Leveraging Ensemble LLMs and Contextual Embeddings for Case-Based Reasoning in the Legal Domain

Ramitha Abeyratne[1,*]

[1]Robert Gordon University, Aberdeen , Scotland

Abstract

This research investigates the integration of Case-Based Reasoning (CBR) with Retrieval-Augmented Generation (RAG) for Large Language Models (LLMs) to enhance the reliability of legal question-answering systems. Thus far, we have developed a structured retrieval mechanism using CBR to improve the contextual relevance of generative outputs. Additionally, we introduced two novel alignment-based evaluation metrics—weighted and unweighted—which demonstrated superior performance over existing baselines in assessing QA responses. Our experimental validation on a legal dataset confirmed the effectiveness of the CBR-RAG approach in improving response accuracy. Moving forward, we aim to refine weighting strategies for alignment metrics and enhance textual representations to improve evaluation robustness. Furthermore, we plan to extend our study beyond the legal domain by conducting a comparative analysis across multiple datasets, ensuring broader applicability of the CBR-RAG framework.

Keywords

CBR, RAG, LLMs, LLMs-as-Judges, Case alignment, Embeddings

1. Introduction

Retrieval Augmented Generation (RAG) has emerged as a powerful technique for enhancing the outputs of Large Language Models (LLMs) by incorporating external knowledge sources [1]. This approach is particularly crucial in specialised domains such as legal question-answering, where responses must be both highly accurate and contextually relevant. Standalone LLMs often suffer from hallucinations, largely due to their limited knowledge coverage and reliance on probabilistic text generation rather than factual verification [2]. However, conventional RAG systems typically depend on generic information retrieval mechanisms, which may not always provide structured or contextually appropriate content [3]. As a result, these limitations can lead to suboptimal outputs, reducing the overall reliability and trustworthiness of such systems.

To address these challenges, Case-Based Reasoning (CBR) presents itself as a structured retrieval framework that leverages past cases to inform new queries [4]. Unlike traditional information retrieval systems, CBR-based approaches contain multiple attributes that facilitate nuanced comparisons between cases [5], making them particularly advantageous for legal applications. By integrating CBR with RAG, retrieval processes can be significantly enhanced through structured similarity-based knowledge extraction to ensure improved contextual alignment and greater accuracy in generated responses [3]. This fusion of methodologies allows for more precise case retrieval, leading to well-informed and legally sound outputs.

Despite the potential benefits of RAG-based legal QA systems, their effectiveness is highly dependent on the availability of high-quality, annotated datasets that enable rigorous performance evaluation [6]. However, the manual annotation of legal datasets is an exceptionally resource-intensive task, requiring not only substantial time but also considerable domain expertise [7]. The complexity of legal texts further magnifies these challenges as annotations must adhere to strict legal interpretations and contextual nuances. Consequently, the scarcity of large-scale annotated datasets poses a significant hurdle in the development and refinement of RAG-based legal QA systems.

ICCBR DC'25: Doctoral Consortium at ICCBR-2025, July, 2025, Biarritz, France

*Corresponding author.

r.abeyratne@rgu.ac.uk (R. Abeyratne)

0009-0008-5582-8311 (R. Abeyratne)

To mitigate these challenges, automated annotation methodologies have emerged as a promising solution for enhancing the evaluation process [8]. Leveraging LLMs for dataset annotation, a concept recently introduced as 'LLM-as-a-Judge', has demonstrated potential in expediting this otherwise laborious task [8]. However, such existing methods are inherently susceptible to biases originating from the applied LLMs themselves, which can lead to erroneous evaluations. Moreover, specific biases—including positional bias, verbosity bias, and self-enhancement bias—can influence the reliability of automated evaluations depending on the type of assessment being conducted [6]. These biases can compromise the validity of the evaluation process and impact the reliability of legal QA systems.

To overcome these limitations, we propose an advanced AI evaluation framework based on ensemble LLMs functioning as collective judges. The framework begins with the development of case alignment-based assessment metrics to provide a structured and unbiased evaluation process. These novel measures avoid the bias issues related to existing methods. Next, the CBR-infused RAG framework is applied to a legal QA system, leveraging case-based reasoning to enhance retrieval and generation accuracy. Finally, the proposed evaluation metrics are used to assess the effectiveness of CBR-RAG, ensuring a balanced and reliable performance analysis. This approach strengthens the robustness and credibility of legal AI systems by refining evaluation methodologies, supporting more dependable and context-aware legal information retrieval.

2. Research Plan

2.1. Research Objectives

- **RO1**: Develop an AI-driven evaluation framework using ensemble LLMs-as-a-judge to reduce bias and enhance automated assessment through CBR-based alignment.
- **RO2**: Improve contextual embeddings from transformer networks to strengthen the representation quality required for reliable alignment-based evaluation.
- **RO3**: Build a CBR-based RAG retrieval approach to provide structured, context-aware inputs for generation, and evaluate its domain generalisability.

2.2. Approach / Methodology

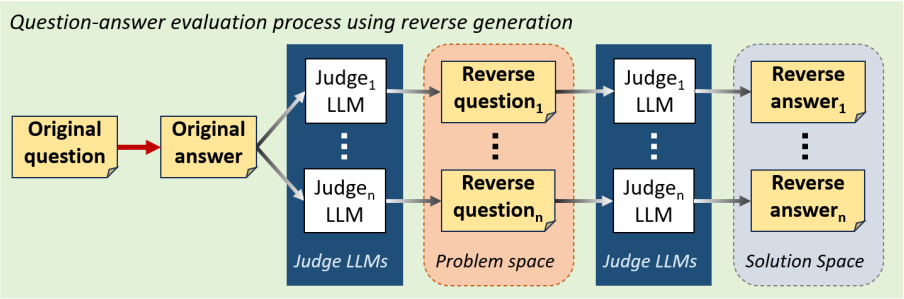

Figure 1: Reverse generation workflow for generative AI based evaluation methods

First, we create the QA evaluation framework using a concept called reverse generation (*RO1*). This means that given an answer to a question, we use the answer to reverse-generate the question using LLMs acting as judges. This reverse-generated question is then used to generate a reverse-generated answer via the sames judges. We establish our first evaluation metric by using the embedded representations of reverse-generated questions and reverse-generated answers to form the problem and solution spaces as shown in Figure 1. Contextual embeddings from the generating model are used to convert the text responses to embeddings. The relevance of an answer to the original question is assessed by evaluating the alignment between the original question and answer within this space. This concept is heavily inspired by case alignment literature [9]. We name this evaluation metric Inter-Language

Model Reconstruction Alignment (ILRAlign). Another metric called Weighted ILRAlign (WILRAlign) is formed by assigning dynamic weights based on problem similarity to improve evaluation reliability.

To address limitations in current alignment reliability, we enhance the alignment-based measures by optimising both weighting mechanisms and contextual embeddings (RO2). This builds directly on RO1, where weighting can be applied to the problem and solution space alignment measurements, as well as during dynamic weighting in WILRAlign. For the initial set of experiments, we utilise contextual embeddings of text from the last hidden layer. We also aim to explore strategies such as basic embeddings and aggregated embeddings from different layers [10]. This is expected to have a significant impact on the results as the evaluation metric relies entirely on numerical representations.

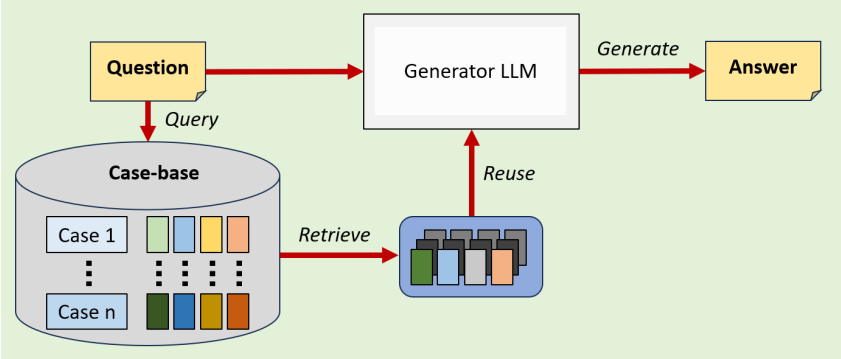

Figure 2: CBR infused Retrieval Augmented Generation architecture

Finally, we implement a Case-Based Reasoning and Retrieval-Augmented Generation (CBR-RAG) architectural method to enable structured retrieval of legal cases and improve the contextual grounding of LLM outputs (RO3). This involves developing a case-based retrieval mechanism that indexes legal cases to leverage similarity knowledge containers, thereby improving retrieval performance. Figure 2 denotes this in the high level where multiple cases are retrieved and fed into the generator. We utilise two types of embeddings for case-base embedding: one optimised for retrieval and another for similarity assessment. Additionally, dynamic weighting of case attributes is applied to enhance retrieval accuracy. RO3 is designed to be disjoint from RO1 and RO2, and is evaluated using the matured output of RO2 to ensure effective integration and factual grounding.

3. Progress Summary

The question and answer evaluation metrics (RO1) - ILRAlign and WILRAlign, were successfully implemented and tested on the Australian Legal Question Answering (ALQA) [11] and Sri Lankan Supreme Court (SLSC) datasets. We generated 10 sets of answers for these datasets by varying the generator model's temperature parameter. Four 7-billion-parameter LLMs were used in the experiments. This included variations from Gemma, Llama, Mistral and Falcon. A leave-one-out strategy was employed for evaluating the judges.

Metric	ALQA	SLSC
Judge answer similarity mean	+ 0.0002	+ 0.0811
Judge reverse question similarity mean	+ 0.3006	+ 0.1986
ILRAlign	+ 0.8299	+ 0.8376
WILRAlign	+ **0.8025**	+ **0.8101**

Table 1
Average correlation on gold standard with Pearson coefficient of alignment based QA evaluation metrics.

Similarity was measured using cosine similarity by comparing the candidate answers to the ground truth. Finally, Pearson's coefficient was used to calculate the correlation between the metrics and

variations in the gold truth similarity. Table 1 presents these results, demonstrating that our novel metrics outperform traditional techniques by a significant margin. The detailed findings will be published (already accepted) as a conference proceeding for International Conference on Case-Based Reasoning (ICCBR 2025).

Document count for RAG	No embeddings	BERT	LegalBERT	AngleBERT
No RAG	0.897	N/A	N/A	N/A
1 - RAG with context	N/A	0.899	0.902	0.912
1 - RAG with full case	N/A	0.907	0.904	0.907
3 - RAG with context	N/A	0.900	0.903	0.909
3 - RAG with full case	N/A	0.900	0.905	**0.914**

Table 2
CBR RAG method evaluation with baselines for BERT, LegalBERT and AngleBERT [3]

Table 2 presents the results of the CBR-RAG experiments (RO3). These show the cosine similarity between the generated answer and ground truth. Generation without the RAG context and retrieval of only the support text as context were used as baselines. Multiple experiments were conducted, querying 1 and 3 documents using different retrieval mechanisms. We weighted these retrieval mechanisms based on empirical analysis. The results demonstrate that case retrieval using AngleBERT yields the best generation results, closely aligning with the ground truth. The details of these experiments were published at the ICCBR 2024 conference with the [3] publication.

4. Future Work

The alignment-based question and answer evaluation metric which utilises an ensemble of LLMs, produced significant improvements compared to existing baseline approaches. This advancement highlights the potential of ensemble-based methodologies in enhancing the accuracy and fairness of automated legal assessments. However, the initial prototype was developed with equal weightings assigned to both the problem space and solution space when computing alignment. While this provided a balanced approach, it may not always reflect the true complexity and nuance of question-answering tasks. A key area for future research involves refining these weightings to optimise alignment calculations and ensure a more contextually appropriate evaluation framework. Additionally, we plan to explore more sophisticated weighting strategies for WILRAlign, moving beyond the simple normalised weighting derived from problem space similarity. By incorporating adaptive and dynamic weighting mechanisms, we aim to enhance the precision of our evaluation process.

Furthermore, we will focus on enhancing the representation of textual data as embeddings, which plays a crucial role in the effectiveness of our evaluation metrics. Given the multiple layers within a transformer model, we intend to conduct a comprehensive study to identify the most effective representation for alignment measurement. A deeper understanding of how different embedding layers influence alignment computations will allow us to obtain improved performance. This investigation will include an exploration of alternative embedding methodologies to optimise representation learning for legal text alignment.

Subsequently, we aim to validate the adaptability of our Case-Based Reasoning-infused Retrieval-Augmented Generation (CBR-RAG) architecture. While initial tests on the Australian Legal QA dataset have yielded promising results, we plan to conduct a comparative study across multiple domains such as healthcare, finance and education to assess its broader applicability and generalisability. Extending our evaluation to different domains will provide deeper insights into the robustness and scalability of our approach. By assessing its effectiveness across diverse knowledge-intensive sectors, we can refine the model to better accommodate the specific requirements of each domain.

5. Conclusion

This study successfully developed an AI-driven evaluation framework that leverages ensemble LLMs to enhance the grading of legal QA systems. Our proposed alignment-based metrics—ILRAlign and WILRAlign—demonstrated superior performance in assessing the overall quality of generated legal responses. Furthermore, by integrating CBR with RAG, we enhanced the retrieval of legal cases, ensuring that generated responses are informed by structured, contextually relevant knowledge. This advancement improves the overall generation quality of LLMs, making them more reliable for legal applications.

While our current evaluation framework applies uniform weightings to the problem and solution spaces, future research will focus on dynamic weighting mechanisms to improve reliability. Additionally, further investigation into transformer-based embedding optimisation will enhance representation learning, with the goal of refining legal text alignment measurements. By continuously improving these methodologies, we aim to develop a highly effective evaluation framework that not only improves legal NLP applications but also serves as a benchmark for AI-driven legal reasoning. This work lays the foundation for more robust AI-driven legal reasoning to advance in the field of legal NLP applications towards greater reliability and precision.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT for the purpose of: grammar and spelling check, paraphrase and reword. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55. URL: http://dx.doi.org/10.1145/3703155. doi:10.1145/3703155.

[3] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, B. Fleisch, Cbr-rag: Case-based reasoning for retrieval augmented generation in llms for legal question answering, 2024. URL: https://arxiv.org/abs/2404.04302. arXiv:2404.04302.

[4] A. Upadhyay, S. Massie, A case-based approach for content planning in data-to-text generation, in: Int. Conf. on CBR, Springer, 2022, pp. 380–394.

[5] I. Watson, A case-based persistent memory for a large language model, 2024. URL: https://arxiv.org/abs/2310.08842. arXiv:2310.08842.

[6] S. Tan, S. Zhuang, K. Montgomery, W. Y. Tang, A. Cuadron, C. Wang, R. A. Popa, I. Stoica, Judgebench: A benchmark for evaluating llm-based judges, 2024. URL: https://arxiv.org/abs/2410.12784. arXiv:2410.12784.

[7] J. Choi, J. Yun, K. Jin, Y. Kim, Multi-news+: Cost-efficient dataset cleansing via LLM-based data annotation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), 2024 EMNLP, ACL, Miami, Florida, USA, 2024, pp. 15–29. URL: https://aclanthology.org/2024.emnlp-main.2/. doi:10.18653/v1/2024.emnlp-main.2.

[8] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, J. Guo, A survey on llm-as-a-judge, 2025. URL: https://arxiv.org/abs/2411.15594. arXiv:2411.15594.

[9] S. Massie, N. Wiratunga, S. Craw, A. Donati, E. Vicari, From anomaly reports to cases, 2007, pp. 359–373. doi:10.1007/978-3-540-74141-1_25.

[10] C. Tao, T. Shen, S. Gao, J. Zhang, Z. Li, Z. Tao, S. Ma, Llms are also effective embedding models: An in-depth overview, 2024. URL: https://arxiv.org/abs/2412.12591. arXiv:2412.12591.

[11] U. Butler, Open australian legal qa, 2023. URL: https://huggingface.co/datasets/umarbutler/open-australian-legal-qa. doi:10.57967/hf/1479.