LawNet AI Search: Performance Evaluation



September 2026

Authors

Mark FISHER, Singapore Academy of Law

Michelle CHUA, Singapore Academy of Law

Jethro LOW, LawNet Technology Services

Justin TAN, LawNet Technology Services

About

SAL's mission is to enable future-ready, trusted legal professionals. The statutory body works with members and stakeholders to set new precedents of excellence in Singapore law through developing legal thought leadership, world class legal technology infrastructure and legal solutions to support our members.

LTS is the technology company behind LawNet, Singapore's leading portal for legal research, information and transactions. A wholly-owned subsidiary of the Singapore Academy of Law, LTS continues to develop bold and innovative products and services that will better serve the needs of the legal community.

Acknowledgements

SAL and LTS would like to thank the following parties for their comments on a draft of this report:

- BizTech Group, Infocomm Media Development Authority
- Dr. Pavandip Singh Wasan, TinyRobot.AI
- Stella Chen, Lee Shen Yang, Rodney Yap and Zeng Hanyi, AI in Law Consultation Group, Singapore Academy of Law

The authors would also like to record their appreciation to Catherine Lim (Intern, Singapore Academy of Law) for her assistance in drafting this paper.

Disclaimer

Whilst every effort has been made to ensure that the information contained in this report is correct, the Singapore Academy of Law and LawNet Technology Services disclaim all liability and responsibility for any error or omission in this report, and in respect of anything, or the consequences of anything, done or omitted to be done by any person in reliance, whether wholly or partially, upon the whole or any part of the contents of this report.

Introduction

Background of LawNet

1. The Singapore Academy of Law (**SAL**) has operated LawNet, Singapore's leading legal research portal, since 1990.[1] LawNet services legal practitioners in Singapore by providing access to authoritative legal research material, including judgments from the Singapore courts, as well as books and secondary material published under SAL's Academy Publishing brand.

Development of LawNet AI Search

2. In 2024, SAL and its technology arm, LawNet Technology Services (**LTS**), entered into a collaboration with the BizTech Group of Infocomm Media Development Authority (**IMDA**) to build **LawNet AI Search**[2], an AI-powered legal research engine that combines a fine-tuned large language model (**LLM**) with retrieval-augmented generation (**RAG**) technology. The result of this was LawNet AI Search, which was first launched in September 2025 and is currently operated by LTS.

3. In simple terms, to use LawNet AI Search, a user keys in a legal research question, and the system generates an answer to that question grounded on documents sourced from LawNet's legal corpus.[3]

4. This paper arises as a result of SAL's ongoing commitment to transparency and the evaluation of LawNet products. It aims to evaluate LawNet AI Search's performance at the time of writing.

Evaluation results

Overview

5. We conducted an evaluation exercise of LawNet AI Search in 2026. The aim of our evaluation exercise was two-fold:

    a. First, to assess the quality of the answers produced by the system. Generally, there are two key components of a RAG system that can be evaluated:

        (i) **Retrieval.** This refers to the system's ability to retrieve the relevant source documents in response to a user query. In the context of LawNet AI Search, this concerns whether the relevant judgment and statute excerpts were retrieved in response to a particular query.

        (ii) **Generation.** This refers to the system's ability to generate an accurate and useful response based on the retrieved information. This includes assessing whether the generated answer correctly addresses the user query and is adequately grounded in the retrieved judgment and statute excerpts.

---

1 "LawNet: 35 Years of Evolution and Innovation", *Singapore Academy of Law,* 11 September 2025.

2 Previously referred to as "GPT-Legal Q&A".

3 These can be from judgments and legislation. Legal books published by SAL's Academy Publishing arm will also be added to the document corpus in 2026.

b. Second, to assess how well LawNet AI Search performed across different areas of legal research, given that its embedding model was fine-tuned based on contract law material. The generative model, on the other hand, was fine-tuned based on synthetic datasets that covered various areas of law.

Findings

6. Our findings are as follows:

   a. The generation component of the RAG system is generally capable of producing relevant and well-grounded legal responses when provided with the correct supporting authorities.

   b. The retrieval quality is a major determinant of end-to-end performance. When supplied with the correct reference contexts, the fine-tuned LLM produces substantially faithful and factually correct answers, indicating scope for improvement through retrieval optimisation.

   c. Although the embedding model was fine-tuned on one area of law, LawNet AI Search performs comparably well for other areas of law (see further below). Across the legal topic areas represented in the evaluation set, we observed no material variation in RAGAS performance.

7. We purposively selected five areas to provide coverage across the identified legal topic areas rather than statistical proof of performance equivalence between them. Given the sample size, the results should be interpreted as evidence that no topic-specific performance degradation was observed in the evaluated cases, rather than evidence that such differences cannot exist.

Evaluation dataset

8. For this evaluation, we used a dataset comprising of 30 query–context–answer (**QCA**) triplets. "Triplets" refers to each QCA pair having three components. The 30 triplets were created by the SAL Legal Research team, and evenly distributed across the top five areas of law for legal research:

   a. civil procedure;

   b. contract law;

   c. criminal law;

   d. tort law; and

   e. statute-based queries.

9. Each triplet comprises of:

   a. a legal question (the "question");

   b. the legal materials from which the answer was derived—this was in the form of excerpts of Singapore judgments and/or legislation (the "context"); and

   c. a written answer grounded in those materials (the "model answer").

10. These triplets were created to cater for common query archetypes including:

    a. definitional questions;

    b. questions about judgments (including holdings);

    c. questions about statutory instruments; and

d. questions regarding purpose or policy.

11. In creating these triplets, we also accounted for variation in:

    a. the context required to answer a query (e.g. the number of document excerpts required to answer a question);

    b. whether the answer is located in a piece of legislation or text of a judgment (or both); and

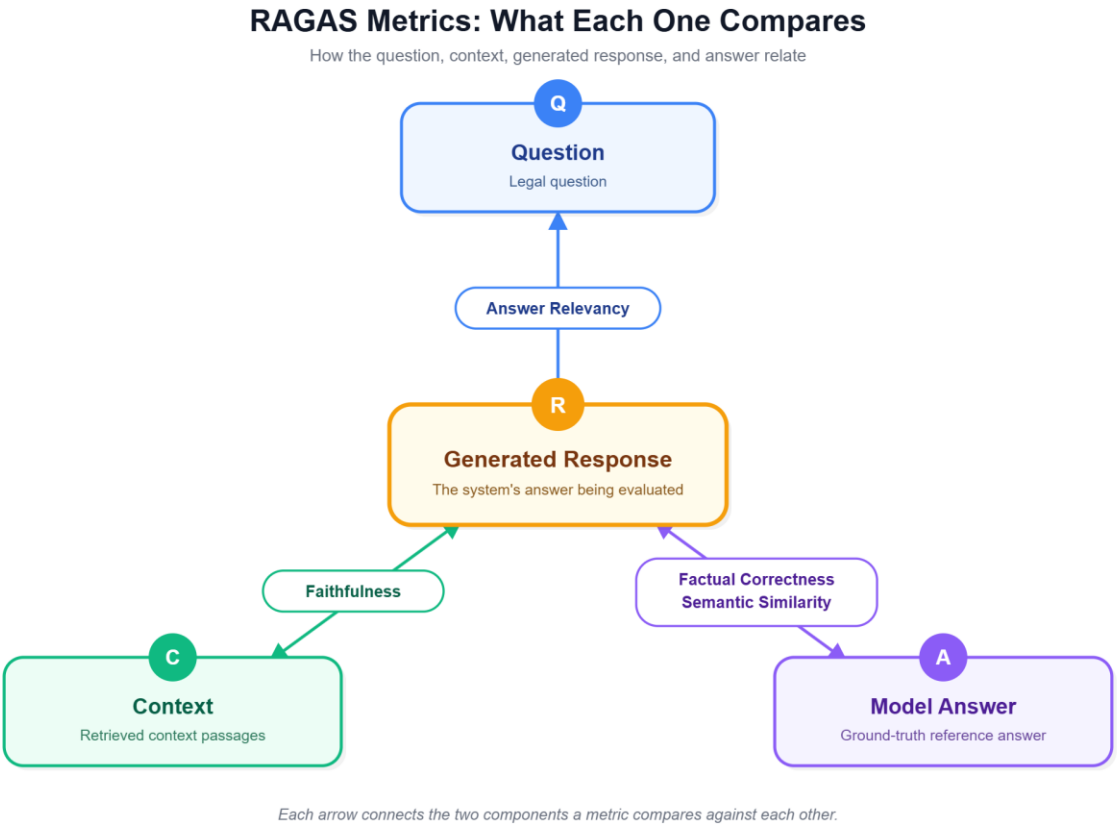
    c. whether the answer required synthesis from different materials.

RAGAS metrics

12. We used Retrieval-Augmented Generation Assessment (**RAGAS**) generation metrics to assess LawNet AI Search.

13. RAGAS is an evaluation framework designed specifically for measuring the quality of RAG systems. Instead of merely checking whether the final answer is correct, RAGAS evaluates both the retrieval and generation steps in a RAG pipeline.

14. The RAGAS generation metrics are set out in the table below.



Metric	Brief description	How it is measured
Factual correctness[4]	Measures the factual accuracy of a generated response against the answer in the relevant QCA triplet.	An LLM decomposes both the generated response and the reference answer into atomic claims. Natural language inference (NLI) is then used to verify overlap in the generated claims against the reference claims.
Faithfulness	Measures the factual consistency of the generated response against the relevant context (whether retrieved by the system or provided by us). A low score may indicate hallucinations or unsupported claims.	An LLM extracts factual claims from the generated response and verifies whether each claim is supported by the retrieved context. The faithfulness score reflects the proportion of claims that can be directly grounded in the retrieved information.
Semantic similarity	Measures the semantic similarity between a generated answer and the answer in the relevant QCA triplet.	Both responses are converted into vector representations using an embedding model, and the score is computed as the cosine similarity[5] between the two embeddings. No LLM judge is involved.
Answer relevancy	Measures the relevance of the generated response against the query.	An LLM is instructed to generate three hypothetical questions that could be answered by the generated response. The score is the average cosine similarity between the vector embedding of each hypothetical question and that of the actual question in the relevant triplet.

[4] In our exercise, RAGAS factual correctness was run in recall mode at high atomicity and coverage settings. Atomicity refers to how much a sentence is broken down into its meaningful claims, while coverage refers to how comprehensively the claims represent the information in the original sentence. High atomicity and coverage will produce claims that provide a detailed breakdown of the original information.

[5] Cosine similarity measures the similarity between two vectors by comparing the cosine of the angle between them, focusing on directionality instead of size.

Evaluation process

15. In relation to factual correctness, faithfulness and answer relevancy (but not semantic similarity), LLM-as-a-Judge was used as part of the evaluation process. This refers to a technique in which an LLM is used to assess one or more metrics, rather than a human, as an objective, scalable, and lower cost alternative. All LLM-judged metrics were computed using GPT-4o[6] as the evaluator model, at temperature 0.0[7], with text-embedding-ada-002[8] as the evaluator embedding model. Each metric was averaged over five runs, to reduce the impact of run-to-run variability of LLM-based evaluations.

16. Evaluation with the RAGAS metrics was carried out on LawNet AI Search under two conditions:

    a. First, "end-to-end", which measures the entire system's performance as a whole (i.e. assessing both LawNet AI Search's retrieval and generation capabilities); and

    b. Second, with supply of the *correct* context to the LLM that generates responses to user queries. This measures the generation capability of the LLM, under the assumption of perfect retrieval (i.e. given that the correct context has already been provided to the LLM). This is referred to below as "reference context supplied".

Results of RAGAS evaluation

17. Our evaluation exercise showed that:

    a. LawNet AI Search performed equally well across all five legal research areas under the "reference context supplied" condition; and

    b. The retrieval process significantly influenced "end-to-end" performance results. This can be inferred from the difference in scores between the "end-to-end" condition and the "reference context supplied" condition.

18. Under the "reference context supplied" condition, answer generation by the LLM in LawNet AI Search received generally high scores. Factual correctness sits in a narrow 65.9%–78.4% band across all five legal research areas. Faithfulness exceeds 90% in four out of five legal research areas (contract law, at 81.21%, is the exception). Semantic similarity exceeds 90% and answer relevancy exceeds 95% in all five legal research areas.

---

6 GPT-4o model version 2024-08-06.

7 Temperature is a model parameter that controls the degree of randomness in an LLM's output. A temperature of 0.0 makes the model more deterministic by strongly favouring the highest-probability response during generation, reducing output run-to-run variability. This is commonly used in LLM-based evaluation settings to improve scoring consistency and reproducibility.

8 Azure text-embedding-ada-002 model version 2.

Areas of law	Factual correctness		Faithfulness		Semantic similarity		Answer relevancy	
	End-to-end	Ref context supplied	End-to-end	Ref context supplied	End-to-end	Ref context supplied	End-to-end	Ref context supplied
Civil Procedure	52.35%	67.91%	89.58%	100.00%	90.89%	93.79%	98.30%	97.71%
Contract law	57.16%	67.64%	87.50%	81.21%	94.16%	95.28%	97.71%	95.58%
Criminal law	34.80%	76.81%	67.26%	100.00%	93.90%	94.40%	99.12%	99.22%
Statute-based queries	65.67%	65.94%	66.67%	94.44%	93.35%	94.46%	97.07%	95.83%
Tort law	38.02%	78.35%	84.44%	97.73%	90.86%	94.00%	98.31%	97.27%
All areas	**49.60%**	**71.33%**	**79.09%**	**94.68%**	**92.63%**	**94.39%**	**98.10%**	**97.12%**

Table 1 – RAGAS generation metrics across areas of law under two conditions ("end-to-end" and "reference context supplied")

19. However, when the "reference context supplied" condition is compared to the "end-to-end" condition:

    a. Semantic similarity (>90%) and answer relevancy (>95%) remain high in both conditions, indicating broad topical and semantic alignment;

    (iii) Factual correctness falls in every area when using LawNet AI Search's retrieved context (i.e. top 15 most relevant excerpts retrieved by the system). This suggests the retrieval process is likely a main contributor to the observed performance gap;

    (iv) Faithfulness scores are varied, being higher than 80% for civil procedure, contract law and tort law, and higher than 65% for criminal law and statute-based queries. This suggests different areas of law are sensitive to retrieval quality in varying degrees. Further investigations will be conducted to verify this.

20. The "reference context supplied" scores suggest that the generation component of the RAG system is generally capable of producing relevant and well-grounded legal responses when provided with the correct supporting authorities. The substantial improvement in faithfulness (79.09% to 94.68%) and factual correctness (49.60% to 71.33%) from the "end-to-end" to the "reference context supplied" condition could suggest that retrieval errors contribute to the end-to-end performance degradation.

Conclusion

21. From the above results, we can conclude that:

    a. The fine-tuned LLM performs consistently well across the different areas of law in the dataset when given the right material (i.e. under the "reference context supplied" condition). It can generate relevant and well-grounded answers for different areas of law even though the embedding model was fine-tuned on contract law material.

    b. The retrieval process is likely the main contributor to performance gaps in the system on the whole. We would thus direct future work to the retrieval layer in order to improve the "end-to-end" performance of the system.

22. Further, the results suggest that "area of law" may not be a technically meaningful dimension along which one should expect the system's performance to vary. The system does not execute a specific retrieval or generation pathway contingent on an identified "area of law". The taxonomy of "area of law" is also not a clean one. Judgments, for instance, can and do address more than one area of law and sometimes the interaction thereof. Such doctrinal labels are helpful for lawyers in organising legal knowledge, but what is being evaluated is the components of RAG system and it is not immediately apparent why different areas of law would be confined within statistically homogeneous classes of language in a retrieval system. Further investigations are planned to validate this hypothesis.

23. This paper represents an initial step in identifying potential areas of improvement to LawNet AI Search, and to that end, provides an empirical basis for prioritising improvements to the retrieval process. Additional work is also required to develop QCA triplets along the dimensions of question archetypes and areas of law.

Contact us

24. If you would like to use SAL's evaluation data set of thirty QCA triplets to evaluate a legal research tool, or provide feedback on this paper, we can be contacted at LawNet_AI@sal.org.sg.

25. Evaluation remains an ongoing effort. If you are a legally trained person and are keen to volunteer to help us develop more QCA triplets, in order to cover a wider range of question topics and archetypes, please contact us at LawNet_AI@sal.org.sg.