RAG System for Supporting Japanese Litigation Procedures: Faithful Response Generation Complying with Legal Norms

Yuya Ishihara[1,*], Atsushi Keyaki[1,*], Hiroaki Yamada[2], Ryutaro Ohara[1,3] and Mihoko Sumida[1]

[1]*Hitotsubashi University, Japan*
[2]*Institute of Science Tokyo, Japan*
[3]*Nakamura, Tsunoda & Matsumoto, Japan*

Abstract

This study discusses the essential components that a Retrieval-Augmented Generation (RAG)-based LLM system should possess in order to support Japanese medical litigation procedures complying with legal norms. In litigation, expert commissioners, such as physicians, architects, accountants, and engineers, provide specialized knowledge to help judges clarify points of dispute. When considering the substitution of these expert roles with a RAG-based LLM system, the constraint of strict adherence to legal norms is imposed. Specifically, three requirements arise: (1) the retrieval module must retrieve appropriate external knowledge relevant to the disputed issues in accordance with the principle prohibiting the use of private knowledge, (2) the responses generated must originate from the context provided by the RAG and remain faithful to that context, and (3) the retrieval module must reference external knowledge with appropriate timestamps corresponding to the issues at hand. This paper discusses the design of a RAG-based LLM system that satisfies these requirements.

Keywords

Retrieval–Augmented Generation, Litigation Procedures, Legal Norms, Expert Knowledge, Information Retrieval

1. Introduction

In recent years, large language models (LLMs) have demonstrated remarkable advancements in their capabilities, leading to a growing movement toward their implementations into professional domains such as medicine and law. Since they are trained on extensive large text corpora, LLMs acquire a broad collection of knowledge throughout the training process [1, 2]. However, LLMs do not retain up-to-date information about events that occurred after their training period, and their knowledge of highly specialized or less common domains is not necessarily adequate. For these reasons, in professional domains such as legal[3] and medicine[4] , recent research has increasingly explored the use of Retrieval-Augmented Generation (RAG) approaches, which exploits external knowledge to generate high-quality responses.

Our research group is studying a RAG-based LLM system to support medical litigation procedures in Japan. Normally, litigation process goes as follows, (i)arranging issues, (ii)fact finding, (iii)legal evaluation, (iv)writing reasons of outcome and (v)writing sentences1. In the process of (i)arrange issues, each claim submitted by the plaintiff and the defendant is examined to distinguish the points of agreement from those in dispute, thereby extracting points that should be the focus points of the litigation. In medical litigation, this process involves technical advisors who are medical professionals. These experts provide technical explanations regarding matters that may require witness examination or expert testimony during (ii)fact finding, attend sessions involving the discovery of evidence or

This paper was reviewed and accepted by the program committee for BREV-RAG 2025: Beyond Relevance-based EValuation of RAG Systems, a SIGIR-AP 2025 workshop, held on December 10, 2025 in Xi'an, China.

*Corresponding author.
*Corresponding author.
yuya.ishihara@r.hit-u.ac.jp (Y. Ishihara); a.keyaki@r.hit-u.ac.jp (A. Keyaki); yamada@comp.isct.ac.jp (H. Yamada); r.ohara@ntmlo.com (R. Ohara); m.sumida@r.hit-u.ac.jp (M. Sumida)
0009-0003-8033-9927 (Y. Ishihara); 0000-0001-6495-117X (A. Keyaki); 0000-0002-1963-958X (H. Yamada); 0009-0000-8018-3895 (R. Ohara); 0000-0002-8531-2964 (M. Sumida)

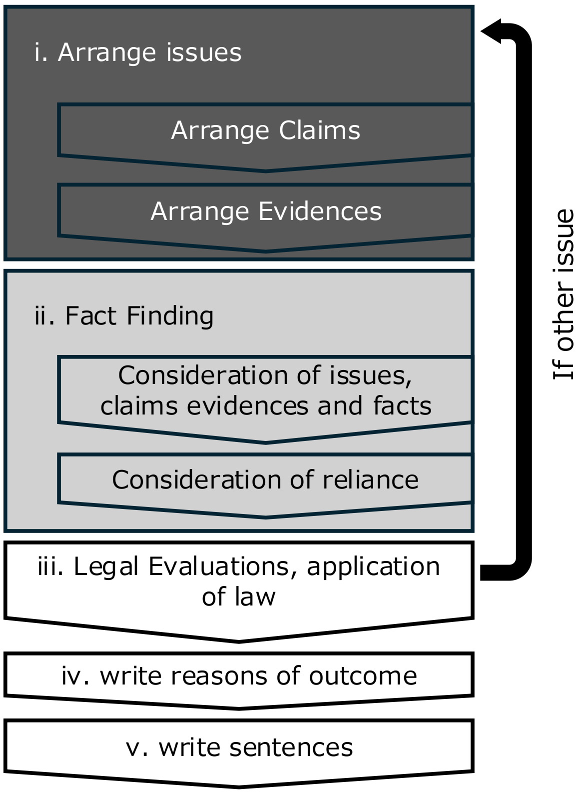

Figure 1: process of civil litigation

witnesses, and offer explanations to judges to aid in assessing the reliability of evidence and witness testimonies presented by the parties. Particularly, because fact-finding by judges in medical litigation requires domain-specific medical expertise, identifying the matters that should be subject to expert testimony demands extensive referencing and analysis of a large volume of legal and medical claims and related documents. Consequently, the workload in this process is extremely high. According to Professor Murata Wataru (Chuo University), a former judge, Japanese courts prepare internal reference materials summarizing the case overview and the issues for expert testimony and opinion to facilitate judicial deliberation. Thus, support by computers, especially by the application of LLMs and RAG, is expected to significantly enhance the efficiency of these processes. Furthermore, in a RAG-based LLM system designed to support medical litigation procedures, it is essential not only to provide accurate information but also to ensure compliance with the judicial system's norms. Based on this premise, we propose a set of requirements that a legal RAG framework should satisfy in order to fulfill those normative principles.

2. Requirements in medical litigations

As a preliminary requirement, it is essential to retrieve external knowledge that is appropriate and relevant to the issues of dispute. Although the accuracy of the retrieval module is one of the established evaluation points in RAG systems, conducting appropriate expert testimony in the context of medical litigation further requires to comply with the norms of civil procedure and consideration of the domain-specific characteristics within the medical litigations.

2.1. Procedural Requirements for Use of Expert Knowledge

In Japan's civil litigation procedure, regardless of whether a case involves a specialized domain, the adversarial principle is adopted. Judges make decisions from a neutral point based solely on the claims and evidence submitted by both parties. Consequently, judges are restricted from relying on issues not submitted or raised by any of the parties or on knowledge not contained within the submitted evidence, as doing so would infringe upon the procedural rights of the parties. This adversarial principle, has a tension with the expectation of the litigation system that judges continually update their understanding of precedents, statutes, and social norms. The extent to which judges should be permitted to conduct judicial investigation and use privately acquired knowledge has thus become a subject of debate, particularly in the context of the ongoing digital transformation of society. This requirement applies not only to medical litigations but also to other types of litigation, particularly those that are highly specialized and involve expert advisory systems, such as intellectual property, construction and system development litigations.

Currently, it is recognized that when judges rely on specialized knowledge, domain-specific expert knowledge can be invoked without the presentation of evidence only if it has undergone critical verification by the relevant expert community, and the parties must be guranteed an opportunity to contest the use of such knowledge[5]. Furthermore, in the civil procedure systems of the United States, the United Kingdom, and Germany, the use of privately acquired expert knowledge is considered to be permitted under certain requirements, such as it is being commonly shared within the relevant expert community, or the implementation of procedural requirements including disclosure to both parties and the provision of an opportunity for comment[6]. Accordingly, in the context of RAG systems, the external knowledge sources should be limited to those that have been critically validated by expert communities, and access to such information must be controlled to ensure that it remains equally available to both parties involved in the proceedings.

2.2. Reliance of Expert Knowledge and Frequent Updates

Judges, by their professional responsibilities, are required to continually update their understanding of statutes, judicial precedents, and social norms. In the medical litigation, moreover, judges are additionally expected not only to address the specialized nature of the cases in charge but also to ensure the reliability of the data and evidence upon which their judgments rely. Furthermore, doctors who serve as technical advisors in medical domains provide technical explanations grounded in their medical expertise, and also update them. Due to continuous advances in medicine, even authoritative data sources such as well-established medical textbooks, peer-reviewed papers, and clinical guidelines issued by professional communities gradually lose their validity over time. According to Professor Shigeto Yonemura (The University of Tokyo), a leading authority in Japanese medical law and also a doctor, approximately twenty percent of such data becomes outdated within five years. While the knowledge acquired through pretraining inevitably becomes obsolete, continuously retraining LLMs to update domain-specific knowledge would be an inefficient approach. Therefore, it is essential that generated responses explicitly derive from and be faithful to the context provided through retrievals.

2.3. Issue-specific Reference Time

One of the reasons why expert testimony in arranging issues can become a complex procedure is that the applicable standard of expert knowledge differs depending on the issue in dispute. For example, when the issue concerns whether a physician was negligent or not, the judgment must be based on the medical knowledge, standard of care, and medical law valid at the time the incident happened. In contrast, when the issue concerns the causal relationship between a medical treatment and its outcome, the judgment should rely on the most up-to-date knowledge available at the close of oral proceedings[7]. Therefore, it is necessary to reference external knowledge corresponding to the appropriate time period relevant to the issue in focus.

Although authoritative data gradually lose their validity over time, the transition to new authorized knowledge does not occur abruptly at once. During the transitional period, multiple streams of expert knowledge that contradict each other may coexist until new data or precedents become widely acknowledged. Therefore, when retrieving appropriate sources in a RAG framework, it is necessary to

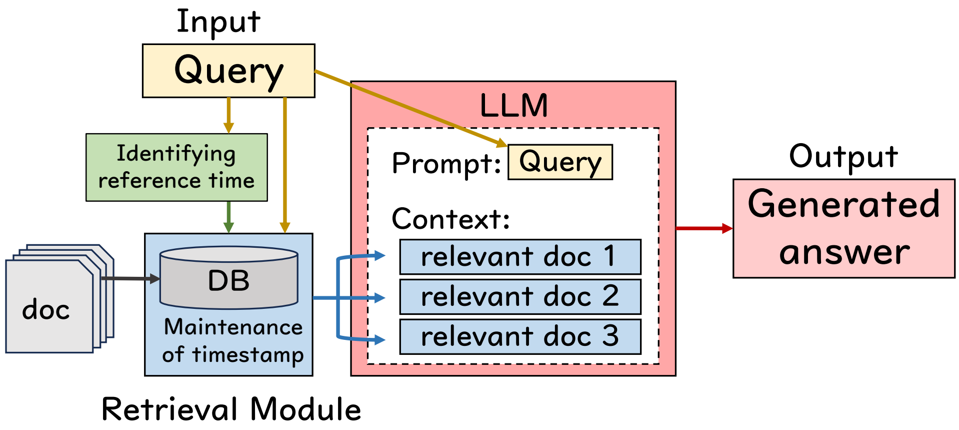

Figure 2: Overview of the a RAG-based LLM System

consider the expert knowledge that was valid at the relevant point in time.

Based on the above, this study addresses the realization of a “norm-compliant RAG” system, focusing on: (1)controlling knowledge sources in compliance with procedural requirements concerning the use of expert knowledge, (2)attribution and faithfulness of generated responses to their information sources; and (3)appropriateness of the published time of referenced sources.

3. Related Work

3.1. Retrieval-Augmented Generation (RAG)

Since large language models (LLMs) are trained on extensive corpora, they acquire various forms of knowledge during the learning process [1, 2]. However, they do not retain up-to-date information, such as current events that occur after model training, and their coverage of specialized or less common knowledge is often insufficient. As a result, LLMs are known to generate responses containing misinformation, commonly referred to as *hallucinations*. Because the training corpora used in constructing LLMs may not adequately include domain-specific expertise, the presence of hallucinations is particularly likely when applying such models to specialized domains.

Several approaches have been proposed to mitigate hallucinations, including improving the quality of training data [8], adjusting decoding strategies [9], enabling self-verification by the model [10, 11], and regenerating responses based on factual verification results [12]. Among these, Retrieval-Augmented Generation (RAG) [13, 14] has emerged as one of the most prominent and widely studied approaches.

An overview of the RAG framework is presented in Figure 2. First, the user’s input to the LLM is used as a query to retrieve relevant documents through a retrieval module. The retrieved documents are then provided to the LLM as contextual information, together with the user’s input. The LLM generates a response while referring to these relevant documents. By leveraging high-quality external information through RAG, previous studies have reported improvements in task performance [13, 14] and reductions in hallucination occurrence [15, 16, 17, 3, 4].

However, completely suppressing hallucinations remains challenging even when using RAG. For example, a study on the application of RAG in the legal domain [3] reported that, although hallucinations can be mitigated through RAG, they cannot be entirely eliminated. Consequently, the study emphasizes the importance of expert responsibility in verifying the texts generated by LLMs when applying AI within the legal field.

In addition, [3] conducted an evaluation of hallucinations based on accuracy and factuality. Therefore, to the best of our knowledge, no prior research has focused on compliance with legal norms or on the appropriateness of the knowledge sources that substantiate such compliance, which constitutes the central challenge addressed in this study.

3.2. Analysis of the Correspondence Between Information Sources and Responses

To verify whether the responses are faithfully generated based on the context provided by the RAG system, possible approaches include analysis using Data Attribution (DA) and evaluation methods related to response faithfulness.

In existing studies on Data Attribution (DA) [18, 19], the focus of analysis has been on the pre-training data of LLMs, known as Training Data Attribution (TDA). In contrast, in the RAG-based LLM system examined in this study, knowledge derived from the RAG component and that originating from the LLM's pre-training data may exist in a competitive relationship, thereby requiring a more complex analytical approach.

Additionally, within the RAG framework, mechanisms have been proposed to evaluate the faithfulness of responses with respect to the provided context. For example, Ragas[1] enables the computation of a Faithfulness Score, which assesses the degree of consistency between the context and the generated response. The Faithfulness Score determines, through natural language inference, whether the content of the generated response is supported by the information contained in the given context. Specifically, the Faithfulness Score is calculated through the following procedure:

1. Identify all the claims in the response.
2. Check each claim to see if it can be inferred from the retrieved context.
3. Compute the faithfulness score using the formula:

$$\text{Faithfulness Score} = \frac{\text{Number of claims in the response supported by the retrieved context}}{\text{Total number of claims in the response}} \tag{1}$$

4. Norm-compliant RAG

In this section, we discuss: (1) controlling knowledge sources in compliance with procedural requirements concerning the use of expert knowledge; (2) attribution and faithfulness of generated responses to their information sources; and (3) appropriateness of the published time of referenced sources, to realize a norm-compliant RAG.

4.1. Controlling Knowledge Sources in Compliance with Procedural Requirements

We restrict the use of external knowledge to sources that are acceptable according to civil litigation norms. We can achieve this control by controlling the scope of documents targeted by RAG retrieval and filtering the results. In assessing this aspect, we could simply label outputs derived from sources that deviate from the predefined scope as inappropriate.

4.2. Attribution and Faithfulness of Generated Responses to Information Sources

Expert knowledge is continuously updated over time. Thus, responses generated by relying solely on the model's knowledge acquired during its pre-training period can become easily outdated. RAG is the solution for this issue. To reinforce the effect of RAG, it is necessary to devise methods that generate responses faithful to the context (or documents) retrieved in the pipeline of the RAG approach. Possible approaches include explicit constraints through prompting and the introduction of chain-of-verification steps that check whether candidate sentences for generation are contained in the context.

To assess this aspect, we need to identify the data source or authority via DA analysis. We check whether the information contained in the outputs originates from RAG-derived knowledge or from pre-training. If it originates from pre-training, it is regarded as an inappropriate answer. Even if the

[1] https://docs.ragas.io/en/stable/

generated output is based on RAG-derived knowledge, if it contradicts the knowledge in the source, it should be considered inappropriate. Thus, it is also necessary to assess the faithfulness of the outputs against the source.

4.3. Valid Time Period of Referenced Sources

The older documents can be outdated if they are overruled due to new discoveries and updates. A naive solution to this issue might be keeping the knowledge always updated to the latest version; however, this solution would not work in our legal RAG setting.

The valid time periods of information sources vary depending on the types of issues raised in trials. For instance, if the issue of interest in a trial is a physician's negligence, the medical knowledge valid at the time of the physician's act can differ from that which is valid at the time of the trial, which is based on newer sources. If a system generates responses only according to the newest sources, it makes up an unrealistic conclusion based on knowledge unavailable at the time of the act in question. Moreover, the validity of time periods for sources matters not only in medical expert knowledge but also in legal expert knowledge, such as precedents and statutes.

Thus, a legal RAG system should be able to recognize and manage the validity of time periods for sources correctly. Managing the time metadata of information sources is important, especially concerning when information is published and when it becomes invalid. We could achieve this by utilizing timestamps and citation networks.

When assessing this aspect, if a generated response is based on information with inappropriate timestamps that do not align with the input query, it is considered unsuitable.

5. Conclusion

We are developing a RAG-based LLM system to support medical litigation proceedings in Japan. Such a system must not just present accurate information, but also provide legally compliant responses to support expert testimony. To accommodate the requirements, we propose aspects of "conformance to the norm" that a legal RAG system should satisfy. Specifically, we propose three aspects: (1) controlling knowledge sources in compliance with procedural requirements concerning the use of expert knowledge; (2) attribution and faithfulness of generated responses to their information sources; and (3) appropriateness of the published time of referenced sources.

Our future work includes proposing methods that satisfy each requirement, refining evaluation metrics, implementing them, and conducting experiments.

Acknowledgments

We appreciate Prof. Shigeto Yonemura (The University of Tokyo), Prof. Wataru Murata (Chuo University), Prof. Shozo Ota (Meiji University), Prof. Simon Deakin (University of Cambridge), and Prof. Felix Steffek (University of Cambridge) for their helpful comments. This work was partially supported by Minji-Funsou-Shori-Kenkyukikin, the Japanese Society for the Promotion of Science Grantin-Aid for Scientific Research (B) (#23H03686, #25K03178) and Scientific Research (C) (#24K15066), and JST PRESTO (#JPMJPR236B).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 and Grammarly for grammar and style suggestions. After using this tool, the authors reviewed and edited the content and take full responsibility for the publication's content.

References

[1] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, A. Miller, Language Models as Knowledge Bases?, in: Proc. of the EMNLP-IJCNLP 2019, 2019.

[2] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, Transactions on Machine Learning Research (TMLR) (2022) 2835–8856.

[3] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, D. E. Ho, Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, Journal of Empirical Legal Studies 22 (2025) 216–242.

[4] Y.-W. Chu, K. Zhang, C. Malon, M. R. Min, Reducing Hallucinations of Medical Multimodal Large Language Models with Visual Retrieval-Augmented Generation , in: Proc. of the AAAI 2025, 2025.

[5] G. Okanari, Saibankan no shichi riyou no kinshi (the prohibition of judge’s use of private knowledge), Hougaku Zasshi :(Journal of Law) of Osaka City University 68 (2021) 1 – 66.

[6] E. Sugiyama, Saibankan niyoru senmonchishiki no shushu to riyou (collection and use of expert knowledge by the judge, symposium: The discipline of civil judges in the exercise of their powers), Minso Zasshi (Journal of Civil Procedure) 69 (2023) 103 – 114.

[7] Y. Shirai, Mijukuji Moumakusyou to Ishi no Kashitu (Misdiagnosis of retinopathy of prematurity and Doctor’s Negligence), Hanrei kara Manabu Minji-Jijitsu Nintei: Jurisuto Zoukan (Special Edition of Journal: Jurist: Learning from Case Law: Civil Fact-Finding) (2006) 252–256.

[8] A. Albalak, Y. Elazar, S. M. Xie, S. Longpre, N. Lambert, X. Wang, N. Muennighoff, B. Hou, L. Pan, H. Jeong, C. Raffel, S. Chang, T. Hashimoto, W. Y. Wang, A Survey on Data Selection for Language Models, arXiv:2402.16827, 2024.

[9] K. Li, O. Patel, F. Viégas, H. Pfister, M. Wattenberg, Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, in: Proc. of the NeurIPS 2023, 2023.

[10] P. Manakul, A. Liusie, M. Gales, SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, in: Proc. of the EMNLP 2023, 2023.

[11] X. Zhang, B. Peng, Y. Tian, J. Zhou, L. Jin, L. Song, H. Mi, H. Meng, Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation, in: Proc. of the ACL 2024, 2024.

[12] Y. Wang, R. G. Reddy, Z. M. Mujahid, A. Arora, A. Rubashevskii, J. Geng, O. M. Afzal, L. Pan, N. Borenstein, A. Pillai, I. Augenstein, I. Gurevych, P. Nakov, Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers, in: Proc. of the Findings of the EMNLP 2024, 2024.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in: Proc. of the NIPS 2020, 2020.

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv:2312.10997, 2023.

[15] S. T. I. Tonmoy, S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, A. Das, A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, arXiv:2401.01313, 2024.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Reducing Hallucination in Structured Outputs via Retrieval-Augmented Generation, in: Proc. of the NAACL 2024, 2024.

[17] W. Zhang, J. Zhang, Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review, Mathematics 13 (2025).

[18] G. Pruthi, F. Liu, S. Kale, M. Sundararajan, Estimating Training Data Influence by Tracing Gradient Descent, in: Proc. of the NeurIPS 2020, 2020.

[19] T. A. Chang, D. Rajagopal, T. Bolukbasi, L. Dixon, I. Tenney, Scalable Influence and Fact Tracing for Large Language Model Pretraining, in: Proc. of the ICLR 2025, 2025.