National College of Ireland

A RAG Framework for Legal Text Retrieval and Generation in Irish Statutory Instruments

MSc Research Project
MSc in Artificial Intelligence

Hassan Ozair
Student ID: 23410663

School of Computing
National College of Ireland

Supervisor: Ade Fajemisin / Paul Stynes

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Ozair Hassan
Student ID:	23410663
Programme:	MSc Artificial Intelligence **Year:** 2025
Module:	MSc (Research) Practicum Part 2
Supervisor:	Ade Fajemisin / Paul Stynes
Submission Due Date:	11th December 2025
Project Title:	A RAG Framework for Legal Text Retrieval and Generation in Irish Statutory Instruments
Word Count:	7519 **Page Count:** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ozair Hassan

Date: 10/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A RAG Framework for Legal Text Retrieval and Generation in Irish Statutory Instruments

Hassan Ozair
23410663

Abstract

Retrieval Augmented Generation (RAG) systems are the backbone of grounding responses from large language models in external knowledge, but ensuring accuracy and efficiency in certain specialized domains remains an ongoing challenge. This research proposes a comparative framework for evaluating the three main RAG approaches those being BM25, vector based and hybrid (BM25 + vector). A custom dataset was prepared from the Irish Statutory Instruments (2020-2025) by extracting the text from the PDFs, applying section-based chunking and encoding the chunks using BAAI/bge-large-en-v1.5. A gold standard evaluation set of 240 question-answer pairs was used for benchmarking. Weaviate database was utilized in the implementation of BM25, vector and hybrid retrievals. Evaluation was primarily focused on the following metrics, Mean Average Precision, recall, Average Entailment Score, Average Contradiction Score, Average Obligation Coverage Score and Average Final Composite Score. Results indicate that the Weaviate based hybrid system with re ranking had a performance increase in retrieval precision of more than 20% over the baseline simple hybrid system. Overall hybrid re rank performed best with 0.702 recall and highest final composite score of 0.35. This research aims to contribute to the practical guidelines for the designing and development of reliable and efficient RAG systems in the legal domain benefitting legal practitioners and policy makers with trustworthy AI tools.

Keywords: Retrieval Augmented Generation, BM25, Vector, Legal document retrieval

1 Introduction

With the advancement of technology and the growth of the internet, there is naturally a rapid expansion of digital information everything from personal information to legal information is readily available at a buttons click. This is had a fundamental change in the root of how legislation is now accessed, interpreted and applied. In this digital age more and more legal practitioners and policy makers have become increasingly dependent on these computational aids and tools to assist in navigation of large volumes of regulatory text. This highlights the fact that the conventional keyword driven searching techniques are now inadequate and struggle to properly encompass the complexity of the documents, ultimately failing to capture the cross-referencing patterns and semantic nuances ultimately making keyword driven techniques obsolete on their lonesome. In the Irish Statutory Instruments (SIs) it can be observed that obligations and conditions are distributed across vast hierarchical sections and complex inter dependent provisions.

This is where Retrieval Augmented Generation (RAG) systems come into play. RAG is the latest development to address these challenges. It leverages the superior generative capabilities of large language models and combines it with complex and advanced retrieval techniques. Lexical ranking such as BM25 are essentially the core retrieval techniques for the foundations of RAG systems (Robertson & Zaragoza., 2009), while more modern techniques from the last decade use semantic vector retrieval approaches that rely on the use of embedding models Levy et al.(2015), and from the most recent years we have the hybrid approach which is a carefully curated approach that utilizes the semantic vector dense score while gracefully balancing the lexical precision using BM25 (Erazímová, 2025; Sawarkar et al. 2024). Benchmark studies especially on large specialized legal datasets have highlighted the importance of accurate retrieval which in turn ensures a reduction in hallucinations and improves factual grounding. (Dahl et al., 2024) Hallucinations in Large Language Models Pipitone & Alami, (2024). Despite all this in depth study there are very few covering the performance of these three staple retrieval strategies when transferred to a new legal domain especially outside of the large regulatory jurisdictions. Accordingly, this study addresses the following research question:

How do BM25, vector and hybrid (BM25 + vector) based Retrieval-Augmented Generation systems compare in terms of retrieval accuracy (recall, mean average precision), generative quality, faithfulness, and citation accuracy within Irish Statutory Instruments (2020–2025)?

To answer this question the primary focus of this research is the development and evaluation of the three staple retrieval strategies, BM25, Vector based and Hybrid (BM25 + Vector) and implement these with the support structure of Weaviate vector database. The Irish SIs dataset will serve as the foundation for testing the three retrieval techniques subject to the same prepressing rules and evaluation queries. A fair comparison will be conducted using a gold standard set of 240 question and answer pairs that was manually created to ensure highest quality and accuracy. Due to the scale of the statutory corpus this Q&A dataset represents approximately 5% of the total material. However, it provides sufficient coverage to allow for a meaningful evaluation across the retrieval strategies.

Evaluation metrics will be aligned with the latest in legal RAG research and will be focused on Mean Average Precision (MAP), Recall along with further evaluation performed using RePASs and using the generated Average Entailment Score, Average Contradiction Score, Average Obligation Coverage Score and an overall Final Composite Score. (Rayo et al. 2025)

This paper is structured as follows. Section 2 is a comprehensive literature review highlighting the details of the state of the art focusing on the work of Rayo et al. (2025). Section 3 covers the research methodology. Section 4 covers the design specifications from preparing the dataset to evaluation design. Section 5 presents the results of experiments conducted. Section 6 discusses the implications of these findings and finally section 7 concludes the research with a discussion of the limitations faced and the directions for future research.

2 Related Work

With the recent advancements in the artificial intelligence technology there have been equally large advances in the research on how to better support the large language models (LLMs) this where the retrieval augmented generation (RAG) has come to play as a very important aspect and invaluable asset. In certain use cases such as the legal domain, this is mostly driven by advances in semantic encoders and it is the key to solve one of the biggest issues that being the falsification of data and hallucinations in the LLMs generated content and response. Ash and Chen (2019) have shown in their research that legal domain texts can be quite effectively and easily represented in high dimensional vector embeddings thus allowing the enabling of similarity based retrieval techniques. Their research revealed that legal corpora distribute spatially in a easily predictable manner, although in their approach their non transformer model struggled with clause level nuances and hierarchical structures of typical legal documents.

With this growing recognition that RAG systems can be used to supplement existing LLMs knowledge base and greatly enhance and improve the quality and accuracy of the response has motivated further benchmarking and studies. Chen et al. (2024) showed in their work that retrieval precision is more important in comparison to model size, although their work mainly focuses on general QA instead of structured legal text. However, even with the advantages of semantic similarity searches classic methods such as lexical retrieval still serve a strong foundational role as seen the work of Robertson and Zaragoza (2009) articulating the robustness and longevity of the BM25 search technique. Recent sparsity learning approaches such as SPLADE (Formal et al.,2021) and SpaDE (Choi et al., 2022) have introduced learned lexical expansions that boost recall. While these studies are a testament to the longevity of BM25 it is however limited by the fact that these models that relies on lexical search need specific domain training data and have a tendency to degrade in performance when used in smaller datasets.

This realization that neither of these system is perfect on its own and a balance is needed between generalisability and specificity has led to the rise of hybrid retrieval. Sawarkar et al. (2024) showed in their work that a blend of BM25 with semantic scores improved the robustness of the RAG system in ambiguous queries while Erazímová (2025) reported consistent gains for the upcoming next generation of hybrid scoring functions still utilizing BM25. Still, both reports came to the same conclusion that even though hybrid gains are quite significant it is unfortunately severely limited due to its dependency on the corpus used and the performance can is quite significantly degraded if transferred and implemented without adaption. Further ongoing studies into the complex domain of distributional similarity Levy et al. (2015) have complicated the landscape more and have led to new models such as the BMX model introduced by Li et al. (2024) which uses entropy weighted semantics to counter balance the problem of dense score saturation and this work is further supported by the work of Lu, Chen and Eickhoff (2025), who prove that cross encoder models are capable of implicitly learning semantic variants of BM25. While these approaches perform brilliantly and accurately on specially curated and structured datasets they still lack the validation on real world messy legal corpora with unpredictable formatting and deep cross referencing.

Legal specific RAG based benchmarks and hallucination sturdies further reinforce the importance of high quality retrieval. LegalBench-RAG (Pipitone & Alami, 2024) showed the

severe effects of even minor retrieval errors on the amplification of LLM hallucinations. Dahl et al. (2024) and later Latif, Y.A., (2025) both emphasize that even minor misstatements in LLM can drastically increase the chance of incorrect material and wrong legal conclusions further heightening the need for strong reliable truth grounding. However, again as see with previous studies these also focus primarily on large regulatory systems such as the US or EU but do not tackle the problem of smaller jurisdictions such as Ireland which end up under represented. Another recent development in RAG systems are Graph based retrieval systems while He et al. (2024) show that graph enhanced RAG improves interconnectivity and multi hop reasoning while the work of Peng et al. (2025) argue that they are only suited for domains with high cross section dependencies such as the legal domain. In the midst of this Raja and Vats (2025) further prove that semantic compression and graph augmentation dramatically improves the vector search efficiency. However, all agree that graph-RAG methods are extremely dependent on proper annotated relation node structures or citation networks which in the context of this research do not at the moment exist for the Irish SIs and would require significant manual preprocessing effort beyond the scope of this study.

While the core of RAG system design lies in the various search techniques there is also another additional line of research that focuses on multistage and re ranking retrieval pipelines. Xu et al. (2023) show in their work that dense vector-based retrieval systems paired with cross encoder reranking consistently outperform single stage simple retrieval pipelines, a similar pattern is observed in the work of Stryszewski et al. (2025), Fathallah et al. (2025), and Sager et al. (2025). However as with RAG systems themselves the effectiveness of these cross encoders again depends on the fine tuning and quality of the legal corpora used. To work around this both Pan et al. (2025) and Salemi and Zamani (2024) argue that retrieval and LLM based semantic query expansion are frequently conflated in evaluation frameworks which leads to obscurity and uncertainty about which component is responsible for downstream errors.

Overall, this literature review revealed several gaps, firstly although the strength of hybrid and dense retrieval systems have shown strong performance, they are still limited in their effectiveness because of being highly domain sensitive and remain untested on smaller legal corpora such as the Irish SIs. Secondly no existing works provide a direct controlled comparison of BM25, vector and hybrid scoring within the contexts of Irish legal setting or evaluate downstream generative consistency in such a domain. Finally, despite extensive concern for hallucinations in LLMs the specific link and showcase of a properly designed and developed system purely focusing on utilizing modern vector databases with support for hybrid retrieval has not been explored.

2.1 State of the art

As the research and development of RAG systems continues classic methods such lexical BM25 still manage to stay relevant in the modern world however their limitations show especially when dealing with the complex domain of legal texts. In the last decade vector-based approaches have steadily increased in performance and use however while they improve semantic matching, they lack the precision needed in legal retrieval. Leading to the new

reliance and further study into hybrid retrieval systems that are can be finely tuned and be a balanced combination of both lexical and the semantic methods.

A standout example of a modern hybrid framework designed for regulatory retrieval is introduced by Rayo, de La Rosa and Garrido (2025). Their system is capable of locating relevant information within a complex and large corpus of regulatory documents. They have achieved this using a combination of BM25 lexical with a fine-tuned SentenceTransformer model a subset of BAAI/bge-small-en-v1.5 and training it on the ObliQA dataset of 27,869 regulatory questions. By balancing this blend of dense semantic similarity with the ability of the BM25 to detect significant lexical cues their hybrid approach consistently outperformed all the standalone retrieval methods with performance metrics measured using Recall@10 and Recall@20 along with Map@10 and Map@20.

They have strengthened their system by carefully tailoring and designing their preprocessing pipeline and fine tuning it to support the regulatory text of the ObliQA dataset. Techniques such as contraction handling, text normalization, stopword removal and tokenisation enhance the performance of the lexical retrieval strategies while MultipleNegativesRankingLoss is implemented to optimise the semantic alignment. Finally using a weighted scoring scheme ($\alpha = 0.65$) helps to ensure that the final score is not lopsided to either lexical or semantic but balanced. Rayo et al.(2025) Go further by then passing their retrieval data to various LLMs. In this case the paid GPT-3.5-Turbo achieved the highest RePASs score (0.57) while the free llama 3 was still able to achieve a score of (0.37)

However, while the strengths of the Rayo et al. (2025) Framework are clear there are also several limitations that remain, particularly concerning generalisability and adaptability. Their approach is entirely focused and dedicated to the ObliQA dataset and does not focus and attempt to capture the linguistic variability and structural convections such as found in Irish SIs. Their system relies on a static preprocessing and monolithic fine-tuned model leaving open questions as its performance on corpuses that depend heavily on section identifiers or jurisdiction specific terminology. These gaps motivate the need to test their system in new environments and compare them against the simpler and easier approach of utilizing the benefits of vector databases-based retrieval systems. This is the primary motivation of the present research, which translates Rayo et al.'s baseline to the Irish SIs corpus and systemically evaluates the performance of BM25, vector and Hybrid retrieval implemented using Weaviate under controlled and comparable conditions.

3 Research Methodology

The methodology for this research is structured into two distinct key phases. (i) Establishing the state-of-the-art baseline by making minor adjustments to the work of Rayo et al (2025) and (ii) development of a original retrieval framework built utilizing Weaviate. This section will be broken down into five subsections each covering and expanding on the adopted methodology from data collection to evaluation.

3.1 Data collection and preprocessing

The data chosen for this research is entirely available online at the Irish Statute Book website, all relevant files required are available in PDF format for download. Step one of the research was to write a simple web scraping script that gathers the complete available files from the official Irish Statutory Instruments webpage (as of September 2025) from 2020 to 2025 amounting to approximately 4000 files in total. The script simply iterates through the various download pages and stores each PDF file in a predefined location.

The next step of the process was to perform the required preprocessing on the downloaded Irish SIs PDF files to extract the text, normalize the text and break into segments divided based on section headers such as reg-1, Explanatory Notes, Schedules etc. A simple preprocess script was developed in Python with the main goal to clean up text by removal of whitespaces and Unicode characters followed by lowercasing and symbol filtering and finished with selective retention of legally meaningful terms such as shall, may and must. Each passage of the file was then extracted based on its section header and assigned a unique identifier with its text being stored in both raw and normalised form. These passages were then saved into JSONL files with one JSONL file containing the contents of one PDF file.

3.2 Gold Standard Question & Answer Dataset

Following the ObliQA dataset pattern from Rayo et al's work a 240 question and answer dataset was manually created for evaluation. The goal was to have questions that are as close to mimicking real legal queries and covering a wide range of the material from the Irish SIs dataset. Each answer is mapped directly to one or more of the segmented passages from the Irish SIs dataset created. While structurally it is similar to the ObliQA dataset content wise it is tailored for the Irish legislations. Figure 3.1 is a sample from the final question & answer file prepared.


```
{
  "QuestionID": "27fb0e3b-cb35-46dc-bbb0-507adf02a3b8",
  "Question": "What legal change did these Regulations introduce regarding the duration of the Covid-19
  restrictions under the Health Act 1947?",
  "Passages": [
    {
      "DocumentID": "si-2020-0128",
      "PassageID": "reg-2",
      "Passage": "2. Regulation 2 of the Health Act 1947 (Section 31A – Temporary Restrictions) (Covid-19)
      Regulations 2020 (S.I. No. 121 of 2020) is amended by the substitution of "the 5th day of May 2020" for "the
      12th day of April 2020"."
    }
  ],
  "Group": 1
},
```


Figure 3.1: Sample from prepared question and answer test dataset

3.3 Retrieval System Implementations

After the completion of the preprocessing and preparation of the test Q&A dataset the next step was the implementation of the RAG systems.

First a baseline was established using the work of Rayo et al. (2025) where there was minimal change done to support the Irish SIs dataset and a slight change in LLM used as the free Groq API was utilized. Finally, a few additional tweaks were made to print presentable graphs to represent the results.

The next step was the primary work, that being the development of the Weaviate based RAG first step was to setup a local instance of the Weaviate vector database using Docker. Next setting up the ingestion layer and finally setting up the retrieval layer for each type of RAG and its relevant evaluation layer.

3.4 Evaluation Methodology

In order to maintain the integrity of the experiments the evaluation metrics were kept identical between both the baseline and the Weaviate implementation. Both pipelines relied on the retrieval of the top-k passages (k=10 and k=20) for every test set. These were measured primarily using: Recall and Mean Average Precision (MAP). While these were sufficient for testing the RAG systems in order to evaluate the legal reasoning the results from the Groq API llama-3 were assessed using RePASs which provided insights into the entailment score, contradictions core and obligation coverage scores allowing for a more in depth analysis and comparison of the baseline against the Weaviate RAG. The primary reasoning behind using the Llama-3 model from Groq API in comparison to a paid version such as GPT-3.5-Turbo is to ensure complete reproducibility without need for anyone to pay or spend funds on any LLM model.

4 Design Specification

The proposed system is designed to follow a modular RAG architecture. A high-level overview of the system shows it is composed of 6 main parts: Preprocessing, Data Ingestion, Retrieval methods (BM25, Vector, Hybrid), Re ranking, LLM answer generator and RePASs. These parts are broken down into 4 main layers: Knowledge layer, Retrieval layer, Re ranking layer and the Answer generation layer

4.1 Why use a Vector Database

While dealing with small sized datasets or simpler texts then using file-based corpora or in memory embeddings are sufficient however when dealing with legal corpora which are always large in size and extremely complex in their structure and linkage it is extremely inefficient to use these simple techniques. This is where vector databases come into play a few of the reasons why it is better to utilize vector databases are discussed below:

Primary reason is the unified management of text and embeddings. Since legal legislation and passages require both lexical and semantic retrieval to get highest performance a vector database has the ability to store both the raw and normalized text, the metadata associated with the file and the generated dense vector embeddings all in one single place this allows for a single source of truth rather than multiple disconnected systems.

Vector databases are designed and explicitly optimized for similarity search operations allowing them to store and index dense embeddings enabling fast nearest neighbour search even across tens of thousands of passages. Without the use of vector databases implementing hybrid BM25 + vector retrieval requires quite a lot more steps as showcased in Rayo et al's implementation whereas the use of a Vector database simplifies the boilerplate code and allows

for more focus on fine tuning the relevant factors which affect retrieval accuracy and performance.

4.2 Why Weaviate was selected

In recent years many vector databases have been developed each with their pros and cons. However, for this research Weaviate was chosen specifically for three main reasons. Firstly, it is fully open source this means Weaviate is designed to be easily run entirely locally on any system utilizing virtual machine tools such as Docker. This allows us to setup a local instance of Weaviate without having to pay for paid API services or cloud hosting. Playing an essential role in ensuring the reproducibility of this research. Second reason is the built in Hybrid Retrieval support Weaviate is one of the few vector databases that provides native support for combined lexical and vectoring scoring which allows for easier evaluation of its performance. Lastly it is developer friendly since Weaviate has a simple ready to use functions available through its official Python client that allows us to setup the ingestion pipeline, the retrieval pipelines with minimal configuration overhead.

4.3 High-Level System Design and Key Components

4.3.1 Knowledge Store Layer

The knowledge store layer as seen in Figure 4.1, can be considered the entry layer for the RAG pipeline. This is where the Irish SI PDFs are scraped and converted into their structured, segmented passages before being normalised annotated. The next step is to ingest these into Weaviate which automatically constructs the index for BM25, vector and hybrid. This serves as the single source of truth for the rest of the pipeline.

4.3.2 Retrieval Layer

The retrieval layer as seen in Figure 4.1, is the layer of the pipeline in which the three retrieval strategies are implemented. Each query from the Q&A dataset is transformed into its lexical and semantic representations. The cleaned text is used for the BM25 retrieval while the semantic embeddings are used for the dense vector search. Finally, a hybrid system is implemented utilizing both the lexical and semantic search techniques.

4.3.3 Re-Ranking Layer

The re ranking layer as seen in Figure 4.1, plays an important role in refining the results from the retrieval layer, utilizing a cross encoder re ranking model it re asses the relevance of each candidate passage cross checking against the context of the query. By re scoring the outputs of all three retrieval methods there is a consistent improvement in the results. The hybrid re ranked output is used as the primary source for the answer generation layer.

4.3.4 Answer Generation Layer

Finally, the answer generation layer as seen in Figure 4.1, completes the pipeline but taking the ranked passages from the re ranking layer and using those to construct the structured prompt. These are passed to the Llama-3 model via the Groq API. The generated responses are then evaluated using RePASs.

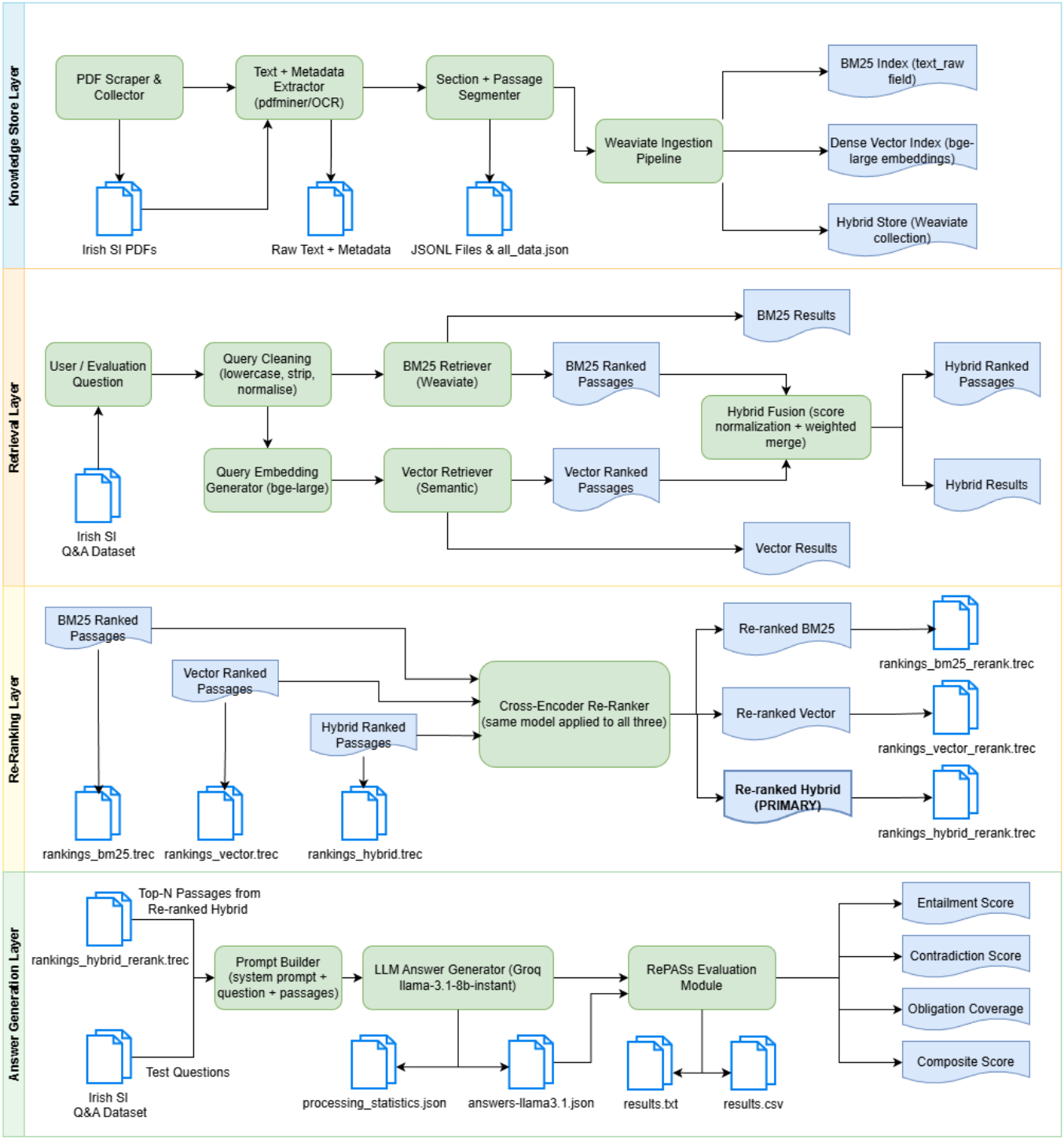

Figure 4.1: System Architecture

5 Implementation

In this section we will be discussing the final RAG system developed and used in this research. In this research several key outputs were produced at key stages, these range from transformed datasets, indexed vector collections, retrieval results, TREC-compatible ranking files etc. Here we will be outlining the tools and technologies used. This section is broken down into five subsections.

5.1 Data preparation and Transformation

The first step of the research involved the fetching and processing of the Irish Statutory Instruments from the last 5 years. The requirements for the final pipeline relied on a single consolidated dataset which would contain all the Irish SI documents broken down section wise. Additionally, a set of 240 manually prepared question and answer pairs was curated to serve as the evaluation benchmark. All passages and queries were normalised using the same techniques to ensure consistency across BM25, vector and hybrid retrieval modes.

5.2 Setup of Weaviate as the Retrieval Backend

The backbone of the proposed system relies on the Weaviate vector database. This serves as the single source of truth and unified database for all three systems: Lexical BM25 search, Dense vector similarity search and Hybrid retrieval search. Weaviate was setup locally using a simple docker-compose.yml file utilizing the “weaviate:1.32.9” image freely available from “semitechnologies” in the following stages of the pipeline the system would connect to the local docker hosting to utilize Weaviate.

5.3 Implementation of Retrieval Methods

In the final system we have three retrieval modes all utilizing Weaviate

5.3.1 Lexical BM25

For the lexical BM25 we use Weaviate’s BM25 index to match and return the passages that most closely matched the query. Figure 5.1 shows the ranked list of passages returned for a sample query. Each passage contains its relevant text and a composite ID which contains the Parent ID and Section ID. As seen in Figure 5.1 “si-2020-0025-reg-12" can be broken down into “si-2020=0025” as the Parent ID and “reg-12” as the Section ID


Query: 'If an eligible applicant later receives a payment outside this scheme for the sam

	ID	DocumentId	PassageId	text_preview
0	si-2020-0025-reg-12	si-2020-0025	reg-12	reg-12 12. (1) Where an application under Regu...
1	si-2020-0025-reg-11	si-2020-0025	reg-11	reg-11 11. (1) Without prejudice to the genera...
2	si-2020-0025-reg-9	si-2020-0025	reg-9	reg-9 9. (1) An applicant who has received a c...
3	si-2024-0103-reg-26	si-2024-0103	reg-26	reg-26 26. (1) Where an awarding authority mak...
4	si-2025-0070-reg-26	si-2025-0070	reg-26	reg-26 26. (1) Where the awarding authority ma...

Figure 5.1: BM25 retrieval sample

5.3.2 Dense Vector Retrieval

For the dense vector retrieval, we use Weaviate’s bult in “near_vector” method which computes to find the closest match using cosine similarity and returns the passages that most closely matched the query. As seen in Figure 5.2 the distance represents the mathematical

similarity between the user's query and the saved embeddings in Weaviate. The lower the distance the more accurate and semantically similar the results.

Query: 'If an eligible applicant later receives a payment outside this scheme for the same damage,

	ID	DocumentId	PassageId	distance	text_preview
0	si-2023-0347-reg-54	si-2023-0347	reg-54	0.243071	reg-54 54. Where I have received or receive in...
1	si-2023-0347-reg-53	si-2023-0347	reg-53	0.249695	reg-53 53. Where a payment otherwise than unde...
2	si-2023-0347-reg-41	si-2023-0347	reg-41	0.251202	reg-41 41. Where I have received or receive in...
3	si-2023-0347-reg-40	si-2023-0347	reg-40	0.253100	reg-40 40. Where a payment otherwise than unde...
4	si-2020-0025-reg-11	si-2020-0025	reg-11	0.260702	reg-11 11. (1) Without prejudice to the genera...

Figure 5.2: Vector retrieval sample

5.3.3 Hybrid Retrieval

For the Hybrid Retriever we use the combination of Weaviate's "hybrid" method which uses a balanced combination of Lexical BM25 and Dense vector retrieval. The balancing can be controlled by adjusting the "alpha - α" value. As seen in Figure 5.3 in hybrid retrieval a score is assigned for all embedding documents that are fetched. The closer the score is to 1 the more accurate the results that are fetched.

	rank	ID	DocumentId	PassageId	score	text_preview
0	1	si-2020-0025-reg-12	si-2020-0025	reg-12	0.899425	reg-12 12. (1) Where an application under Regu...
1	2	si-2020-0025-reg-11	si-2020-0025	reg-11	0.758766	reg-11 11. (1) Without prejudice to the genera...
2	3	si-2023-0347-reg-54	si-2023-0347	reg-54	0.660624	reg-54 54. Where I have received or receive in...
3	4	si-2023-0347-reg-41	si-2023-0347	reg-41	0.620439	reg-41 41. Where I have received or receive in...
4	5	si-2020-0025-reg-10	si-2020-0025	reg-10	0.506959	reg-10 10. (1) An eligible applicant who has r...
5	6	si-2023-0347-reg-53	si-2023-0347	reg-53	0.491673	reg-53 53. Where a payment otherwise than unde...
6	7	si-2023-0347-reg-40	si-2023-0347	reg-40	0.476325	reg-40 40. Where a payment otherwise than unde...
7	8	si-2024-0441-reg-13	si-2024-0441	reg-13	0.289555	reg-13 13. (1) If an accountable person, other...
8	9	si-2023-0406-reg-13	si-2023-0406	reg-13	0.289555	reg-13 13. (1) If an accountable person, other...
9	10	si-2020-0025-reg-9	si-2020-0025	reg-9	0.284690	reg-9 9. (1) An applicant who has received a c...

Figure 5.3: Hybrid (un ranked) retrieval sample

5.4 Re Ranking Using a Cross-Encoder

While the results were satisfactory and an improvement over traditional methods there was still a lacking and room for improvement. Here is where the CrossEncoder re ranking came into play. The role of this was to re-evaluate each candidate passage and recheck its relation to the query. The model chosen was 'BAAI/bge-reranker-large' this was applied to all three retrieval types and resulted in much improved outputs. Figure 5.4 shows the results of reranking.

Table 5.1: Result of CrossEncoder re-ranking

Re Ranked Performance Table (MAP + Recall)				
System	**MAP@10**	**MAP@20**	**Recall@10**	**Recall@20**
BM25	0.396	0.396	0.504	0.504
Vector	0.446	0.448	0.658	0.689
Hybrid	0.446	0.448	0.667	0.702

5.5 Integration with LLM Based Answer Generation

To further test and compare the results the next step was to pass the top ranked passages for each of the 240 evaluation questions to an LLM for testing the responses. The chosen LLM was the Llama 3 model using Groq API. The answers generated from the LLM were saved to a JSON file. This file was then passed to RePASs, an automated evaluation framework. RePASs computes key evaluation metrics that play an important role in legal and regulatory reasoning. These metrics are Entailment Score, Contradiction Score, Obligation Coverage Score and Composite Score. These metrics allow us to quantify the quality of each generated answer. For this stage the best performing retrieval system which was the hybrid system was selected. A complete evaluation report was generated for all the 240 LLM generated answers and per question metric values were saved in a csv file for further in-depth analysis.

6 Evaluation

In this section we will cover the results obtained from both the baseline system and the proposed newer Weaviate based system. The goal is to see how effective each system is in supporting retrieval augmented question answering for the Irish SI dataset.

6.1 Benchmark (Baseline)

The baseline system selected for this research was an adapted version of Rayo et al. (2025) in which they developed a pipeline for using BM25, dense vector retrieval and hybrid approach to test the ObliQA dataset. Their approach was slightly adjusted to work with the Irish SI dataset and was used as the benchmark for this study.

6.1.1 Baseline MAP Performance

Figure 6.1 shows the Mean Average Precision (MAP) values for all three retrieval systems. Some of the key observations are that the BM25 achieves higher MAP scores in comparison to the Vector retrieval system. BM25 achieves MAP@10 = 0.299 and MAP@20 = 0.303 in comparison to the vector systems MAP@10 = 0.253 and MAP @20 = 0.256. This suggest that the semantic drift is higher in the Irish SI dataset and that the lexical precision is slightly better. Finally, in can be seen that the best results are for the Hybrid system which achieves MAP@10 = 0.299 and MAP@20 = 0.304 which is a small but overall improvement over both systems independently.

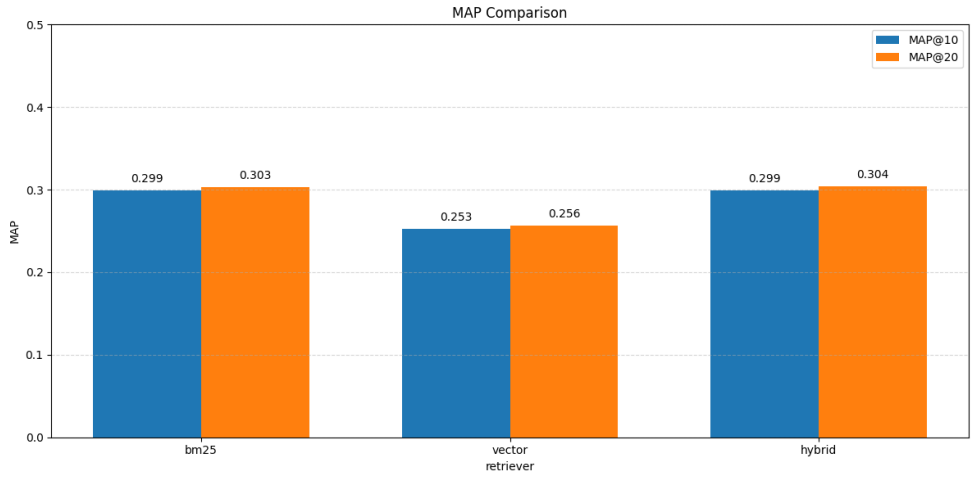

Figure 6.1: Baseline MAP results

6.1.2 Baseline Recall Performance

Figure 6.2 shows the Recall values for all three retrieval systems. Some of the key observations are that again the BM25 achieves higher Recall scores in comparison to the Vector retrieval system. BM25 achieves Recall@10 = 0.487 and Recall@20 = 0.550 whereas the vector retrieval system falls short again in performance with scores of Recall@10 = 0.446 and MAP @20 = 0.5. Finally, in can be seen that the best results are for the Hybrid system which achieves Recall@10 = 0.496 and Recall@20 = 0.562 which is a small but overall improvement over both systems independently.

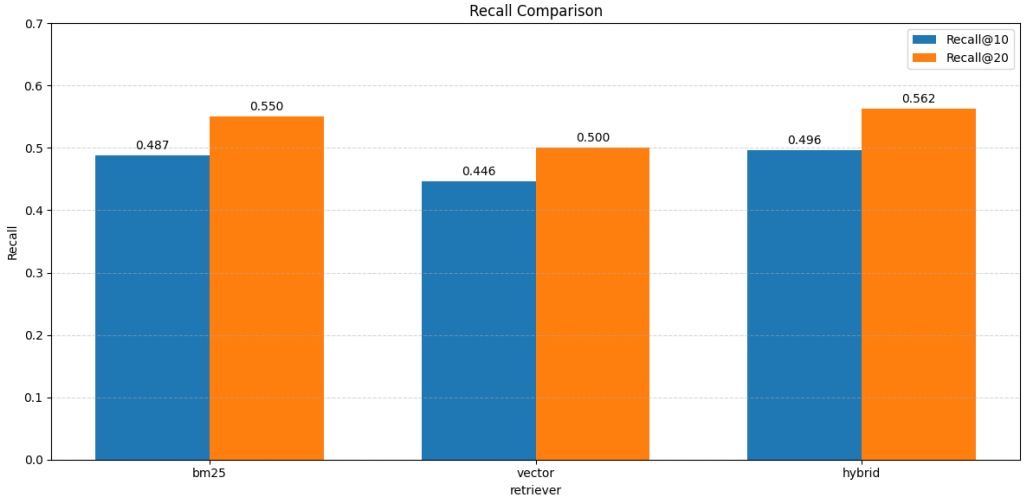

Figure 6.2: Baseline Recall results

6.2 Weaviate approach

In this section we will cover the results obtained from the proposed Weaviate based system. This will cover both the results obtained from the simple system and the results obtained from the cross encoder re ranking system.

6.2.1 Weaviate MAP Performance

Figure 6.3 shows the MAP values for all three retrieval systems comparing both the simple versions with the re ranked versions. As seen in the figure BM25 achieves similar results as the baseline with MAP@10 = 0.294 however the real improvement is seen in the vector retrieval with the MAP@10 = 0.375 compared to the MAP@10 = 0.253 of the baseline. This further highlights the importance of using a vector database specifically designed to handle large embeddings from stronger embedding models. Further in the hybrid retrieval we see a performance bump with MAP@10 = 0.357 as well.

While overall an improvement can be seen as compared to the baseline the biggest boost in results comes from the second stage re ranking. With the highest gains observed in the hybrid retrieval re ranked with MAP@10 = 0.446 and MAP@ 20 = 0.488 a significant improvement of more than 20% over the baseline performances.

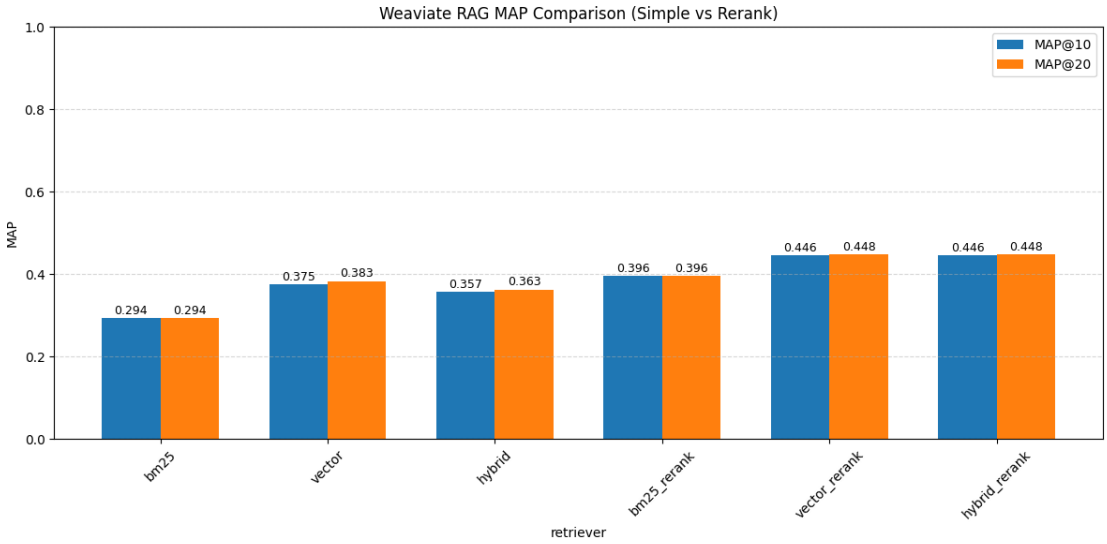

Figure 6.3: Weaviate RAG MAP results

6.2.2 Weaviate Recall Performance

Figure 6.4 shows the Recall values for all three retrieval systems comparing both the simple versions with the re ranked versions. As seen in the figure BM25 achieves similar results as the baseline with Recall@10 = 0.500 however the real improvement is seen in the vector retrieval with the baseline Recall@10 = 0.496 compared to the new Weaviate approach Recall@10 = 0.600. Further in the hybrid retrieval we see a performance bump with Recall@10 = 0.571 compared the baseline Recall@10 = 0.496.

While overall an improvement can be seen as compared to the baseline the biggest boost in results comes from the second stage re ranking. With the highest gains observed in the vector and hybrid retrieval systems with re ranking. The vector re-rank Recall@10 = 0.658 with the hybrid re-rank system performing even better with Recall@10 = 0.667 this represents a significant improvement of over 20%

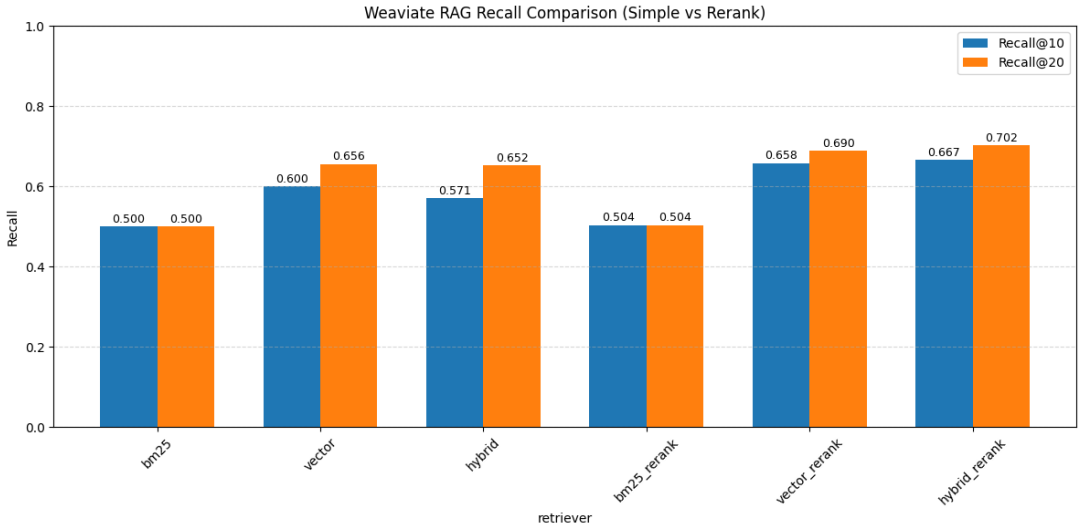

Figure 6.4: Weaviate RAG Recall results

6.3 Discussion

The results obtained in this research clearly show that the Weaviate based system significantly outperforms the baseline system. This due to the key combination of utilizing the modern vector database with a strong re ranking system this results in far better semantic matching with the balance in lexical matching than the original simpler hybrid retrieval system.

From an academic point of view these findings reinforce the and further the support the understanding that these dense retrieval methods are more efficient in the complex and semantically rich datasets. The improvements are consistent with what was expected from all the prior literature where it's been seen that the pure lexical approaches like BM25 struggle when direct simple word to word matching is not enough. Furthermore, the results also show that simply relying on dense retrieval is not enough either but instead a balanced hybrid approach is recommended with a strong re ranking stage as introduced in this study. This shows that simply increasing model size or pretraining efforts such as in model size or pretraining efforts such as in the work of Rayo et al (2025) alone are not enough to have a significant shift in performance.

From the viewpoint of a legal practitioner these improvements have a big impact on real world legal and compliance environments, organisations that rely on statutory interoperation could benefit from adapting hybrid retrieval search systems with second stage re ranking. The improvements observed show that with the Weaviate based system the chances of false positives and hallucinations is greatly reduced as the results fetched are more grounded and accurate greatly benefiting practitioners as accuracy is key in legal application.

6.3.1 Baseline vs Weaviate Retrieval on a Sample Query


```
{
  "QuestionID": "1d39b2f7-5c90-4457-8ce3-cac5d0317b4c",
  "Question": "According to the Explanatory Note, what is the substantive effect of S.I. No. 128 of 2020 on the operation period of the Covid-19 Temporary Restrictions Regulations?",
  "Passages": [
    {
      "DocumentID": "si-2020-0128",
      "PassageID": "explanatory-note",
      "Passage": "These Regulations amend the Health Act 1947 (Section 31A-Temporary Restrictions) (Covid-19) Regulations 2020 (S.I. No. 121 of 2020) to provide that the Regulations shall remain in operation until 5 May 2020."
    }
  ],
  "Group": 1
},
```


Figure 6.5: Sample Query

To show the practical impact of what we mean when showing the better MAP and Recall of the Weaviate based system, we choose a random question from the 240 samples we have. The chosen query comes from “si-2020-0128” and the section “explanatory-note” as seen in Figure 6.5. Yet when tested against the baseline hybrid system it fails to fetch the correct document and section completely even when tested in the top 20 results. Instead, it assigns the wrong document an incorrect score of “0.982755” which means that the system will be getting the incorrect document as seen in Figure 6.6. However, here is where the newer and improved Weaviate re ranked system shines as seen in Figure 6.7 it not only fetches the correct document and correct passage, it also shows how initially the score was 0.748943 and could have put this result lower down on the list however when passed through the re ranker it correctly increases its score to 0.949510 thus ensuring that indeed the correct document and passage is used. This is further supporting evidence of why the small 10-20% improvements in performance are so important and can shape and structure the entire efficiency and accuracy of the system.

Query: 'According to the Explanatory Note, what is the substantive effect of S.I. No. 128 of 2020 on the operation period of the Covid-19 Temporary Restrictio
Alpha: 0.5 (Balance between BM25 and Vector)

	rank	ID	DocumentId	PassageId	score	text_preview
0	1	si-2020-0327-reg-1	si-2020-0327	reg-1	0.982755	reg-1 1. These Regulations may be cited as the Health Act 1947 (Section 31A – \n(Revocation) \n\n(Relevant Counties) \n\n(Covid-19) \n\nTemporary Restrictions) \nRegulations 2020..
1	2	si-2022-0048-explanatory-note	si-2022-0048	explanatory-note	0.909668	explanatory-note EXPLANATORY NOTE \n\n(This note is not part of the Instrument and does not purport to be a legal \ninterpretation.) \n\nThese Regulations amend the Health Act 1947 (Section 31A – Temporary.
2	3	si-2021-0273-explanatory-note	si-2021-0273	explanatory-note	0.908407	explanatory-note EXPLANATORY NOTE \n\n(This note is not part of the Instrument and does not purport to be a legal \ninterpretation.) \n\nThese Regulations amend the Health Act 1947 (Section 31A – Temporary.
3	4	si-2020-0701-explanatory-note	si-2020-0701	explanatory-note	0.897534	explanatory-note EXPLANATORY NOTE \n\n(This note is not part of the Instrument and does not purport to be a legal \ninterpretation.) \n\nThese Regulations revoke the Health Act 1947 (Section 31A - Temporar.
4	5	si-2020-0121-reg-1	si-2020-0121	reg-1	0.889828	reg-1 1. (1) These Regulations may be cited as the Health Act 1947 (Section 31A \n\n-Temporary Restrictions) (Covid-19) Regulations 2020. \n\n(2) These Regulations shall come into operation on the 8th day.

Figure 6.6: Baseline Hybrid response for sample query

Query: 'According to the Explanatory Note, what is the substantive effect of S.I. No. 128 of 2020 on the operation period of the Covid-19 Temporary Restrictio
Alpha: 0.5 (Balance between BM25 and Vector)
Re-ranking with Cross-Encoder...

	rank	ID	DocumentId	PassageId	original_score	rerank_score	text_preview
0	1	si-2020-0128-explanatory-note	si-2020-0128	explanatory-note	0.748943	0.949510	explanatory-note EXPLANATORY NOTE \n\n(This note is not part of the Instrument and does not purport to be a legal \ninterpretation). \n\nThese Regulations amend the Health Act 1947 (Section 31A-Temporary \n...
1	2	si-2021-0273-explanatory-note	si-2021-0273	explanatory-note	0.718136	0.850439	explanatory-note EXPLANATORY NOTE \n\n(This note is not part of the Instrument and does not purport to be a legal \ninterpretation.) \n\nThese Regulations amend the Health Act 1947 (Section 31A – Temporary...
2	3	si-2022-0048-explanatory-note	si-2022-0048	explanatory-note	0.727485	0.842393	explanatory-note EXPLANATORY NOTE \n\n(This note is not part of the Instrument and does not purport to be a legal \ninterpretation.) \n\nThese Regulations amend the Health Act 1947 (Section 31A – Temporary...
3	4	si-2021-0513-explanatory-note	si-2021-0513	explanatory-note	0.699032	0.528600	explanatory-note EXPLANATORY NOTE \n\n(This note is not part of the Instrument and does not purport to be a legal \ninterpretation.) \n\nThese Regulations amend the Health Act 1947 (Section 31A - Temporary...
4	5	si-2021-0276-explanatory-note	si-2021-0276	explanatory-note	0.480185	0.334487	explanatory-note EXPLANATORY NOTE \n\n(This note is not part of the Instrument and does not purport to be a legal \ninterpretation.) \n\nThese Regulations extend until 19th July 2021 the effective date of ...

Figure 6.7: Weaviate Re Ranked Hybrid response for sample query

6.3.2 RePASs Evaluation

In addition to the evaluation of the retrieval systems the RePASs evaluation also provides enhanced insights into the performance of the systems and what the benefits of using the Weaviate based system are. Figures 6.8 and 6.9 present the contradiction scores and positive evaluation metrics respectively

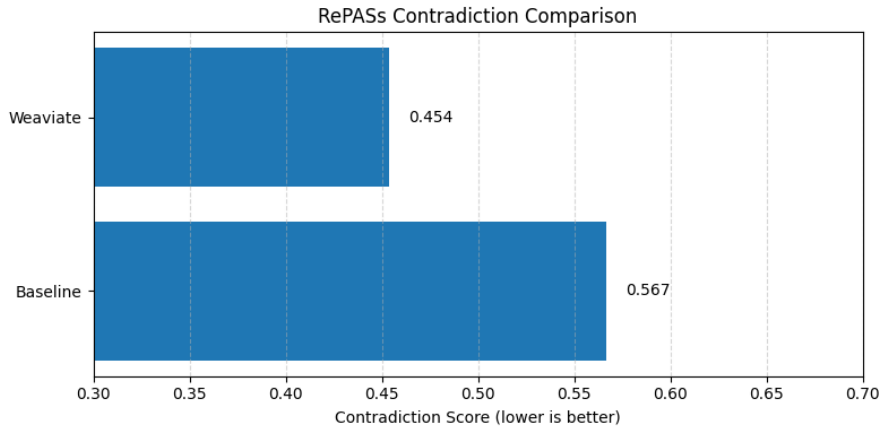

Figure 6.8: RePASs Contradiction Comparison

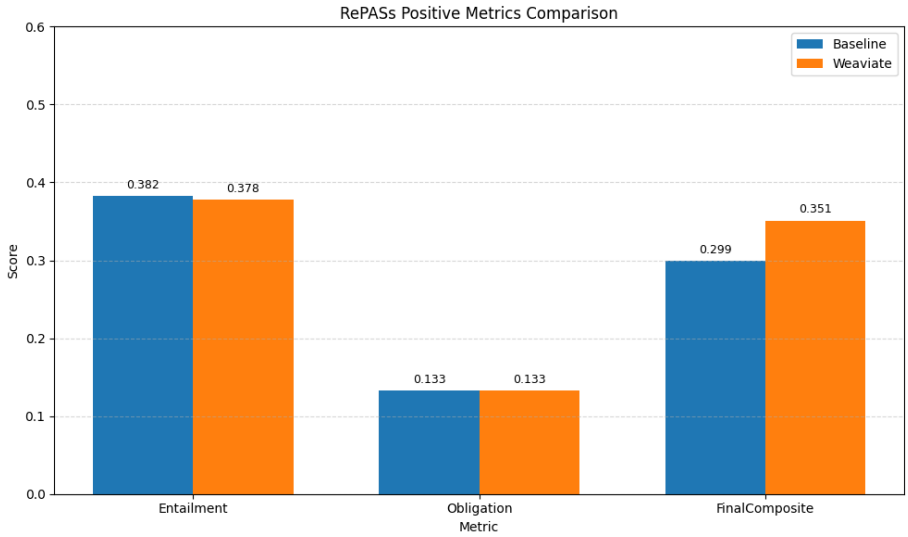

Figure 6.9: RePASs Metrics Comparison

The results of RePASs show several key trends. First and foremost, the most important improvement for the Weaviate based system is in the contradiction score, with a significant improvement of more than 10% as the results from the bassline were 0.567 whereas the results from the Weaviate based system are 0.454. This suggests that the LLM generated more consistent answers with higher quality and most importantly more relevant context passages. Secondly the final composite score saw an increase of 6% going from 0.299 to 0.351 for the Weaviate based approach, this also further showcases better grounding and coherence in the answers generated from the Weaviate based system.

An important point of note is that the entailment and obligation coverage scores remained similar between both systems, this was expected as both the systems relied on the same Irish SI corpus however the improved contradiction and composite scores clearly show that the Weaviate based system fetches more reliable results for the LLM thereby reducing the chances of hallucinations and inconsistences. Overall, the results show the importance of retrieval quality in RAG systems. The table below summarizes the findings of this study:

Table 6.1: Performance Table of Baseline vs Weaviate

Performance Table (MAP + Recall)				
System	**MAP@10**	**MAP@20**	**Recall@10**	**Recall@20**
Baseline Hybrid (Rayo et al.)	0.299	0.304	0.496	0.562
Weaviate Hybrid	0.357	0.363	0.571	0.652
Weaviate Hybrid + Re-ranking	0.446	0.488	0.667	0.702

Although developed specifically for this thesis, the resulting pipeline functions as a modular RAG framework: each layer from the Knowledge layer, Retrieval layer, Re ranking layer and the Answer generation layer is designed to be interchangeable and replicable across regulatory datasets beyond Irish Statutory Instruments.

7 Conclusion and Future Work

This research set out to determine whether a modern approach to RAG systems using a vector database driven retrieval architecture could have any sort of improvement in retrieval accuracy and answer generation for the Irish Statutory Instruments compared to using a state-of-the-art baseline system that does not utilize a vector database. More specifically, the aim of the research was to answer the research question of how BM25, vector, and hybrid retrieval strategies compare in terms of retrieval accuracy, precision, generative faithfulness, and citation quality within the Irish SI corpus. The plan was to build a structured Irish SI dataset and then recreate the baseline hybrid system to set a benchmark for the new Weaviate based retrieval framework. Evaluation of both systems was done across multiple metrics ranking from MAP and Recall to the results from the RePASs evaluation model.

The findings obtained in this study showed that the Weaviate based system delivered superior performance across all the selected evaluation metrics. The most significant improvement was in the dense vector retrieval system both the MAP@10 jumping from 0.299 to 0.357 and Recall@10 increasing from 0.446 to 0.600 showcasing the advantages of using a system based around a vector database. As the initial baseline system was primarily focused on hybrid approach a similar hybrid approach was adopted in the Weaviate based system again showing more than 10% improvement in results going from Recall@10 of 0.496 to 0.571. Finally, the addition of a strong re ranking stage boosted the results to a further 10% improvement resulting in the final proposed Weaviate based system achieving results with an improvement of over 20% as seen in the hybrid system increasing from Recall@10 = 0.496 to Recall@10 = 0.667

These results were further supported by the RePASs evaluation of the LLM answer quality. Showing a significant improvement in contradiction score for the Weaviate based system dropping from 0.567 to 0.454 an improvement of more than 10% in false positives. The final composite score increased from 0.299 to 0.351 showing better grounding and less hallucinations in the LLM answers.

Although the research produced strong results there are still several limitations, first and foremost the biggest limitation is the size of the Q&A testing dataset as this was manually curated it leaves a lot of room for further additions and enhancements. Secondly while the reranking model used showed a significant improvement in results and performance the model used however was a generic CrossEncoder rather than a fine-tuned model for legal use. Another limiting factor was the generic prompting strategy used for the LLM, while it was effective and served the purpose it still relied heavily on heuristics rather than using a formally trained legal

summarisation model. With these constraints in mind while the system is robust it is far from a representation of what can be achieved with more specialised modelling.

Future work can easily address these limitations by utilizing and developing the legal domain reranking model that is specific for usage in legal texts, more work can be done on expanding the Q&A dataset with more input from legal practitioners. Also, in this research the primary reasoning behind using the Llama-3 model from Groq API in comparison to a paid version such as GPT-3.5-Turbo was to ensure complete reproducibility without need for anyone to pay or spend funds on any LLM model however it's quite simple to swap in any LLM model as desired by the researcher. Additionally, there is a lot of room for building a proper deployable legal search assistant that can leverage the strengths of the implemented system and expand support for real time day to day legal research. An application of that scale will have clear benefits for both commercial and governmental usage especially when factoring in the low operating cost and strong performance from the Weaviate based system.

References

Ash, E. & Chen, D.L. (2019) Case Vectors: Spatial Representations of the Law Using Document Embeddings. Law as Data, 11.

Chen, J., Lin, H., Han, X. & Sun, L. (2024) 'Benchmarking Large Language Models in Retrieval-Augmented Generation', Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), pp. 17754–17762.

Choi, E., Lee, S., Choi, M., Ko, H., Song, Y.-I. & Lee, J. (2022) 'SpaDE: Improving Sparse Representations using a Dual Document Encoder for First-Stage Retrieval', Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22), pp. 272–282.

Dahl, M., Magesh, V., Suzgun, M. & Ho, D.E. (2024) 'Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models', Journal of Legal Analysis, 16(1), pp. 64–93.

Erazímová, L. (2025) BM42 vs. Conventional Methods: Evaluating Next-Generation Hybrid Search Techniques for Information Retrieval. Dissertation.

Formal, T., Piwowarski, B. & Clinchant, S. (2021) 'SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking', Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), pp. 2288–2292.

He, X. et al. (2024) G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering.

Latif, Y.A., (2025). Hallucinations in large language models and their influence on legal reasoning: Examining the risks of ai-generated factual inaccuracies in judicial processes. Journal of Computational Intelligence, Machine Reasoning, and Decision-Making, 10(2), pp.10-20.

Levy, O., Goldberg, Y. & Dagan, I. (2015) 'Improving Distributional Similarity with Lessons Learned from Word Embeddings', Transactions of the Association for Computational Linguistics, 3, pp. 211–225.

Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y. & Tang, S. (2025) 'Graph Retrieval-Augmented Generation: A Survey', ACM Transactions on Information Systems, Just Accepted (November 2025).

Pipitone, N. & Alami, G.H. (2024) LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain.

Rayo, J., de La Rosa, R. & Garrido, M. (2025) A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts.

Robertson, S. & Zaragoza, H. (2009) 'The Probabilistic Relevance Framework: BM25 and Beyond', Foundations and Trends in Information Retrieval, 3(4), pp. 333–389.

Salemi, A. & Zamani, H. (2024) 'Evaluating Retrieval Quality in Retrieval-Augmented Generation', Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), pp. 2395–2400.

Sawarkar, K., Mangal, A. & Solanki, S.R. (2024) 'Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers', 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 155–161.

Askari, A., Abolghasemi, A., Pasi, G., Kraaij, W. & Verberne, S. (2023) 'Injecting the BM25 score as text improves BERT-based re-rankers', in European Conference on Information Retrieval, March, pp. 66–83. Cham: Springer Nature Switzerland.

Li, X., Lipp, J., Shakir, A., Huang, R. & Li, J. (2024) Bmx: Entropy-weighted similarity and semantic-enhanced lexical search. arXiv preprint arXiv:2408.06643.

Lu, M., Chen, C. & Eickhoff, C. (2025) Cross-Encoder Rediscovers a Semantic Variant of BM25. arXiv preprint arXiv:2502.04645.

Raja, R. & Vats, A. (2025) 'Beyond Nearest Neighbors: Semantic Compression and Graph-Augmented Retrieval for Enhanced Vector Search', in Proceedings of the 3rd International Workshop on Deep Multimodal Generation and Retrieval, October, pp. 39–45.

Stryszewski, J., Prosowicz, W., Kawiak, T. & Jaśkowiec, A. (2025) AGH IR at LongEval: Improving Scientific Information Retrieval with Dense Representations and Cross-Encoder Re-ranking.

Fathallah, M., El-Makky, N.M. & Torki, M. (2025) 'AlexUNLP-FMT at ClimateCheck Shared Task: Hybrid Retrieval with Adaptive Similarity Graph-based Reranking for Climate-related Social Media Claims Fact Checking', in Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), July, pp. 288–292.

Sager, P.J., Kamaraj, A., Grewe, B.F. & Stadelmann, T. (2025) Deep Retrieval at CheckThat! 2025: Identifying Scientific Papers from Implicit Social Media Mentions via Hybrid Retrieval and Re-Ranking. arXiv preprint arXiv:2505.23250.

Pan, M., Xiong, W., Zhou, S., Gao, M. & Chen, J. (2025) 'LLM-Based Query Expansion with Gaussian Kernel Semantic Enhancement for Dense Retrieval', Electronics, 14(9), p. 1744.

Xu, G., Zhang, Y., Zhang, L., Long, D., Xie, P. & Guo, R. (2023) Hybrid Retrieval and Multi-stage Text Ranking Solution at TREC 2022 Deep Learning Track. arXiv preprint arXiv:2308.12039.

Saparov, A., Pang, R.Y., Padmakumar, V., Joshi, N., Kazemi, M., Kim, N. & He, H. (2023) 'Testing the general deductive reasoning capacity of large language models using OOD examples', Advances in Neural Information Processing Systems, 36, pp. 3083–3105.