs have significantly impacted various domains, including the legal industry. This gives rise to the occurrence of legal datasets, such as the Challenge of AI in Law (CAILf)[4], LeCaRD (Ma et al., 2021), JEC-QA (Zhong et al., 2020) and EQUALS (Chen et al., 2023). These datasets cover document classification, summarization, question answering, information extraction, similar document retrieval, and other popular NLP tasks in the Chinese legal domain. To the best of our knowledge, this paper is the first work to exam the zero-shot performances on these legal datasets.

6.2 Chinese Legal LLMs

As for Chinese legal LLMs, recent work utilizes a paradigm of continual learning in the legal domain, and a substantial number of instruction fine-tuning datasets were constructed to augment the proficiency in rendering legal advice. Particularly, the series of LaWGPT (Song, 2021) has been developed by leveraging foundational models such as Chinese-LLaMA-7B (Cui et al., 2023b), ChatGLM (Du et al., 2022), and Chinese-alpaca-plus-7B (Cui et al., 2023b). Lawyer LLaMa (Huang et al., 2023) base on the more advanced Chinese-LLaMa-13B (Cui et al., 2023b), On the other hand, LexiLaw (Hai, 2023), built on the foundation of ChatGLM-6B (Du et al., 2022), underwent training through the application of three distinct methods, namely LoRA (Hu et al., 2022), P-tuning (Liu et al., 2021), and fine-tuning. Furthermore, Chatlaw (Cui et al., 2023a) received training based on both Ziya-LLaMA-13B-v1 (IDEA-CCNL, 2023) and Anima-33B (lyogavin, 2023). DISC-LawLLM (Yue et al., 2023) adopted legal syllogism prompting strategies to construct supervised fine-tuning datasets and fine-tune LLMs with legal reasoning capability. A primary reason hindering us from utilizing such existing models is that they have often been trained on those publicly legal tasks already. Therefore the zero-shot capabilities can not be truly reflected. We thus continue training the general Baichuan 7B on legal data by ourselves.

6.3 Retrieval-augmented Inference

In scenarios where language models are confronted with tasks necessitating an infusion of external knowledge, a retriever mechanism can be used to provide evidence. The Retrieval-Augmented Generation (RAG) (Lewis et al., 2020b) system incorporates a BERT-based (Devlin et al., 2019) Document Retrieval Process (DRP) and utilizes BART (Lewis et al., 2020a) for answer generation. Analogously, the EMDR2 (Yu et al., 2023) employs the expectation-maximization algorithm to account for multiple retrieved documents. The Atlas (Izacard et al., 2022b) builds upon the EMDR2 framework, and by synergistically training the retriever and reader components, it demonstrates few-shot learning capabilities commensurate with the 540B PaLM (Chowdhery et al., 2022). RETRO (Borgeaud et al., 2022) benefits from retrieval mechanisms on expansive corpora during its pre-training phase and exhibits performance in close alignment with those of GPT-3 (Brown et al., 2020b).

7 Conclusions

In this paper, we reformulate the zero-shot domain content generation of large language models as an adapt-retrieve-revise procedure. This approach combines the merits of efficient continual training of a smaller 7B LLM for domain adaptation, robustly retrieving the supporting evidence from an external knowledge base, and effectively leveraging the evidence-assessing and revision capabilities of GPT-4. Our proposal substantially enhances the zero-shot performance on the Chinese legal tasks.

[4] http://cail.cipsc.org.cn/index.html

8 Limitations

While this paper manages to validate the effectiveness of the proposal in the Chinese legal domain, the adapt-retrieve-revise method itself is a flexible framework, which is expected to be adapted to a wide range of domains. We leave it as future work. Due to the substantial costs of the GPT-4 API, we could only sample a subset of test data during the evaluation. Resolving the trade-off between the growing experimental costs and the validity of evaluation remains a challenge for the GPT-4 research in the future.

References

Baichuan-inc. 2023. Baichuan-7b. https://github.com/baichuan-inc/Baichuan-7B.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. 2023. EQUALS: A real-world dataset for legal question answering via reading chinese laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL 2023, Braga, Portugal, June 19-23, 2023*, pages 71–80. ACM.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

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

Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Tao Hai. 2023. Lexilaw. https://github.com/CSHaitao/LexiLaw.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report.

IDEA-CCNL. 2023. Fengshenbang-lm. https://github.com/pengxiao-song/LaWGPT.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022b. Atlas: Few-shot learning with retrieval augmented language models.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

lyogavin. 2023. anima. https://github.com/lyogavin/Anima.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: A legal case retrieval dataset for chinese law system. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2342–2348. ACM.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt-embeddings. https://platform.openai.com/docs/guides/embeddings/what-are-embeddings.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4.

Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, QiaoQiao She, Jing Liu, Hua Wu, and Haifeng Wang. 2022. DuReader-retrieval: A large-scale Chinese benchmark for passage retrieval from web search engine. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5326–5338, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation.

Pengxiao Song. 2021. Lawgpt. https://github.com/IDEA-CCNL/Fengshenbang-LM.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9701–9708. AAAI Press.

A Appendix

A.1 Legal Instruction Tuning

We build our legal instruction dataset by human experts. Due to privacy concerns, we are not allowed to disclose the annotated instruction data. However, we will release the instruction annotation guideline along with our 7B legal LLM. We show a template with a toy example below.

- **Due to the Article** x **in the law** y: [the corresponding content in the law]
- **Considering the fact that** [the fact]
- **The judgment is** [the conclusion]

A toy example could be:

- **Due to the article 91 of the Road Traffic Safety Law of the People's Republic of China:** [Whoever drives a motor vehicle after drinking alcohol shall be imposed upon the penalty of temporary seizure of his motor vehicle driving license for not less than 1 month but not more than 3 months, and be imposed upon a fine of not less than 200 Yuan but not more than 500 Yuan as well; whoever drives a motor vehicle when he is drunk shall be restricted by the traffic administrative department of the public security organ until he becomes sober, be detained for not more than 15 days, be imposed upon the penalty of temporary seizure of his motor vehicle driving license for not less than 3 months but not more than 6 months, and be imposed upon a fine of not less than 500 Yuan but not more than 2000 Yuan as well. Whoever drives a commercial operating motor vehicle after drinking alcohol shall be imposed upon the penalty of temporary seizure of his motor vehicle driving license for 3 months, and be imposed upon a fine of 500 Yuan as well; whoever drives a commercial operating motor vehicle when he is drunk shall be restricted by the traffic administrative department of the public security organ until he becomes sober, be detained for not more than 15 days, be imposed upon the penalty of temporary seizure of his motor vehicle driving license for 6 months, be imposed upon a fine of 2000 Yuan as well. Where anyone is penalized for twice or more within one year due to his drunken driving as prescribed in the preceding two paragraphs, his motor vehicle driving license shall be canceled, and he shall not drive a commercial operating motor vehicle within 5 years.]
- **Considering the fact that** [the man was riding a motorbike when drunk.]
- **The judgment is** [to be restricted by the traffic administrative department of the public security organ until he becomes sober, be detained for not more than 15 days, be imposed upon the penalty of temporary seizure of his motor vehicle driving license for not less than 3 months but not more than 6 months, and be imposed upon a fine of not less than 500 Yuan but not more than 2000 Yuan as well.]

A.2 Retrieval Modules

we leveraged multilingual E5-large, which is the SOTA family of text embeddings, which has been reported to outperform BM25, Contriever (Izacard et al., 2022a), and GPT embeddings (OpenAI, 2023b). Since the improvements of our method are consistent and substantial (+33.3% vs vanilla, +15.4%/23.9% vs retrieval baselines), we believe these gaps have shown sufficient robustness in our proposal.

However, we agree that including more established retrieval modules can enhance the robustness of our findings. Therefore, we added extra experiments and compared them with the current SOTA Chinese retrieval module CoROM following (Qiu et al., 2022) on the LegalQA dataset, the same setting as in Section 5.1. Table 3 shows the comparison results. We find that: (1) multilingual e5-large has a competitive performance with CoROM on query-based retrieval and vastly outperforms CoROM on answer-based retrieval; (2) for both modules, the answer-based retrieval primarily improves the retrieval quality than the query-based setting.

Retriever	Retrieval	Top-1	Top-5	Top-10
Multilingual E5-large	Query-based	45.8	61.3	70.5
Multilingual E5-large	Answer-based	65.3	84.5	88.5
CoROM	Query-based	47.5	60.8	71.5
CoROM	Answer-based	58.8	72.5	80.5

Table 3: **Results of different retrieval seteps**.

A.3 Datasets

Please refer to Table 4 for the statistics.

Datasets	CP	LCR	LegalQA	JEC-QA
# test	965, 219	965, 219	1,000	13,341

Table 4: **Statistics of datasets**.