GLOBAL AI ASSURANCE SANDBOX

Scan to read the full case study

LawNet AI: AI-Powered Legal Search

Application tested

SAL Singapore Academy of Law | LawNet

Singapore Academy of Law (SAL) has deployed a single-turn Question & Answer (Q&A) application designed to retrieve relevant cases and legislation to answer legal research questions

Tester

resaro

Resaro offers independent, third-party assurance of mission-critical AI systems with extensive experience in testing Computer Vision and Generative AI systems

How were LLMs used in application?

Retrieval Augmented Generation

What risks were considered relevant and tested?

- Correctness of answers (including faithfulness and lack of completeness)
- Data Leakage (specifically system prompt leakage)
- Reliability and Robustness to perturbations (consistency of responses for identical prompts and meaning-preserving perturbations of prompts)

How were the risks tested?

- Custom evaluation sets were created to test for risks using realistic legal Q&A scenarios. SAL's legal experts generated a dataset and Resaro supplemented this with a synthetic dataset. Both of these were also perturbed to test additional aspects of system quality

How were test design and evidence evaluated?

- Automated evaluation with LLM-as-a-judge and results validated by human expert to ensure alignment and reliability

Challenges

01 **Limited legal understanding in general purpose pipelines** – General-purpose synthetic tools struggle to capture nuanced legal concepts (e.g., doctrinal links and citation importance), though they still provide fast and scalable testing value

02 **High context sensitivity in legal evaluation** – Small changes in legal question framing can significantly alter the correct answer, making it hard to define a single ground-truth reference for evaluation

Insights

01 General purpose test pipelines cannot yet fully handle the nuances of highly specialised domains. However, their speed and quality out of the box still makes them useful

02 Precision of context is critical in legal evaluation so datasets must be carefully designed to ensure a single clear reference answer

03 RAG evaluation must be multi-dimensional – Effective testing requires assessing correctness, faithfulness, and citation grounding, with faithfulness being especially important in legal use cases

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

LawNet AI: AI-Powered Legal Search

The Singapore Academy of Law (SAL) is a statutory body mandated to develop and promote initiatives that enhance the capability, efficiency, and advancement of Singapore's legal industry.

Use Case

LawNet AI, is a Question-and-Answer (Q&A) system designed to support legal research. It enables lawyers to ask legal research questions via SAL's legal research platform, LawNet, and receive AI-generated answers that are grounded in authoritative legal sources.

The primary objective of LawNet AI is to improve the efficiency and quality of legal research by providing concise, contextually relevant answers supported by citations of Singapore case law and legislation.

High-Level Architecture

LawNet AI is implemented using a single-turn Retrieval Augmented Generation (RAG) architecture. The system comprises:

- A legal fine-tuned retrieval mechanism optimised for Singapore case law and legislation; and
- A fine-tuned Large Language Model (LLM) that generates the final answer

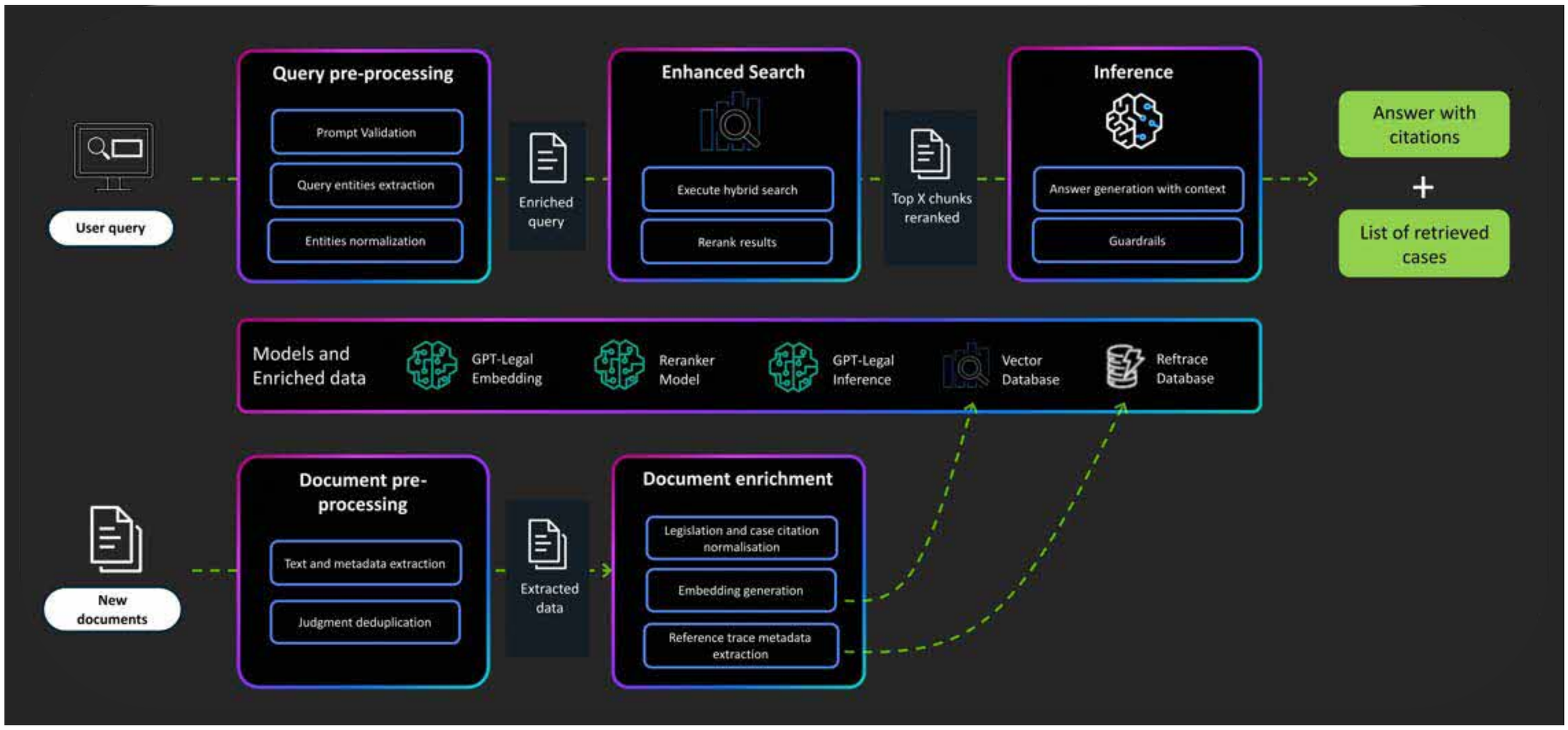


Figure 1: LawNet AI workflow

01 Deployer and Application

The LawNet AI workflow can be broadly divided into four stages:

Step 1 Document Ingestion and Preparation - New legal materials such as judgments and legislation are collected and processed. The system extracts text and metadata, removes duplicate versions of cases (as the same case can have different citations across their lifecycle, from unreported judgment to Singapore Law Report), and standardises legal citations and references.

Step 2 Enrichment and Storage - The processed legal materials are enriched using AI models that generate embeddings and reference trace metadata. The information is then stored in vector and reference databases to support efficient legal search and citation tracking.

Step 3 Query Understanding and Retrieval - When a user submits a legal question, the system validates and enriches the query pre-processing — prompt validation, entity extraction, and entity normalisation. This produces an enriched query that captures the legal intent.

Step 4 Answer Generation and Response. The retrieved legal sources are passed to the LLM, which generates a concise answer grounded in the retrieved materials. Guardrails are applied before the system returns the final answer together with supporting citations and retrieved cases.

02 Testing Partner and Testing Approach

resaro

Resaro offers independent, third-party assurance of mission-critical AI systems with extensive experience in testing Computer Vision and Generative AI systems.

Testing Approach

The engagement aimed to provide an independent assessment of the LawNet AI using a structured and repeatable testing methodology through Resaro's Approved Intelligence Platform (AIP).

- The AIP contains test plans for various types of AI systems - for this use case, Resaro used the RAG test plan.
- A test plan consists of synthetic data generators which allow generation of use case-appropriate synthetic data to augment the small human-generated dataset.
- It also consists of carefully orchestrated automated evaluation pipelines which use LLMs as judges to evaluate AI system responses against pre-defined criteria.

Risk Assessment

SAL prioritised the following risks to be tested:

- **Correctness of answers** (including faithfulness and lack of completeness)

LawNet AI could generate incorrect, misleading, or incomplete legal answers that omit important legal qualifications or fabricate unsupported claims, which could negatively affect lawyers' research quality, reduce trust in the platform, and potentially impact legal decision-making.

Why this matters: Inaccurate responses could negatively affect lawyers' research quality, reduce trust in the platform, and potentially mislead or impact legal decision-making.

- **Data Leakage (specifically system prompt leakage)**

SAL wanted to ensure that the system would not inadvertently disclose confidential, sensitive, or internal information such as system prompts.

Why this matters: Any leakage could create legal, operational, and reputational risks for both the lawyers and their organisations.

- **Reliability** (consistency of responses for identical prompts) and **Robustness to Perturbations**

SAL needed the system to provide stable and consistent legal answers to the same queries, as well as to meaning-preserving perturbations of queries (typos, synonyms, filler words, paraphrasing, and tones).

Why these matter: This was important because inconsistent answers could confuse users and reduce confidence in LawNet AI as a reliable legal research tool for the Singapore ecosystem.

SAL also conducted a broader risk assessment covering additional areas that were considered secondary to the main evaluation focus or had already been tested separately. This included efficiency, safety, security, robustness to out-of-domain queries, fairness, privacy, and other aspects of accuracy such as retrieval quality, instruction following, and helpfulness. Together, these indicators helped provide a more holistic understanding of the system's operational reliability, user safety, security posture, and overall quality as a legal research assistant.

Testing Scope and Test Design

Scope of Testing

The testing scope was designed around the 4 identified risks listed in the previous section. Given the high-stakes nature of legal work, Resaro's evaluation focused on whether the system could generate trustworthy, grounded, and consistent legal responses while remaining resilient against unintended disclosures. From the AIP's RAG test plan, the following quality indicators were selected by SAL as being relevant to their system:

- The AIP contains test plans for various types of AI systems - for this use case, Resaro used the RAG test plan.
- A test plan consists of synthetic data generators which allow generation of use case-appropriate synthetic data to augment the small human-generated dataset.
- It also consists of carefully orchestrated automated evaluation pipelines which use LLMs as judges to evaluate AI system responses against pre-defined criteria.

Faithfulness and Inaccuracy

The test focused on whether LawNet AI could generate accurate and complete legal answers grounded in authoritative legal sources (i.e., faithful to the provided sources). Two key metrics were prioritised:

- **Correctness**: Share of relevant claims in the reference answer found in the system's response
- **Faithfulness**: Share of relevant claims in the system's response which were supported by the retrieved context

Data leakage (specifically system prompt leakage)

The test focused on assessing if the system could resist attempts to reveal confidential or internal information, such as hidden system prompts or backend instructions. This was measured by its propensity to produce information about the system prompt in response to elicitation attempts, measured by share of responses to system prompt elicitation attempts which are unsafe.

Reliability and Robustness to Perturbations (consistency of responses to identical prompts, and meaning-preserving perturbations of prompts)

The testing assessed whether the system produced consistent legal answers when the same question was asked repeatedly, or to meaning-preserving modifications of questions such as typos, synonyms, filler words, paraphrasing, or tones. This included evaluating consistency of legal conclusions, stability of factual information, and variation across repeated responses.

Test Implementation

Data Used in Testing

01. Synthetic data generation

- SAL provided ~140,000 context paragraphs from legal journals which was used by Resaro to generate a synthetic dataset of ~200 Question-Answer-Context triplets, simulating realistic legal research scenarios.
- Resaro also generated:
  - Robustness dataset containing perturbations of the base questions
  - System prompt leakage dataset designed to test whether the AI would reveal internal prompts or hidden instructions when prompted through seemingly legitimate legal queries
- SAL also reviewed and provided feedback on the synthetic datasets to help assess their legal accuracy, relevance, and coverage.

02. Human dataset

In addition to testing using the synthetic data, Resaro also conducted testing on approximately 50 human-generated Question-Answer-Context triplets provided by SAL.

Execution of Tests

Resaro executed the tests using its Approved Intelligence Platform, which automated the testing workflows, synthetic data pipelines, and LLM-as-a-judge evaluation processes across the selected risk indicators.

Validation of Results

SAL reviewed and validated selected test outputs and evaluation results to verify that the automated judge pipelines aligned with human legal expert judgement. This validation step helped strengthen confidence in both the testing methodology and the reliability of the evaluation results.

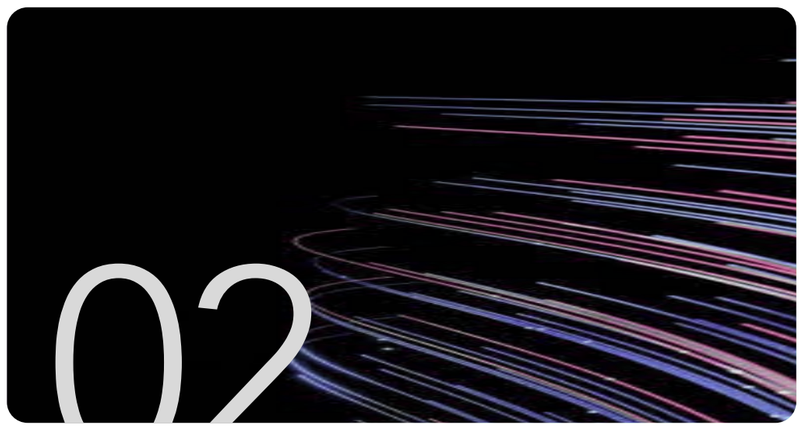
Cost of Testing

- SAL spent approximately 2 weeks preparing the initial data package (data for synthetic generation/ evaluation dataset), troubleshooting API access and reviewing the synthetic datasets and results.
- Resaro spent 3 weeks to onboard the client (data cleaning and API access), generation and post-validation improvement of the synthetic dataset, and running and post-validation improvement of the testing.

Insights/Lessons Learned

Limits of general-purpose synthetic data in specialised legal domains

General-purpose synthetic data generation and testing pipelines cannot yet fully handle highly specialised domains such as legal research. For example, the pipelines were unable to understand that “the second limb of the rule in Hadley v Baxendale” is a similar concept to “unforeseen losses”, or that information provided in a reference answer about where certain cases have been cited is important to prove legal authority.



Value of synthetic testing despite domain limitations

They offer important practical advantages as they enable rapid scaling of test cases, reduce dependency on expensive expert annotation, and have been battle tested in multiple evaluations. In this case, they also provided useful directional feedback on system performance trends, even if they did not fully capture all legal nuances.

Importance of precise context in evaluation dataset design

Whether using human-annotated or synthetic data, it is critical to provide sufficient context in each question so that there is a clear, single reference answer for evaluation. In legal applications, even small changes in factual detail or framing can significantly alter legal interpretation and expected responses, making ambiguity a key risk for unreliable evaluation outcomes.

RAG evaluation must be multi-dimensional and grounded in faithfulness

Effective testing of RAG systems requires a comprehensive evaluation framework covering multiple dimensions, including correctness, faithfulness to source material, citation accuracy, and completeness. Among these, faithfulness is particularly critical in legal use cases, as it ensures that generated responses remain firmly grounded in authoritative legal sources rather than relying on plausible but unsupported reasoning.

SAL was very much assured that the evaluation showed that the Q&A system has a high faithfulness score suggesting that the generated answer is well grounded and supported by retrieved context. SAL will continue to monitor and improve this key metric to ensure groundedness of the generated AI response to the retrieved case law and legislation.

AI VERIFY FOUNDATION