Supplemental Enhancement of Action Segments: A Retrieval Optimization for Large Language Models in the Legal Domain

Anonymous ACL submission

Abstract

Utilizing the Retrieval-Augmented Generation (RAG) framework with large language models for question answering often results in low retrieval precision and recall rates. A solution to address this issue involves retrieving external knowledge at various granularities. However, this strategy typically suffers from decreased precision in coarse-grained retrieval and omissions in fine-grained retrieval. To overcome these challenges, we introduce a novel framework designed for the legal domain, named Supplemental Enhancement of Action Segments (SEAS). SEAS utilizes few-shot prompting to extract action segments from legal texts, which are then used to enhance the retrieval of complete legal texts. In the Japanese Law Retrieval task, SEAS significantly enhances the performance of three distinct embedding models. Furthermore, in the Chinese Legal Question Answering task, SEAS outperforms all baselines across all metrics.

1 Introduction

When using large language models (LLMs) for question answering, the Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2020) has become one of the most popular frameworks for reducing hallucinations (Zhang et al., 2023). Despite its advantages, the framework often encounters challenges with low precision and recall rates (Gao et al., 2023) in its retrieval processes. Recent studies have explored various strategies to enhance retrieval, including adjustments in retrieval granularity (Ram et al., 2023) and retrieval frequency (Izacard et al., 2022). Our research explores effective retrieval granularity within the legal question answering context, targeting statute law. Subsequently, we propose a novel method that integrates various retrieval granularities to improve retrieval precision.

There are two main challenges in this work: (1) defining and extracting effective retrieval granularity for legal question answering, and (2) balancing the trade-offs between granularity levels, where coarse granularity provides broader context but less precision, while fine granularity offers precise content but risks overlooking relevant details.

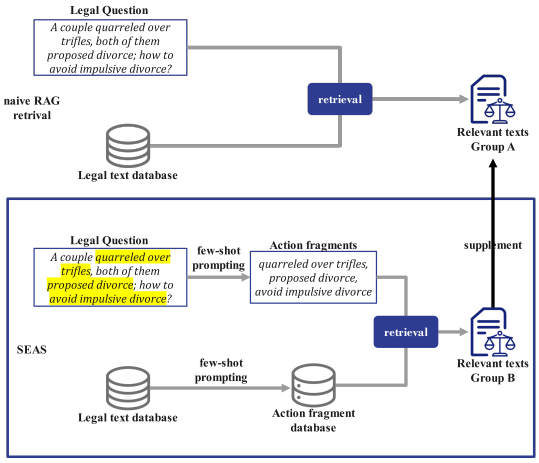


Figure 1: Overview of Supplemental Enhancement of Action Segments: (1) Extracting action segments from the questions and legal texts using few-shot prompting. (2) Supplementing and enhancing the retrieval results of complete legal texts with these action segments.

To address these challenges, we introduce a new framework, Supplemental Enhancement of Action Segments (SEAS) (see Figure 1), which utilizes "action segments" as fine-grained retrieval units within the legal domain. The concept of "action segments" is derived from the jurisprudential concept of "legal actions", which refers to actions implemented by individuals that produce legal effects. For example, in the legal question "*A couple quarreled over trifles, both of them proposed divorce; how to avoid an impulsive divorce?*" the action segments identified are "*quarreled for trifles,*" "*proposed divorce,*" and "*avoid impulsive divorce.*" These segments serve as fine-grained document chunks that supplement the retrieval of coarse-grained chunks, encompassing complete questions

and legal texts.

We evaluate SEAS through two experiments: (1) Japanese Law Retrieval from COLIEE's Task 3[1], and (2) Chinese Legal Question Answering. Our experiments show that incorporating action segments to enhance retrieval improves RAG performance significantly. In the Japanese Law Retrieval task, SEAS boosts the retrieval performance of three embedding models: BAAI/bge-large-en-v1.5[2], OpenAI text-embedding-3-small[3] and text-embedding-3-large[3]. In the Chinese Legal Question Answering task, SEAS enhances the performance of GPT-3.5-Turbo+RAG and GPT-4+RAG, with Accuracy (ACC) (Yue et al., 2023) improvements of 2.9% and 2%, respectively.

The main contributions of this paper are the proposal of a new framework, Supplemental Enhancement of Action Segments (SEAS). SEAS introduces two main innovations:

- To the best of our knowledge, this is the first method that uses action segments as retrieval granularities to enhance retrieval performance in legal question answering domain.
- SEAS combines retrievals of document chunks at different granularities, exploring optimization paths in the RAG framework.

2 Related work

A line of studies (Huang et al., 2023; Cui et al., 2023; Louis et al., 2023) has extended the RAG framework in the context of legal question answering. However, these methods suffer from low precision and recall rates in retrieval (Lewis et al., 2020). Research on the granularity of RAG retrieval (Khandelwal et al., 2019; Nishikawa et al., 2022; Kang et al., 2023) and chunking strategies (Langchain, 2023; Yang, 2023) seeks to improve precision and efficiency by using text chunks of varying sizes. Our framework improves overall precision by combining retrieval results of different granularities.

Determining appropriate granularity and obtaining fine-grained document chunks are two key challenges in our framework. Several studies (Min et al., 2023; Kamoi et al., 2023; Chen et al., 2023a) have investigated semantic representations of text at the propositional level. Building on this foundation, Chen et al. (2023b) have effectively utilized propositions as retrieval units. Inspired by these advancements, our approach integrates propositions from legal texts—action segments—as fine-grained retrieval units to address the first challenge.

Recent research has leveraged content generated by LLMs for retrieval and enhancement tasks (Gao et al., 2023), as demonstrated in studies by Wang et al. (2023), Yu et al. (2022), and Cheng et al. (2023). These studies highlight the innovative use of data sources within the RAG framework. Inspired by these developments, we have employed few-shot prompting with LLMs to extract fine-grained document chunks, thus addressing the second challenge.

3 Supplemental Enhancement of Action Segments

We introduce a novel framework, Supplemental Enhancement of Action Segments (SEAS), as illustrated in Figure 1. First, we devise the Action Segment Extraction (Section 3.1), which extracts action segments from legal texts via few-shot prompting. Then, using these action segments, we implement the Supplemental Enhancement (Section 3.2). This process supplements the retrieval results of complete text chunks with the results of action segments to produce the final relevant legal texts for the legal question.

3.1 Action Segment Extraction

Action Segment Extraction involves extracting text that describes actions from a legal text database and legal questions. We use few-shot (3-shot) prompting to extract text describing actions from each legal article in the database, thereby creating an action segment database. Similarly, we use a similar few-shot (3-shot) prompting method to extract text describing actions from the legal questions.

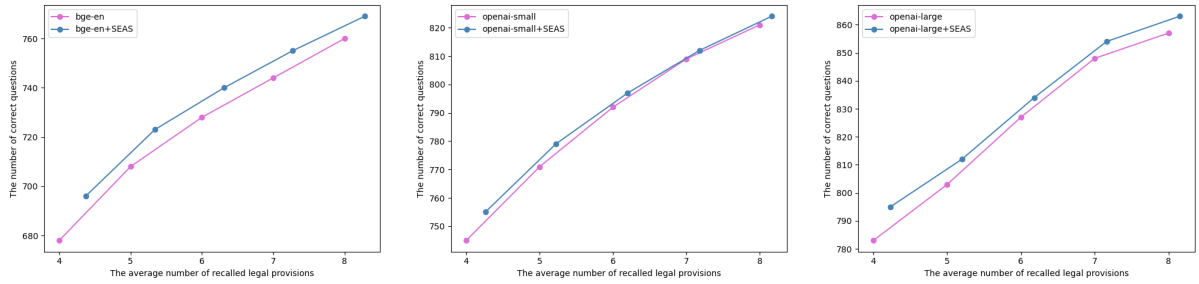
3.2 Supplemental Enhancement

First, we use an embedding model to encode the complete legal texts and legal questions, retrieving texts relevant to the legal question and selecting the top X legal texts. Next, we use the same embedding model to encode the action segment database along with the action segments extracted from legal questions, aiming to retrieve and select the top Y

[1]Competition on Legal Information Extraction/Entailment: https://sites.ualberta.ca/~rabelo/COLIEE2024/

[2]https://huggingface.co/BAAI/bge-large-en-v1.5

[3]Openai embedding model: https://platform.openai.com/docs/guides/embeddings



(a) bge-en (b) openai-small (c) openai-large

Figure 2: Evaluation results for the Japanese Law Retrieval task from COLIEE. The baselines are three embedding models: bge-en, openai-small, and openai-large. This figure shows the retrieval performance of these models after enhancement with SEAS.

action segments from legal texts. We then identify the original legal texts corresponding to these action segments, integrate these Y legal texts with the initially selected X legal texts, and perform deduplication to compile the final set of $[X,\ X+Y]$ relevant legal texts.

4 Experiment

We conducte two experiments to evaluate the retrieval effectiveness of SEAS and its impact on downstream task performance. The first experiment, Japanese Law Retrieval, was inspired by COLIEE's Task 3. This task involves extracting a subset of Japanese Civil Code Articles to answer a yes/no Japanese legal bar exam question, assessing the retrieval effectiveness of SEAS. In the second experiment, Chinese Legal Question Answering, we used SEAS to retrieve relevant articles and added them to the prompt for GPT-3.5-turbo (OpenAI, 2022) and GPT-4 (OpenAI, 2023) to generate answers to Chinese legal questions, evaluating the quality of the answers to assess the impact of SEAS on downstream generation tasks.

4.1 Japanese Law Retrieval

We evaluate the enhancement effects of SEAS on embedding models. In this experiment, a legal question can be related to multiple articles (in this dataset, a legal question is related to no more than six articles, with an average of 1.28 articles per question). The model identifies several articles related to the question. If these articles include all relevant ones, it is considered a successful retrieval. Our analysis focused on whether the SEAS-enhanced embedding model could identify a greater number of relevant articles while recalling the same total number of articles.

Datasets The data comes from Task 3 (English version) of the COLIEE 2023 datasets (Goebel et al., 2024), including the Japanese Civil Code texts and 1,097 yes/no Japanese legal bar exam questions from the training set, along with the articles relevant to each question.

Evaluation Metrics The retrieved articles are primarily used to enhance the downstream generative tasks of LLMs. Our testing shows that LLMs have the capability to select the correct statutes, making it particularly crucial that the retrieved articles comprehensively cover all relevant laws. Consequently, we evaluate retrieval effectiveness by counting the number of questions n for which the retrieved articles cover all relevant articles:

$$n = \sum_{i=1}^{1097} P(i) \tag{1}$$

$$P(x) = \begin{cases} 1 & Q_x \subseteq R_x \\ 0 & Q_x \nsubseteq R_x \end{cases} \tag{2}$$

where i represents the number of the question, $P(x)$ is a function to count, Q_x represents the set of relevant articles for the x-th question, R_x represents the set of retrieved articles for the x-th question in Equation (1, 2).

Baselines We choose three embedding models as baselines: BAAI/bge-large-en-v1.5 (output dimension 1024), OpenAI text-embedding-3-small (output dimension 1536), and text-embedding-3-small (output dimension 3072). We compare the number of questions for which the relevant articles are correctly retrieved before and after enhancement with SEAS.

Results The experiment evaluates the effectiveness of integrating the top 1 article retrieved

Model	ACC	CPL	CLR
General LLMs			
GPT-3.5-turbo	1.97	1.83	2.71
GPT-4	2.10	2.10	3.09
Chinese Legal LLMs			
DISC-LawLLM	2.43	2.22	3.08
Tongyi Farui	3.12	2.93	3.52
General LLMs with RAG			
GPT-3.5-turbo + RAG	3.14	2.76	3.70
GPT-4 + RAG	3.50	3.30	4.11
(SEAS; Ours)			
GPT-3.5-turbo + RAG + SEAS	3.23	2.82	3.72
GPT-4 + RAG + SEAS	**3.57**	**3.33**	**4.14**

Table 1: Evaluation results for the Chinese Legal Question Answering task. Baselines are General LLMs, Chinese Legal LLMs and General LLMs with RAG Framework. This figure shows the effectiveness of SEAS-enhanced LLMs in generating task outcomes.

by the SEAS with the top 4, 5, 6, 7, and 8 articles retrieved by the BAAI/bge-large-en-v1.5, text-embedding-3-small, and text-embedding-3-small models (see Figure 2). Upon incorporating the supplemental articles retrieved by SEAS, all three models improved performance, correctly retrieving relevant articles for more questions. Notably, the enhancement effect of SEAS was greatest for bge-en, followed by openai-large, and least for openai-small.

4.2 Chinese Legal Question Answering

In this section, we evaluate the impact of the SEAS-enhanced legal retrieval model on downstream answer generation tasks. Specifically, we use BAAI/bge-large-zh-v1.5[4] as Chinese embedding model and integrate articles retrieved for Chinese legal questions into the prompts and use GPT-3.5-turbo and GPT-4 to generate answers. The quality of these generated answers is evaluated using the DISC-LawLLM-eval (Yue et al., 2023) method, which involves inputting the question, the generated answer, and a reference answer into LLMs. Considering the proficiency of powerful LLMs like GPT-4 in aligning with human judgments—demonstrating more than 80% consistency (Zheng et al., 2023)—we employ GPT-4 to evaluate the quality of the generated answers.

Datasets The dataset comprises 222 Chinese civil law text questions along with their reference answers (Chinese version), including 62 questions from DISC-LawLLM-eval and 160 questions from Chinese legal consultations, justice-related publications, and other sources.

Evaluation Metrics We use the evaluation metrics from DISC-LawLLM-eval, including accuracy, completeness and clarity. (1) **Accuracy (ACC):** The consistency of the content and semantics of the answer with the reference answer. (2) **Completeness (CPL):** The answer do not omit any details compared to the reference answers. (3) **Clarity (CLR):** The juridical logic analysis of the answer is rigorous and clear, and the sentences are well-organized.

Baselines We select GPT-3.5-turbo, GPT-4, GPT-3.5-turbo + RAG, GPT-4 + RAG, and the Chinese legal large language models DISC-LawLLM (Yue et al., 2023) and Tongyi Farui[5] (commercial model) as baselines.

Results The experiment evaluates the effectiveness of integrating the top 3 articles retrieved by the SEAS with the top 3 articles retrieved by BAAI/bge-large-zh-v1.5 for generating answers. (see Table 1). The answers generated after supplementing with the top 3 articles retrieved by SEAS surpassed those generated by the unenhanced LLMs and the LLMs with RAG. Notably, GPT-4+RAG+SEAS achieved the highest performance, surpassing the generation effects of the Tongyi Fashui and DISC-LawLLM models. Specifically, the ACC of SEAS-enhanced GPT-3.5-turbo+RAG increased by 2.9%, and the ACC of SEAS-enhanced GPT-4+RAG increased by 2%.

5 Conclusion

In this work, we propose a novel framework, Supplemental Enhancement of Action Segments (SEAS). It generates fine-grained retrieval units—action segments—as retrieval granularity for legal domain questions through few-shot prompting and uses these segments to supplement and enhance the retrieval results of coarse-grained retrieval units—complete legal texts. Our framework combines the advantages of different granularity document chunks, optimizing the retrieval process. Experimental results show that SEAS improves the retrieval performance of various embedding models and guides downstream LLMs to generate better answers. We hope this work provides insights into optimizing RAG retrieval and can be applied to real-world scenarios.

[4] https://huggingface.co/BAAI/bge-large-zh-v1.5

[5] https://tongyi.aliyun.com/farui

Limitation

Despite SEAS being a model-agnostic framework that can be combined with other components, our study is limited in demonstrating generalizability across different types or scales of embedding models. Additionally, although the framework focuses on improving RAG retrieval and is domain-agnostic, our experiments were conducted only on two legal datasets, lacking tests in other domains. While SEAS effectively retrieves text chunks through few-shot prompting with LLMs, the generation cost becomes significant when the datasets are large.

Ethics Statement

The acquisition of data for this study was conducted with the explicit permission of the publisher, ensuring full compliance with all applicable legal and ethical standards. This project was conducted as a collaborative effort, and we fully acknowledge the contributions of each collaborator, ensuring a transparent and ethical process throughout the entire collaboration.

References

Sihao Chen, Hongming Zhang, Tong Chen, Ben Zhou, Wenhao Yu, Dian Yu, Baolin Peng, Hongwei Wang, Dan Roth, and Dong Yu. 2023a. Sub-sentence encoder: Contrastive learning of propositional semantic representations. *ArXiv*, abs/2311.04335.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Dong Yu, and Hongming Zhang. 2023b. Dense x retrieval: What retrieval granularity should we use? *ArXiv*, abs/2312.06648.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023. Lift yourself up: Retrieval-augmented text generation with self memory. *ArXiv*, abs/2305.02437.

Jiaxi Cui, Zongjia Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview and discussion of the competition on legal information, extraction/entailment (coliee) 2023. *The Review of Socionetwork Strategies*, 18(1):27–47.

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

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *ArXiv*, abs/2208.03299.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Conference on Empirical Methods in Natural Language Processing*.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *ArXiv*, abs/2305.18846.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *ArXiv*, abs/1911.00172.

Langchain. 2023. Recursively split by character.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Antoine Louis, G. van Dijck, and Gerasimos Spanakis. 2023. Interpretable long-form legal question answering with retrieval-augmented large language models. In *AAAI Conference on Artificial Intelligence*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *ArXiv*, abs/2305.14251.

Yuichi Nishikawa, A. Holobar, Kohei Watanabe, Tetsuya Takahashi, Hiroki Ueno, Noriaki Maeda, Hirofumi Maruyama, Shinobu Tanaka, and Allison S. Hyngstrom. 2022. Detecting motor unit abnormalities in amyotrophic lateral sclerosis using high-density surface emg. *Clinical Neurophysiology*, 142:262–272.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Sophia Yang. 2023. Advanced rag 01: Small-to-big retrieval.

W. Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *ArXiv*, abs/2209.10063.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *ArXiv*, abs/2309.11325.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.