Development and Evaluation of a Simple Embedding-Based System for Jurisprudence Retrieval in Financial Dispute Resolution Contexts

Nikki Bieleveldt
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
n.m.bieleveldt@student.utwente.nl

***Abstract*—Legal professionals struggle with inefficient jurisprudence retrieval using traditional keyword searches. Efficient retrieval of precedents in Dutch financial dispute resolution is critical for informed decision-making. This thesis investigates whether embedding-based retrieval systems can outperform existing keyword methods for Dutch financial legal documents. An embedding-based retrieval system was developed and evaluated on Dutch financial legal documents. Experiments on a curated test set demonstrate that the keyword-based retrieval methods achieve superior F_1 performance compared to purely semantic and hybrid methods, indicating that traditional approaches remain more effective than embedding-based methods for this specialized legal domain. The findings contradict broader research trends and highlight the need for context-specific evaluation when deploying AI systems in legal practice.**

***Index Terms*—Jurisprudence Retrieval, Legal Information Retrieval, Embedding Models, Dutch Financial Dispute Resolution, Vector Embeddings**

1. Introduction

Jurisprudence, the body of court rulings that functions as precedent, forms the bedrock of Dutch legal reasoning. In the Dutch financial dispute resolution system, the quality, fairness, and consistency of the rulings hinge on the ability to retrieve relevant precedents. As financial products grow more complex, jurisprudence becomes even more crucial. Het Klachteninstituut Financiële Dienstverlening (Kifid) is the Dutch primary dispute resolution body for complaints against financial service providers. The consistency and quality of their decisions heavily depend on efficient access to relevant precedents. However, jurisprudence retrieval presents significant practical challenges. This research has identified three key challenges: (1) Limited source accessibility with many unpublished rulings creating bias in the publically available data [1], (2) inconsistent search results from subjective coding schemes [2], and (3) semantic gaps where keyword search misses conceptually similar cases using different terminology [3].

Vector embeddings enable direct modeling of semantic similarity, potentially improving search results. Recent research shows how large-language-model (LLMs) systems are reshaping the legal field [4]. Broad benchmarking confirms that these models scale well across jurisdictions and tasks, from multilingual statute interpretation to judgment prediction, highlighting their capacity to accelerate several aspects of the legal work [5].

Although such systems have demonstrated promise in general legal domains, their effectiveness in specialized contexts, such as Dutch financial jurisprudence, remains unexplored.

This thesis contributes to both theoretical and practical domains by (1) providing an empirical evaluation of Dutch embedding models for jurisprudence retrieval, (2) developing a retrieval framework, and (3) establishing evaluation metrics adapted to final dispute contexts. The findings will directly support Kifid's technological development of improved jurisprudence strategies while advancing the broader understanding of AI applications in specialized legal domains. The remainder of this thesis is organized as follows: Section 2 outlines the research problem and presents the research questions. Section 3 reviews relevant literature. Section 4 describes the research methodology. Section 5 details the implementation and experiment design. Section 6 presents the findings organized by research question, followed by a discussion in Section 7. Section 8 concludes the report and outlines directions for future research.

2. Problem Statement

The challenges named in Section 1 outline a critical gap in current jurisprudence retrieval capabilities. This gap impacts the quality and efficiency of financial dispute resolution. Traditional keyword-based systems often fail to capture the semantic relationships between legal concepts, resulting

43rd Twente Student Conference on IT, July 4th, 2025, Enschede, The Netherlands.

in labor-intensive manual keyword searches. Embedding-based retrieval systems present a promising approach for capturing semantic textual meaning. However, their efficacy in Dutch financial legal contexts has not been thoroughly examined.

2.1. Research Question

To research whether embedding-based retrieval systems are a solution to this problem statement, the following research question was designed:

Can a simple embedding model-based retrieval system for jurisprudence be developed and evaluated to boost the relevance of retrieved cases for Financial Dispute Resolution?

This research question will be answered using the following three sub-questions as guides:

1) What types and volumes of jurisprudence data are available from Kifid and external sources, and how can these be prepared and processed for embedding-based retrieval systems?
2) Which embedding models and vector search techniques best capture semantic similarities between financial dispute cases?
3) How does the developed system perform in terms of technical performance metrics and practical utility for legal professionals at Kifid?

The first two subquestions form the base for the development of the proposed retrieval system. The third subquestion leads to an answer to the evaluation part of the research question.

While the developed system may have applications in other legal domains, the evaluation and optimization in this research will be tailored to financial dispute resolution contexts.

3. Related Work

This section examines the current state of research in legal information retrieval, exploring established methodologies and their evaluation frameworks to establish a foundation for this research. The section also specifies the identified research gaps.

3.1. Legal Information Retrieval

Legal Information Retrieval (LIR) is the task of retrieving relevant legal documents, such as court decisions, based on queries [6] that often involve complex legal jargon [7]. The field has evolved significantly over the past two decades, with computational legal studies emerging as a distinct discipline that promises data-driven insights into legal processes, while presenting unique challenges related to domain-specific language and reasoning patterns [8]. Finding the correct legal documents is crucial for Kifid's ruling process, as they use previous cases as a guideline for their rulings.

The application of machine learning (ML) to legal process tasks has shown great potential, with foundational work [9] showing that computational approaches can effectively predict Supreme Court behaviour using relatively simple features. This early success paved the way for more advanced AI applications. For example, Medvedeva et al. [10] successfully applied ML to predict decisions of the European Court of Human Rights.

However, legal jargon, synonym usage, and a lack of context understanding in current systems [11] hinder legal professionals.

3.2. Retrieval Methodologies

There are several different types of retrieval methods used in LIR. This subsection will focus on sparse, dense, and hybrid retrieval, respectively. Traditional legal search tools, for example, rechtspraak.nl and Legal Intelligence, are based on sparse retrieval methods. Sparse retrieval methods are techniques that index exact term occurrences in inverted lists to identify documents containing specific keywords. Historical approaches to legal text similarity research relied heavily on case-based reasoning, with early work by K.D. Ashley (1990) [12] established baseline approaches for measuring similarity in legal texts.

Examples of currently used sparse retrieval methods are TF-IDF, SVM, or BM-25 [6] [13]. These methods are effective for precise querying, but fail to capture the semantic meaning of legal language, especially when considering synonyms [14].

Dense retrieval methods represent text as vector embeddings using pretrained language models. These embeddings capture semantic properties of words or phrases by mapping them into a high-dimensional vector space [7]. Therefore, vector embeddings allow for the comparison of documents based on semantic and conceptual similarity [15], even when different wording is used. This capability is valuable in the process of jurisdiction, where semantically similar cases may differ significantly in surface language.

Pretrained language models, such as the BERT (Bidirectional Encoder Representations from Transformers) models, have been proven successful for legal case retrieval [6]. Early work on legal-domain BERT derivatives (e.g., LegalBERT, RechtBERT) explored further pre-training on legal corpora. Still, a master's thesis by M.A. Looijenga [16] found that such models do not surpass generic BERTs on downstream legal tasks due to architectural and data imbalance factors.

Recent research by J Savelka and K.D. Ashley [17] has verified this challenge by Looijenga on traditional assumptions about the necessity of fine-tuning for legal applications. This finding has significant implications for model selection, suggesting that non-fine-tuned models are more effective than previously assumed.

Most recently developed techniques combine sparse and dense retrieval [18], with very promising results for legal documents. These methods, also known as hybrid retrieval,

capture the advantages of both retrieval methods, combining the precision of sparse retrieval with the semantic understanding of dense retrieval. In the legal domain, such combinations have been found to improve both recall, the proportion of the retrieved cases that fit the query, and relevance, making them particularly suitable for precedent retrieval tasks [14] [6].

In addition to these so-called pure retrieval approaches, alternatives have emerged that utilize structured legal knowledge. Sovrano et al. [19] present knowledge graph-based approaches for legal question-answering systems, demonstrating other ways of extracting and utilizing legal knowledge for LIR purposes. These types of approaches offer an alternative to pure retrieval methods by utilizing legal reasoning structures and relationships between legal concepts. Due to these legal relationships and hierarchies, these methods offer a more transparent and interpretable approach to retrieving legal information.

3.3. Evaluation of LIR Systems

Examining how these previously introduced approaches are evaluated and measured for their effectiveness in real-world legal applications is crucial.

In the 2024 COLIEE Case Law Retrieval task, Goebel et al. [20] frame evaluation as predicting, for each query case, its set of gold cases. They compare system outputs against these mappings using standard Information Retrieval (IR) metrics (Precision, Recall, F_1) over a pooled candidate set rather than annotating an entire dataset. By pairing each query with its golden cases, this approach gives a fully reproducible, statistically sound test collection. These metrics for evaluating LIR systems are corroborated by Ma et al. [6] and Nguyen et al. [21]. The Friedman test in combination with the Wilcoxon test was determined to be the most reliable test to determine the statistical significance for IR systems by Parapar et al [22].

While benchmarks and metrics provide standardized evaluation frameworks, the legal field also requires consideration of real-world gray areas. As stated by Mukund et al. [23], standard metrics fail to capture the factual and legal fidelity expected in the legal field. Therefore, they introduce a Legal Coverage score assessed using a 5-point Likert scale by legal annotators, which measures how well retrieved documents align with their queries.

3.4. Identified Research Gaps

Even though the research reveals an ever-evolving field with established methodologies and frameworks for LIR and LIR evaluations, several gaps remain relevant for the Dutch jurisprudence context:

- Domain-specific evaluation: While general legal IR methods show promise, their effectiveness in specialized contexts like Dutch financial jurisprudence remains underexplored.
- Practical implementation frameworks: Most research focuses on algorithmic performance rather than practical integration into existing legal workflows and neglects the specific needs of dispute resolution purposes.
- Dutch language model evaluation: Limited empirical evaluation exists for Dutch embedding models in legal contexts, particularly for financial dispute resolution.
- Specialized domain performance: The assumption that semantic search advantages hold across all legal domains requires empirical validation in specialized contexts.

4. Methodology

In this section, the methodology chosen to develop and evaluate the proposed jurisprudence system is outlined. The goal of this section is to explore the theoretical basis behind the experiment design (Section 5). First, the technical framework for processing, embedding, and retrieving documents is presented. Next, the evaluation approach and metrics are outlined. Finally, the mixed-methods research design is described.

4.1. Embedding Model Selection Approach

A comparative evaluation approach was adopted to compare three pre-trained models. The following models were selected based on their potential for Dutch-language processing: RobBERT-base as the state-of-the-art Dutch BERT model [24], Multilingual E5 for its cross-lingual capabilities relevant to legal terminology [25], and RobBERTje-merged as a computationally efficient alternative [26]. The evaluation approach was based on a multi-criteria assessment framework that focused on computational and semantic performance. The framework assessed five metrics: memory footprint, speed per chunk, stability, Pearson r, and classification accuracy. A statistical comparison using non-parametric tests was chosen to handle the ordinal nature of the performance rankings. The model demonstrating the best balance of efficiency and semantic performance in these experiments was determined.

4.2. Retrieval Framework Overview

The developed framework comprises three stages: document chunking, embedding generation, and similarity search in a vector database. The chunking methodology utilizes a 512-token window length to comply with model capacity. For judicial rulings that exceed this limit, a sliding window approach with a 50-token overlap is implemented. Because practitioners often phrase constraints directly in their query (e.g. “Hoge Raad-arresten uit 2024”, “maximaal 5 uitspraken”), we introduce an LLM-based parser (parse_filters) that converts the raw text into a JSON object (Appendix B.1). A system prompt (Appendix B.2) contains the filter extraction rules. The LLM chosen to extract these filters is GPT-4.1-nano, as it is lightweight and cost-effective. The

system employed top-k retrieval with $k = 10$, returning the 10 most relevant documents to reflect conventional search behavior where users predominantly focus on the first 10 results.

$k \times 3$ chunks are retrieved per collection to allow de-duplication of the chunks on the case level. Three different retrieval methods were implemented and evaluated:

Sparse retrieval uses BM-25 scoring, which relies on term frequency and inverse document frequency weights:

$$\text{BM-25}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \text{TF}(t, d)$$

Dense retrieval maps texts to high-dimensional vectors, with semantic closeness measured using cosine similarity:

$$\text{Dense}(q, d) = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Hybrid retrieval combines sparse and dense scores through linear mixing:

$$\text{Hybrid}(q, d) = \alpha \cdot \text{BM-25 Score}(q, d) + (1 - \alpha) \cdot \text{Dense}(q, d)$$

The hybrid parameter, denoted by α, was set at 0.6 as it maximized retrieval accuracy [27]. BM-25 term-frequency saturation, denoted by k_1, was set to 1.2, and BM-25 length normalization, denoted by b, was set to 0.75 as these values are solid in many circumstances [28]. Further tuning of these parameters was deferred due to time constraints on the research.

4.3. System Evaluation

The system evaluation has two components: (1) the technical performance of the system evaluation based on standard IR metrics, and (2) a human-centered relevance assessment using a legal professional.

4.3.1. Technical Performance Evaluation. The first component evaluates the technical performance of the system using a 21-query test set with gold-label ECLIs (European Case Law Identifiers) or Kifid cases [20], which was created by Kifid professionals. For each query, five gold-label cases were determined based on expertise and legal reasoning. Performance is measured using the following three standard Information Retrieval (IR) metrics:

- Precision $= \dfrac{|R \cap S|}{|S|}$
- Recall $= \dfrac{|R \cap S|}{|R|}$
- F1-score $= 2\,\dfrac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Here, R is the set of relevant (gold-label) documents, and S is the set of results returned by the system. Note that with only five gold documents per query, the maximum value for precision is 0.5 (50%).

The Kifid professionals also provided the first 10 documents resulting from searching these queries on their current system. An identical evaluation was run on these results to obtain a baseline against which the system could be compared. Note that this evaluation only happened once, whereas the system analysis will be run thirty times, as this is the minimum sample size for reliable mean estimation with confidence intervals [29].

4.3.2. Human-centered Relevance Assessment. The second component involves a qualitative assessment to determine the relevance of retrieved documents that were not included in the gold-label cases, as they may still be relevant. For 10 queries, a collection was made of the responses for each method. Unique cases for each method were considered.

The collection of responses was blind reviewed by a legal professional based on relevance, where relevance is defined as: "pertaining to the matter at hand" by M. van Opijnen and C. Santos (2017) [30]. The review followed an approach based on the "Legal Coverage" protocol proposed by Mukund and Easwarakumar (2025) [23] for evaluating retrieval-augmented legal systems. The review also adapted the five-point Likert scale (Table 1) from this "Legal Coverage" protocol.

Table 1. RELEVANCE RUBRIC (1 = LOWEST, 5 = HIGHEST).

Score	Interpretation
1	Off-topic or factually unrelated to the query.
2	Touches on the general legal domain but not on the specific issue.
3	Partially overlaps with the issue; would require substantial additional analysis.
4	Directly addresses the legal issue but omits key facts or a controlling precedent.
5	Fully on point; a practitioner would likely cite this authority verbatim.

To mitigate bias, cases were randomly ordered, blind presented, and de-duplicated. Additionally, this research employs an externally validated rubric (Table 1) to ensure validity. These protections align with best practices recommended in recent legal-IR human evaluation work to prevent biases [23].

4.4. Research Design

This study employs a mixed-methods design to answer the sub-questions in Section 2.1, where technical system building addresses sub-question 1, experimental evaluation addresses sub-question 2, and sub-question 3 combines benchmark accuracy with qualitative utility assessment. Figure 1 illustrates the end-to-end workflow of the proposed system, progressing from data collection and pre-processing, through embedding generation and vector storage, to query processing and result ranking.

5. System Implementation and Experiment Design

This section details the technical implementation procedures used to carry out the methodology, specifying the

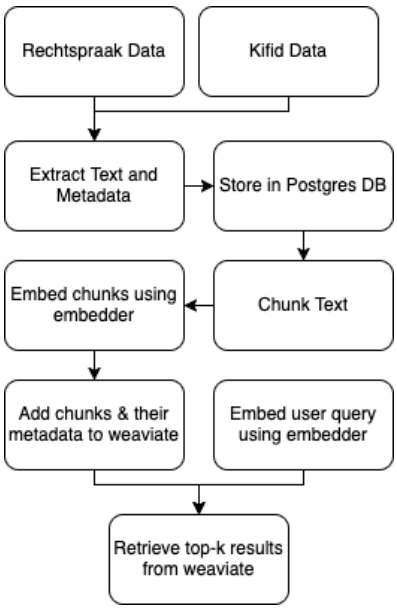

Figure 1. End-to-end workflow of the simple AI system.

data collection methods, evaluation metrics, and testing procedures.

5.1. Data Sources and Collection

This research uses two sources of legal data: Rechtspraak.nl and Kifid's rulings. These data sources were selected based on literature research and conversations with Kifid professionals, as they are the two publicly available sources of legal data that Kifid professionals use for jurisprudence. The cases in both sources are compliant with GDPR rulings. Rechtspraak.nl contains cases on a wide range of legal categories, but for this research, only the categories relevant to Kifid's line of work were considered. The data is provided through an Atom feed and XML documents, though bulk data retrieval is not directly supported. Data collection utilized the `rechtspraak-js` tool, modified to fetch data from specific legal categories identified by Kifid professionals.

Kifid also publishes its rulings on its website (www.kifid.nl/uitspraken). The full rulings are accessible in PDF format. However, no structured or API-based access is provided. Therefore, to compile a comprehensive dataset of Kifid rulings, a custom web crawler was developed to systematically find valid URLs using naming conventions based on year, month, and case numbering. Then, a script was designed to download and parse the information from these PDF files using a combination of `PyPDF2` and `pdfminer`.

5.2. Data Preparation and Analysis

To prepare the collected data for embedding, several techniques were used. For the Rechtspraak data, the complete information of each case was parsed to extract metadata, including case date, court, and the main textual content of the decision. This data was stored in a PostgreSQL database. For each Kifid PDF, text was extracted and key metadata identified by using regular expression patterns. The extracted information was cleaned to remove formatting artifacts and other noise. The processed data was inserted into a PostgreSQL database table for further processing. The specific composition of the databases can be found in Appendix C.

5.2.1. Dataset Limitations. As explained in the introduction (Section 1), many law cases are settled before reaching a ruling. The details of these settlements remain unpublished and can therefore not be used as jurisprudence. This creates a systematic bias toward cases that could not be resolved, as only unsettled cases that proceeded to formal ruling are included. Despite the identified biases, the combined dataset provides a robust foundation for the experimental retrieval task due to: (1) substantial document volume (13,082 total cases), (2) complete full-text availability, (3) rich metadata for filtering, and (4) comprehensive coverage spanning recent Dutch legal history. The identified biases will be taken into account when interpreting the results.

To prepare legal texts for embedding, a token-aware chunking strategy that preserves semantic context while staying inside model limitations was implemented as outlined in Section 4.2.

5.3. Embedding Model Selection Protocol

An empirical comparison was conducted to evaluate the model's performance in terms of accuracy and computational efficiency. Random sampling of 200 Rechtspraak text chunks was performed. From these, 100 pseudo-labeled sentence-pair examples were constructed (50 similar, 50 dissimilar) by joining rows in the dataset based on whether the rows had overlapping sources and subjects. Each model was warmed up and then assessed under identical hardware and preprocessing conditions. Embedding speed was quantified as the mean time (in milliseconds) required to embed a single chunk across batch sizes 1 – 64; five timed runs were made after two warm-up runs, outliers beyond two standard deviations were discarded, and 95% confidence intervals were reported. Peak changes in resident memory (ΔRSS, MB) were recorded for small (10-chunk), medium (50-chunk), and large (100-chunk) batches, after which per-chunk memory usage and confidence intervals were derived.

Embedding stability was assessed by measuring the minimum cosine similarity obtained under three identical repeated runs, varying batch sizes, and shuffled input order, with thresholds of ≥ 0.999, ≥ 0.995 and ≥ 0.999 respectively.

Semantic-similarity quality was evaluated against the 100 sentence pairs, producing Pearson r (with 95% confidence bounds) and median-threshold classification accuracy.

Finally, Wilcoxon signed-rank tests, paired t-tests, and Cohen's d were applied to every model pair, and normalized scores for speed, memory, stability, semantic quality,

and downstream accuracy were combined into a composite ranking that determined the model chosen for the retrieval pipeline.

5.4. Vector Search and Retrieval Method

Weaviate was selected as the vector database due to its open-source nature and support for sparse, dense, and hybrid search.

For the evaluation of the system, a corpus of validation data was created by the legal professionals of Kifid. They provided queries and golden cases, which are cases that one would expect to occur in the retrieved documents of a well-functioning retrieval system for a given query. Ultimately, the corpus comprised 21 queries, each with a total of 99 gold-label cases, resulting in an average of approximately 4.7 cases per query. The corpus is presented in Table 11 (Appendix G). This corpus was used to validate the system's results using standard IR metrics (Precision, Recall, and F_1). This evaluation was conducted for sparse retrieval, dense retrieval, and hybrid retrieval to determine which method performs best.

For each query, the system returns 10 documents. Then, Precision, Recall, and F_1 are computed to evaluate the performance of the system. The results of this system are compared to results from the current searching method to analyze whether retrieval performance improved.

5.4.1. Human Evaluation. To verify the system's results, a human evaluation was conducted. Ten queries from the validation dataset were selected for further evaluation. For these ten evaluation queries, the top 10 results returned by each retrieval method (sparse, dense, hybrid) were pooled after removing the gold-label cases. Duplicates found between methods for a single query were de-duplicated on ECLI/case-number, returning a union set of $N = 254$ unique candidate cases, averaging ≈ 25.4 cases per query. The precise numbers per query are listed in Table 6 (Appendix D). A legal professional assessor reviewed each case and assigned Likert scale scores on a 5-point scale in a spreadsheet format. At the end of each query batch, the expert could review earlier scores if later readings changed their assessment. For analysis, the mean Likert score over unique hits ($\bar{L}$) was computed for each retrieval variant, along with proportions of 'high coverage' cases ($L \geq 4$) and 'irrelevant' cases ($L \leq 2$). Statistical differences between variants were tested using the non-parametric Wilcoxon signed-rank test (paired on query) to handle the ordinal nature of Likert data.

6. Results

This section presents the findings across the three key areas of investigation introduced by the sub-questions: first, the data characteristics and processing outcomes for the Dutch financial jurisprudence corpus; second, a comparative analysis of embedding model performance; and third, an evaluation of different retrieval methods for legal document similarity search.

6.1. Data Characteristics and Processing

The data collection process compiled a dataset totaling 13,082 Dutch financial jurisprudence documents from the two sources discussed in Section 5:

- Kifid Dataset (4,791 documents): Financial dispute resolution decisions spanning May 11, 2011, to June 2, 2025.
- Rechtspraak.nl Dataset (8,291 documents): Court judgments from March 13, 1936, to June 2, 2025.

The specific composition of datasets can be found in Appendix C.

The data preparation pipeline successfully transformed legal documents into an embedding-ready format using two distinct processing strategies. The processing strategy successfully prepared both PDF and XML sources for embedding-based retrieval, maintaining legal document structure while ensuring compatibility with the selected embedding model. The resulting dataset offers comprehensive coverage, making it suitable for semantic similarity-based jurisprudence retrieval in Dutch financial dispute contexts.

6.2. Embedding Model Performance

Table 2 presents the computational performance metrics for the models. The results reveal a clear trade-off between computational efficiency and model complexity. RobBERTje-merged is the fastest, processing each chunk in only 87.87 ms, approximately 2.1 times faster than RobBERT-base. The multilingual E5 model is 3.6× slower than RobBERT-base while consuming 4.5× more memory. All models exhibit excellent embedding stability, with a minimal cosine similarity of ≥0.995 across repeated runs, batch size changes, and reordering.

Table 2. COMPUTATIONAL PERFORMANCE COMPARISON OF EMBEDDING MODELS

Metric	RobBERT-base	RobBERTje-merged	Multilingual E5
Memory footprint (MB)	49.5	70.5	220.9
Speed (ms per chunk)	185.97	87.87	669.17
Stability (min. cosine sim.)	≥ 0.995	≥ 0.995	≥ 0.995

Table 3 shows the semantic similarity evaluation results. The multilingual E5 model reaches the highest Pearson correlation with human judgments ($r = 0.8266$), followed by RobBERT-base ($r = 0.7851$) and RobBERTje-merged ($r = 0.6901$). Downstream 10-fold cross-validated similarity-classification accuracy is highest for RobBERT-base (75%), outperforming multilingual E5 (60%) and RobBERTje-merged (50%). Paired Wilcoxon signed-rank tests on Pearson scores confirm that all model pairs differ significantly ($p < 0.001$).

6.3. Retrieval Method Comparison

Table 4 presents the average performance metrics per method, including the evaluation on current systems

Table 3. SEMANTIC PERFORMANCE COMPARISON OF EMBEDDING MODELS

Metric	RobBERT-base	RobBERTje-merged	Multilingual E5
Pearson r	0.7851	0.6901	0.8266
Classification accuracy (%)	75	50	60

(Keyword Baseline). All figures, except for the baseline, are reported as mean ± 95% confidence interval across 30 independent runs.

Metric	Keyword Baseline	Vector	Hybrid
Precision	0.2298	0.0190 ± 0.0000	0.062 ± 0.0005
Recall	0.4143	0.0381 ± 0.0000	0.1248 ± 0.0011
F_1 Score	0.2918	0.0254 ± 0.0000	0.0832 ± 0.0007

Table 4. AVERAGE PERFORMANCE METRICS FOR VECTOR, KEYWORD, AND HYBRID RETRIEVAL METHODS.

Vector retrieval scores the worst in every section, achieving only 1.9% precision compared to the established baseline of 22.98% with a Confidence Interval (CI) of 0, which is due to the indexing of the vectors by Weaviate. Hybrid retrieval fails to leverage the strengths of either method, achieving an F_1 score of only 8.32%. Most notably, the traditional keyword baseline outperforms both embedding-based approaches by significant margins.

An exploratory evaluation of BM-25 keyword retrieval revealed substantially superior performance compared to vector and hybrid methods. Table 5 presents the comparative results including BM-25. The expanded comparison confirms the dominance of keyword-based approaches in this domain. The statistical analysis found no significant difference between the two keyword-based methods.

Human evaluation revealed that for non-golden-case retrieval, baseline significantly outperformed hybrid (1.5), and vector (1.24) on the mean Likert ($\bar{L}$) in pair-wise Wilcoxon signed-rank tests (all $p < 0.02$). However, no significant difference was found between keyword ($\bar{L} = 2.27, p = 0.4961$) and baseline. Baseline was also the system with the highest share of high-coverage results ($L \geq 4$), at 29%. The full results can be found in Tables 7 and 8 (Appendix D).

Metric	Keyword Baseline	Vector	Hybrid	BM-25
Precision	0.2298	0.0190 ± 0.0000	0.062 ± 0.0005	0.1746 ± 0.0014
Recall	0.4143	0.0381 ± 0.0000	0.1248 ± 0.0011	0.3008 ± 0.0027
F_1 Score	0.2918	0.0254 ± 0.0000	0.0832 ± 0.0007	0.2166 ± 0.0019

Table 5. AVERAGE PERFORMANCE METRICS INCLUDING BM-25 KEYWORD RETRIEVAL.

The per-query breakdown further illuminates the consistent superiority of keyword methods. Comparison between the baseline, hybrid and BM-25 methods showed 0 queries where hybrid had the highest score, 6 queries with BM-25 performing the best, 10 queries where baseline outperformed the other methods, and 5 queries with ties (2 among all three methods, 2 between hybrid and BM-25, and 1 between baseline and BM-25). Table 9 (Appendix E) shows the full comparison per query.

To test for overall differences across the four methods (keyword baseline, vector, hybrid, and BM-25), the Friedman omnibus test on per-query F_1 scores was applied. The test resulted in $\chi^2_F(3) = 23.413, \quad p = 0.0000$, indicating a significant difference between at least two retrieval methods.

Post-hoc pairwise Wilcoxon signed-rank tests with Holm correction showed that every comparison, except keyword versus baseline, was significant at $\alpha < 10$. The full test results are summarized in Table 10 (Appendix F).

7. Discussion

This study evaluated an embedding-based jurisprudence retrieval system for Dutch financial dispute resolution, comparing sparse, dense, and hybrid approaches against traditional keyword search methods. This section contains: a summary of key findings, further interpretations of the results, a comparison to the existing literature, the limitations and edge cases of the research, and practical implications of the results.

7.1. Summary of Key Findings

The findings answer each sub-research question: (1) The study compiled and processed 13,082 Dutch financial jurisprudence documents from Kifid and Rechtspraak.nl sources; (2) The embedding model with the best combination of compute and semantic quality was RobBERTje-merged, offering superior computational efficiency (87.87 ms per chunk) while trading only limited semantic quality reduction (r = 0.6901); (3) The evaluation of system performance and practical utility found that embedding-based methods did not boost retrieval relevance, with pure vector retrieval performing poorly (F_1 = 0.0254), hybrid retrieval achieving no improvement over the keyword baseline (F_1 = 0.0832 vs. 0.2918 respectively), and traditional BM-25 keyword retrieval demonstrating significantly better performance (F_1 = 0.2166) across all metrics except for the baseline, where no significance was achieved. Human evaluation confirmed these findings, with the baseline achieving the highest share of high-coverage results (29%) and significantly outperforming both hybrid and vector methods in pairwise comparisons. Therefore, while a functional embedding-based retrieval system was successfully developed, the research concludes that the developed system does not significantly boost retrieval relevance compared to conventional keyword-based approaches in Dutch Financial Dispute Resolution contexts.

7.2. Interpretation of Results

Several domain-specific factors can explain the superiority in performance of keyword-based methods. First, legal professionals appear to formulate queries using specific terminology and entity names that directly match document content. This reduces the value of semantic similarity matching. The query analysis revealed that keyword search excelled when queries contained specific legal entities such as court names (e.g., "Hoge Raad"), legal citations (e.g., article numbers), or specialized financial terms (e.g., "American depositary receipts"). Second, financial legal language might be more standardized and consistent than assumed. Financial dispute documents appear to use relatively consistent terminology across cases, sources, and periods, which disadvantages methods aiming to close the semantic gap, such as embeddings. As mentioned in the results (Section 6), hybrid retrieval does not outperform the baseline and keyword methods on any query. Further analysis of the queries reveals a possible explanation. Keyword search excelled when queries contained specific legal entities (court names, ECLI citations, specialized legal terms), while hybrid retrieval tied with other methods on queries where the language used was general and the query did not include specific entities or highly specialized legal language. Queries 4 and 20 in Table 11 (Appendix G) are good examples of this.

7.3. Comparison to Existing Literature

These findings contrast with the general literature on legal information retrieval, where embedding-based systems typically outperform keyword approaches [6] [14]. Several studies report significant improvements when applying dense retrieval methods to legal case collections, particularly for English-language corpora. However, several factors may explain this contrast:

- Domain Specificity: Financial dispute resolution may require more precise, entity-focused retrieval than general legal domains and other types of legal applications
- Language Characteristics: Dutch legal language in financial contexts may be more standardized, reducing the benefits of semantic understanding
- User Requirements: Dispute resolution professionals may prefer precise, predictable results over broader semantic matches

7.4. Limitations and Edge-Cases

Several limitations arose in this research. Only published rulings are used in the dataset, as legal data inherently contains bias. The evaluation was conducted on a relatively small gold standard set of 21 queries corresponding to 99 relevant cases. While this dataset enabled initial performance assessment, the limited scope may not fully represent the diversity of information needs in legal practice, potentially undermining the statistical robustness of the conclusions drawn. The chosen evaluation metrics, Precision, Recall, and F_1, may not fully capture practical relevance. The evaluation in this thesis mainly relied on technical performance metrics without incorporating user studies. This limitation arose partly from the practical constraints of accessing legal practitioners for validation purposes, as such collaboration requires significant time investment from busy professionals. The absence of user evaluation represents a considerable gap between technical performance and practical utility that should be addressed in future research. Finally, due to time constraints, only three embedding models were evaluated, which may have resulted in the use of a suboptimal embedding model.

7.5. Practical Implications

The results suggest that Kifid's current keyword-based search approach may be more effective than embedding-based retrieval systems for their specific use case. Implementing advanced keyword matching algorithms on datasets that contain only relevant data for financial dispute jurisprudence could provide improvements over the current system. For the legal technology community, this study contributes to a more in-depth understanding of when AI, and specifically embeddings, can enhance legal information retrieval. The findings suggest that domain-specific evaluation is crucial, as general performance claims may not hold in specialized legal contexts.

8. Conclusion

This research investigated the development and evaluation of an AI-based jurisprudence retrieval system for Dutch financial dispute resolution contexts. A functional embedding-based retrieval system was successfully developed. However, the research concludes that this system does not significantly boost retrieval relevance compared to conventional keyword-based approaches in Dutch Financial Dispute Resolution contexts. Therefore, this research highlights the importance of evaluating AI implementations on a per-use-case basis in specialized legal contexts.

Future research should focus on expanding the size of the gold-label set by involving more Kifid practitioners and creating 50–100 varied queries, each with five gold cases, thereby stabilizing metrics and allowing for the testing of the influence of query formulation. User studies should be conducted with 5–10 legal experts using the system on real tasks, collecting both qualitative feedback (ease of finding cases and trust in results) and quantitative data (time taken and number of clicks) to measure the practical impact. Alternative hybrid schemes should be explored, such as experimenting with different α values or utilizing learning-to-rank on both sparse and dense features. Finally, investigating whether fine-tuned embedding models improve performance would address the applicability of the conclusions of Looijenga [16] to this specific application.

References

[1] F. Peeraer and R. van Gestel, "Systematische jurisprudentieanalyse als uitdaging voor onderwijs en onderzoek," in *Methoden van systematische rechtspraakanalyse*, M. Vols, Ed. n.p., 2024, pp. 185–209.

[2] P. T. J. Wolters, "Mogelijkheden en beperkingen bij de kwantitatieve analyse van jurisprudentie voor de bestudering van geldend recht," in *Methoden van systematische rechtspraakanalyse*, M. Vols, Ed. n.p., 2024, pp. 47–67.

[3] A. Dyevre, "Exploring and searching judicial opinions with top2vec," in *Methoden van systematische rechtspraakanalyse*, M. Vols, Ed. n.p., 2024, pp. 143–161.

[4] J. Ni, Y. Fan, J. Merane, E. Salimbeni, Y. Huang, M. Akhtar, F. Geering, and et al., "LEXam: Benchmarking legal reasoning on 340 law exams," *arXiv e-prints*, May 2025. [Online]. Available: https://arxiv.org/abs/2505.12864

[5] J. Wang, H. Zhao, Z. Yang *et al.*, "Legal evaluations and challenges of large language models," *arXiv e-prints*, October 2024. [Online]. Available: https://arxiv.org/abs/2411.10137

[6] Y. Ma, Y. Wu, Q. Ai, Y. Liu, Y. Shao, M. Zhang, and S. Ma, "Incorporating structural information into legal case retrieval," *ACM Transactions on Information Systems*, vol. 42, no. 2, p. Article 40, 2023. [Online]. Available: https://doi.org/10.1145/3609796

[7] C. Sansone and G. Sperlí, "Legal information retrieval systems: State-of-the-art and open issues," *Information Systems*, vol. 106, p. 101967, 2022. [Online]. Available: https://doi.org/10.1016/j.is.2021.101967

[8] G. Van Dijck, M. Vols, and B. Schäfer, "Computational legal studies: The promise and challenge of data-driven research in law," in *Computational Legal Studies*, J. Medina and M. van Opijnen, Eds. Edward Elgar Publishing, 2021, ch. 10, pp. 186–204. [Online]. Available: https://doi.org/10.4337/9781788977456.00017

[9] D. M. Katz, M. J. Bommarito, and J. Blackman, "A general approach for predicting the behavior of the supreme court of the united states," *PLOS ONE*, vol. 12, no. 4, p. e0174698, 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0174698

[10] M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the european court of human rights," *Artificial Intelligence and Law*, vol. 28, no. 2, pp. 237–266, 2020. [Online]. Available: https://doi.org/10.1007/s10506-019-09255-y

[11] F. Ariai and G. Demartini, "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges," *arXiv preprint arXiv:2410.21306*, 2025, version v2.

[12] K. D. Ashley, *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*, ser. Bradford Books. Cambridge, MA: MIT Press, 1990.

[13] I. Schepers, M. Medvedeva, M. Bruijn, M. Wieling, and M. Vols, "Predicting citations in dutch case law with natural language processing," *Artificial Intelligence and Law*, vol. 32, pp. 807–837, 2024.

[14] Y. Feng, C. Li, and V. Ng, "Legal case retrieval: A survey of the state of the art," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. n.p., 2024.

[15] R. Kalra, Z. Wu, A. Gulley, A. Hilliard, X. Guan, A. Koshiyama, and P. Treleaven, "Hypa-rag: A hybrid parameter-adaptive retrieval-augmented generation system for ai legal and policy applications," *arXiv preprint arXiv:2409.09046*, 2025, version v2; preprint.

[16] M. S. Looijenga, "Rechtbert: Training a dutch legal bert model to enhance legaltech," M.Sc. thesis, University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science, Enschede, The Netherlands, Dec. 2024.

[17] J. Savelka and K. D. Ashley, "The unreasonable effectiveness of large language models in zero-shot semantic similarity tasks for legal texts," *Frontiers in Artificial Intelligence*, vol. 6, p. 1079794, 2023. [Online]. Available: https://doi.org/10.3389/frai.2023.1079794

[18] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2024, version v5.

[19] F. Sovrano, M. Palmirani, and F. Vitali, "Legal knowledge extraction for knowledge graph based question-answering," *Artificial Intelligence and Law*, vol. 28, no. 3, pp. 365–399, 2020. [Online]. Available: https://doi.org/10.1007/s10506-020-09265-8

[20] R. Goebel, Y. Kano, M.-Y. Kim, J. Rabelo, K. Satoh, and M. Yoshioka, "Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024," in *JSAI-isAI 2024*, ser. Lecture Notes in Artificial Intelligence, T. Suzumura and M. Bono, Eds., vol. 14741. Springer Nature Singapore Pte Ltd., 2024, pp. 109–124.

[21] H. Nguyen, H. Nguyen, T. Pham, M. Nguyen, A. Trieu, D.-T. Do, N.-K. Le, and L.-M. Nguyen, "Jnlp at coliee 2025: Hybrid large language model–based framework for legal information retrieval and entailment," in *Workshop on the 12th Competition on Legal Information Extraction and Entailment (COLIEE'2025) at the 22nd International Conference on Artificial Intelligence and Law (ICAIL)*, Chicago, USA, Jun. 2025. [Online]. Available: https://coliee.org/

[22] J. Parapar, D. E. Losada, M. A. Presedo-Quindimil, and A. Barreiro, "Using score distributions to compare statistical significance tests for information retrieval evaluation," *Journal of the Association for Information Science and Technology*, 2019, submitted Oct 11, 2017; Accepted Jan 11, 2019.

[23] S. A. Mukund and K. S. Easwarakumar, "Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation," *Symmetry*, vol. 17, no. 5, p. 633, May 2025. [Online]. Available: https://doi.org/10.3390/sym17050633

[24] P. Delobelle and F. Remy, "Robbert-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion," *Computational Linguistics in the Netherlands Journal*, vol. 13, pp. 193–203, 2024.

[25] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual e5 text embeddings: A technical report," Microsoft Corporation, Tech. Rep., 2023, released in mid-2023. [Online]. Available: https://github.com/microsoft/unilm/tree/master/e5

[26] P. Delobelle, T. Winters, and B. Berendt, "Robbertje: A distilled dutch bert model," *Computational Linguistics in the Netherlands Journal*, vol. 11, pp. 125–140, 2021. [Online]. Available: https://www.clinjournal.org/clinj/article/view/131

[27] H.-L. Hsu and J. Tzeng, "Dat: Dynamic alpha tuning for hybrid retrieval in retrieval-augmented generation," 2025. [Online]. Available: https://arxiv.org/abs/2503.23013

[28] S. E. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm-25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, January 2009.

[29] S. G. Kwak and J. H. Kim, "Central limit theorem: the cornerstone of modern statistics," *Korean Journal of Anesthesiology*, vol. 70, no. 2, pp. 144–156, 2017. [Online]. Available: https://doi.org/10.4097/kjae.2017.70.2.144

[30] M. van Opijnen and C. Santos, "On the concept of relevance in legal information retrieval," *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 65–87, 2017. [Online]. Available: https://doi.org/10.1007/s10506-017-9195-8

Appendix A. AI Disclosure

During the development of this research, I utilized ChatGPT and Claude to gain guidance on how to structure my thesis, search for information, and aid in the writing

process. I also used Gemini to aid in coding. Lastly, I used Grammarly to improve my writing. Other search engines or applications used in the process may also have employed AI technologies without my explicit knowledge. After using these tools and services, I thoroughly reviewed and edited the content as needed, taking full responsibility for the outcome.

Appendix B. Code

B.1. JSON Format

Listing 1. JSON Format LLM Response

```
{
    "court": "...", "date_from": "...", "date_to": "...",
    "limit": 5, "source": "kifid|rechtspraak", "article":
         "...",
    "text_query": "semantic␣core␣of␣q"
}
```

B.2. LLM Interaction

The following prompt was used:

Listing 2. Prompt used for natural-language filter extraction

```
"""
    Extract structured filters from the following legal
         query. Return a JSON object with these fields:
    - court: Court name (e.g., "Hoge Raad", "Gerechtshof
        Amsterdam", "Rechtbank Amsterdam")
    - date_from: Start date in YYYY-MM-DD format
    - date_to: End date in YYYY-MM-DD format
    - limit: Number of requested results (e.g., "3", "5")
    - source: Source of the ruling ("kifid" or "
        rechtspraak")
    - article: Article reference (e.g., "7:929 lid 1 BW")
    - text\_query: The main search query without filters

    Rules for date handling:
    - "meest recente" or "laatste" means from {
        two_years_ago_str} to today
    - "afgelopen twee jaar" or "afgelopen 2 jaar" means
        from {two_years_ago_str} to today
    - "afgelopen X jaar" means last X years from today
    - "uit YYYY" or "in YYYY" means the full year YYYY (
        from YYYY-01-01 to YYYY-12-31)
    - "uit YYYY en ZZZZ" means from YYYY-01-01 to ZZZZ
        -12-31
    - "van na 1 januari YYYY" or "vanaf YYYY" means from
        YYYY-01-01 to today
    - "van voor YYYY" or "tot YYYY" means until YYYY
        -12-31

    Rules for source determination:
    - If query mentions "Kifid", "Geschillencommissie", "
        Geschillencommissie Kifid", "GC", set source to "
        kifid"
    - If query mentions "Rechtbank", "Gerechtshof", "Hoge
         Raad", "civiele rechter", "ECLI", set source to
        "rechtspraak"
    - If both sources are mentioned or no source is
        specified, set source to null

    Rules for article references:
    - Extract full article references like "7:929 lid 1
        BW", "6:162 BW"
    - Always include the code (e.g., "BW", "Rv") if
        mentioned

    Rules for court extraction:
    - Extract specific courts like "Rechtbank Amsterdam",
         "Hoge Raad", "Gerechtshof Den Haag"
    - If "civiele rechter" is mentioned without a
        specific court, don't set a court

    Rules for limit extraction:
    - Always look for numbers like "3 uitspraken", "geef
        me 5", "zoek 10 cases"
    - If a number is found before words like "uitspraken
        ", "zaken", or after "geef me", extract as limit
    - If no limit is mentioned, don't set a limit value

    Rules for text query extraction:
    - The text\_query should contain the main semantic
        content without filters
    - Remove phrases like "geef me", "geef mij", "zoek",
        "uitspraken over", etc.
    - Keep specific legal concepts like "spoofing", "
        moment van onbedachtzaamheid", "normale
        voorzichtigheid"
    - Always include article content in text\_query if it
        's relevant to the search, not just a filter
    - For queries about multiple concepts (e.g., "normale
         voorzichtigheid waarbij sprake is van..."),
        include all concepts

    Current year is {current_year}.

    Example queries and extractions:

    Query: "Geef de meest recente uitspraken over de
        normale voorzichtigheid waarbij sprake is geweest
         van een moment van onbedachtzaamheid"
    Result: {{"court": null, "date_from": "{
        two_years_ago_str}", "date_to": "{current_date.
        strftime('\%Y-\%m-\%d')}", "limit": null, "source
        ": "kifid", "article": null, "text_query": "
        normale voorzichtigheid moment van
        onbedachtzaamheid"}}

    Query: "Geef me 5 uitspraken van de afgelopen twee
        jaar over ambtshalve toetsen van de
        terhandstelling van de verzekeringsvoorwaarden"
    Result: {{"court": null, "date_from": "{
        two_years_ago_str}", "date_to": "{current_date.
        strftime('\%Y-\%m-\%d')}", "limit": 5, "source":
        null, "article": null, "text_query": "ambtshalve
        toetsen terhandstelling verzekeringsvoorwaarden
        "}}

    Query: "Geef me 5 uitspraken van Kifid waarbij fraude
         door de consument bij het indienen van een
        schadeclaim niet is vast komen te staan"
    Result: {{"court": null, "date_from": null, "date_to
        ": null, "limit": 5, "source": "kifid", "article
        ": null, "text_query": "fraude consument indienen
         schadeclaim niet vast komen staan"}}

    Query: "Geef 3 uitspraken uit 2025 waarbij de duur
        van een EVR-registratie moet worden aangepast"
    Result: {{"court": null, "date_from": "2025-01-01", "
        date_to": "2025-12-31", "limit": 3, "source":
        null, "article": null, "text_query": "duur EVR-
        registratie aangepast"}}

    Query: "Geef 3 uitspraken uit 2024 van de civiele
        rechter waarbij de interne registraties moeten
        worden doorgehaald"
    Result: {{"court": null, "date_from": "2024-01-01", "
        date_to": "2024-12-31", "limit": 3, "source": "
        rechtspraak", "article": null, "text_query": "
        interne registraties doorgehaald"}}

    Query: "{query}"

    Return only the JSON object, no other text.
"""
```

Appendix C. Dataset Analysis

The Kifid dataset comprises 4791 financial dispute cases published between May 11, 2011, and June 2, 2025. These documents represent consumer complaints against financial institutions that have proceeded to formal decisions. The dataset is characterized by:

- Document Length: Relatively concise documents with an average length of 2383 words (σ=1254)
- Categorization: Five unique categories in the metadata
- Product Types: 15 distinct financial products referenced in disputes

- Keywords: 16 unique keywords for content classification
- Decision Types: Primarily "Bindend advies" (binding advice) decisions (3363 cases)

The Rechtspraak dataset comprises 8,291 court cases published between March 13, 1936, and June 2, 2025, with the vast majority occurring after 2012. These represent formal court proceedings across various Dutch courts. The dataset features:

- Document Length: Substantially longer documents with an average length of 9498 words (σ=13834)
- Court Distribution: Cases from 26 unique courts, with Gerechtshof 's-Hertogenbosch (1259 cases), Rechtbank Limburg (868 cases), Rechtbank Rotterdam (834 cases), Gerechtshof Den Haag (556 cases) and Hoge Raad (537 cases) most represented
- Subject Areas: 22 unique civil law subjects, with "Internationaal Privaatrecht" (international private law, 457 cases) and "Goederenrecht" (property law, 399 cases) most common
- Procedural Types: 27 distinct procedural classifications, including "eersteAanlegEnkelvoudig" (first instance single-judge, 558 cases) and "hogerBeroep" (appeals, 417 cases)

Temporal analysis shows increasing publication rates for both datasets in recent years, with particularly high volumes in 2021–2022, suggesting greater transparency in the Dutch legal system over time.

Appendix D. Human Evaluation

#	Query	Number of returned cases
1	Geef de meest recente Kifid uitspraken over de normale voorzichtigheid waarbij sprake is geweest van een moment van onbedachtzaamheid	27
2	Geef de meest recente Kifid uitspraken over de normale voorzichtigheid waarbij geen sprake is geweest van een moment van onbedachtzaamheid	25
3	Geef me 3 Kifid uitspraken van de afgelopen twee jaar over ambtshalve toetsen van de terhandstelling van de verzekeringsvoorwaarden	27
4	Geef me 5 uitspraken van Kifid waarbij fraude door de consument bij het indienen van een schadeclaim niet is vast komen te staan	28
5	Geef 3 uitspraken uit 2025 waarbij de duur van een EVR-registratie moet worden aangepast	32
6	Geef 3 uitspraken van Kifid waarbij de verzekeraar de schadevrije jaren van de consument moest herstellen	20
7	Geef mij 3 uitspraken die de Hoge Raad in 2024 en 2025 heeft gedaan op het gebied van verzekeringsrecht waarbij je de artikel 81 RO-zaken niet meeneemt	17
8	Geef me alle uitspraken (gewezen in 2025) waarin een vordering wegens verjaring is afgewezen	30
9	Zoek de uitspraken die het afgelopen jaar over levensverzekeringen zijn gedaan	20
10	Zoek arresten van de Hoge Raad over de uitleg van begunstiging bij levensverzekeringen	28
	Total	**254**

Table 6. OVERVIEW OF THE NUMBER OF CASES PER QUERY IN THE REVIEW

Method	**Mean Likert**	$L \geq 4$	$L < 4$
Baseline	2.61	29%	52%
Keyword (BM-25)	2.27	14%	61%
Hybrid	1.50	6.1%	85%
Vector	1.24	2.0%	93%

Table 7. LIKERT-SCALE RESULTS FROM HUMAN EVALUATION OVER 10 QUERIES

Table 8. PAIRWISE WILCOXON SIGNED-RANK TEST RESULTS FOR HUMAN EVALUATION

Comparison	W	p
Keyword vs. Hybrid	3.000	0.0391
Keyword vs. Vector	1.000	0.0156
Keyword vs. Baseline	16.00	0.4961
Hybrid vs. Vector	1.000	0.0156
Hybrid vs. Baseline	0.000	0.0039
Vector vs. Baseline	1.000	0.0039

Appendix E. Per-Query Analysis

Query	Baseline	Hybrid ΔF1	BM-25 ΔF1	Highest Scoring Method
1	0.267	-0.133	0.133	BM-25
2	0.000	0.133	0.533	BM-25
3	0.143	-0.010	-0.010	Baseline
4	0.000	0.000	0.000	Tie (Baseline, Hybrid, BM-25)
5	0.133	-0.133	0.000	Tie (Baseline, BM-25)
6	0.286	-0.286	-0.143	Baseline
7	0.267	-0.267	-0.267	Baseline
8	0.400	-0.133	0.600	BM-25
9	0.600	-0.200	-0.200	Baseline
10	0.533	-0.400	0.133	BM-25
11	0.000	0.000	0.267	BM-25
12	0.000	0.133	0.133	Tie (Hybrid, BM-25)
13	0.533	-0.533	-0.533	Baseline
14	0.267	-0.267	-0.267	Baseline
15	0.400	-0.400	-0.400	Baseline
16	0.133	0.000	0.295	BM-25
17	0.667	-0.533	-0.533	Baseline
18	0.667	-0.667	-0.667	Baseline
19	0.000	0.133	0.133	Tie (Hybrid, BM-25)
20	0.000	0.000	0.000	Tie (Baseline, Hybrid, BM-25)
21	0.833	-0.833	-0.833	Baseline

Table 9. PER-QUERY BASELINE F_1, THE ΔF_1 FOR HYBRID AND BM-25 (COMPUTED AGAINST THE BASELINE), AND THE HIGHEST-SCORING METHOD.

Appendix F. Wilcoxon Signed-Test Results

Table 10. PAIRWISE WILCOXON SIGNED-RANK TEST RESULTS WITH HOLM-ADJUSTED p-VALUES

Comparison	W	p_{adj}
keyword vs hybrid	0.000	0.0252
keyword vs vector	0.000	0.0069
keyword vs baseline	64.000	0.5534
hybrid vs vector	0.000	0.0252
hybrid vs baseline	12.000	0.0146
vector vs baseline	2.500	0.0063

Appendix G. Queries

Table 11: All Queries With Baseline and Golden Case Responses

ID	Query	Responses	Golden Case Responses
1	Geef de meest recente Kifid uitspraken over de normale voorzichtigheid waarbij sprake is geweest van een moment van onbedachtzaamheid	Kifid 2025-0283, Kifid 2021-0482, Kifid 2020-590, Kifid 2020-461, Kifid 2020-430, Kifid 2020-191, Kifid 2019-961, Kifid 2019-841, Kifid 2019-616, Kifid 2019-615	Kifid 2025-0283, Kifid 2025-0101, Kifid 2024-1082, Kifid 2024-0804, Kifid 2021-0482
2	Geef de meest recente Kifid uitspraken over de normale voorzichtigheid waarbij geen sprake is geweest van een moment van onbedachtzaamheid	Kifid 2021-0482, Kifid 2020-590, Kifid 2020-461, Kifid 2020-430, Kifid 2020-191, Kifid 2019-961, Kifid 2019-841, Kifid 2019-616, Kifid 2019-615, Kifid 2019-581	Kifid 2025-0420, Kifid 2025-0305, Kifid 2025-0283, Kifid 2025-0266, Kifid 2024-0041
3	Geef me Kifid uitspraken van de afgelopen twee jaar over ambtshalve toetsen van de terhandstelling van de verzekeringsvoorwaarden	Kifid 2024-0678, Kifid 2024-0041, Kifid 2024-0419, Kifid 2024-0041, Kifid 2022-0575, Kifid 2022-0013, Kifid 2022-0398, Kifid 2022-0102, Kifid 2022-0095, Kifid 2021-0861	Kifid 2025-0266, Kifid 2024-1112, Kifid 2024-1126, Kifid 2024-0101, Kifid 2024-0041
4	Geef me uitspraken van Kifid waarbij fraude door de consument bij het indienen van een schadeclaim niet is vast komen te staan	Kifid 2025-0254, Kifid 2024-0083, Kifid 2024-0368, Kifid 2024-0106, Kifid 2024-0001, Kifid 2023-0957, Kifid 2023-0031, Kifid 2023-0453, Kifid 2023-0379, Kifid 2023-0255	Kifid 2025-0199, Kifid 2024-0913, Kifid 2024-0812, Kifid 2023-0106, Kifid 2023-0094
5	Geef uitspraken uit 2025 waarbij de duur van een EVR-registratie moet worden aangepast	Kifid 2025-0396, Kifid 2025-0365, Kifid 2025-0034, Kifid 2025-0346, Kifid 2025-0325, ECLI:NL:GHARL:2020:3464, ECLI:NL:RBGEL:2020:2971, ECLI:NL:GHARL:2017:10752, ECLI:NL:GHSHE:2016:5577, ECLI:NL:GHARL:2025:1565	Kifid 2025-0254, Kifid 2025-0022, Kifid 2024-0083, ECLI:NL:GHARL:2025:1565, ECLI:NL:GHARL:2025:850
6	Geef uitspraken uit 2024 van de civiele rechter waarbij de interne registraties moeten worden doorgehaald	ECLI:NL:RBGEL:2024:9241, ECLI:NL:RBAMS:2024:1545, ECLI:NL:RBDHA:2024:7064, ECLI:NL:RBAMS:2023:7376, ECLI:NL:RBMNE:2024:4225, ECLI:NL:RBAMS:2024:2881, ECLI:NL:RBAMS:2024:2745, ECLI:NL:RBMNE:2024:3622, ECLI:NL:GHARL:2023:10092, ECLI:NL:RBDHA:2024:6584	ECLI:NL:RBMNE:2024:4122, ECLI:NL:RBMNE:2024:3622, ECLI:NL:RBAMS:2024:2881, ECLI:NL:RBAMS:2024:2633
7	Geef uitspraken van Kifid waarbij de verzekeraar de schadevrije jaren van de consument moest herstellen	Kifid 2024-1124, Kifid 2024-0660, Kifid 2023-0075, Kifid 2022-0871, Kifid 2022-0635, Kifid 2022-0539, Kifid 2022-0151, Kifid 2021-1085, Kifid 2021-1063, Kifid 2021-0979	Kifid 2024-1036, Kifid 2024-0433, Kifid 2023-0746, Kifid 2023-0075, Kifid 2021-1063
8	Geef mij uitspraken van de Rechtbank Amsterdam over spoofing of helpdeskfraude van na 1 januari 2024	ECLI:NL:RBAMS:2024:6094, ECLI:NL:RBAMS:2024:6356, ECLI:NL:RBMNE:2023:6561, ECLI:NL:RBAMS:2025:1751, ECLI:NL:RBAMS:2025:0629, ECLI:NL:RBOBR:2025:2694, ECLI:NL:RBAMS:2024:0441, ECLI:NL:RBMNE:2024:4459, ECLI:NL:GHAMS:2024:2798, ECLI:NL:RBROT:2024:10756	ECLI:NL:RBAMS:2025:1751, ECLI:NL:RBAMS:2025:1498, ECLI:NL:RBAMS:2024:6356, ECLI:NL:RBAMS:2024:6094, ECLI:NL:RBAMS:2024:441

continued on next page

continued from previous page

ID	Query	Responses	Golden Case Responses
9	Geef mij uitspraken van de Geschillencommissie Kifid uit 2024 of 2025 over EVR-registraties naar aanleiding van hypotheekfraude	Kifid 2025-0256, Kifid 2025-0072, Kifid 2024-0530, Kifid 2024-0337, Kifid 2024-0006	Kifid 2025-0427, Kifid 2025-0256, Kifid 2025-0072, Kifid 2024-0709, Kifid 2024-0530
10	Geef mij uitspraken die de Hoge Raad in 2024 en 2025 heeft gedaan op het gebied van verzekeringsrecht waarbij je de artikel 81 RO-zaken niet meeneemt	ECLI:NL:HR:2025:186, ECLI:NL:HR:2024:1797, ECLI:NL:HR:2024:1719, ECLI:NL:HR:2024:1664, ECLI:NL:HR:2024:1545, ECLI:NL:HR:2024:1543, ECLI:NL:HR:2024:1173, ECLI:NL:HR:2024:1160, ECLI:NL:HR:2024:1022, ECLI:NL:HR:2024:726	ECLI:NL:HR:2024:1022, ECLI:NL:HR:2024:1160, ECLI:NL:HR:2025:186, ECLI:NL:HR:2024:258, ECLI:NL:HR:2024:726
11	Geef me alle uitspraken gepubliceerd in de afgelopen twee jaar van de geschillencommissie over een tekortenprocedure	Kifid 2024-0304, Kifid 2023-0222	Kifid 2023-0456, Kifid 2023-0524, Kifid 2023-0536, Kifid 2023-0848, Kifid 2024-0521
12	Zoek recente uitspraken van lagere rechters over advisering over effectenlease zonder vergunning, waarbij de aanbieder aansprakelijk is voor de schade van eiser	(geen resultaten)	ECLI:NL:GHAMS:2024:184, ECLI:NL:GHARL:2024:7752, ECLI:NL:GHARL:2024:7264, ECLI:NL:RBDHA:2025:8561, ECLI:NL:RBDHA:2025:6896
13	Geef mij uitspraken van Kifid over rechtsbijstandverzekeringen waarin de consument een beroep doet op de vrije advocaatkeuze en waarbij DAS Rechtsbijstand de uitvoerder is.	Kifid 2025-0422, Kifid 2023-0317, Kifid 2022-0818, Kifid 2022-0809, Kifid 2022-0284, Kifid 2022-0207, Kifid 2022-0072, Kifid 2021-0042, Kifid 2021-0300, Kifid 2020-908	Kifid 2025-0422, Kifid 2023-0317, Kifid 2022-0809, Kifid 2022-0795, Kifid 2021-0042
14	Geef mij de meest recente uitspraken van de Hoge Raad of Gerechtshoven over overkreditering bij een hypothecaire geldlening.	ECLI:NL:GHDHA:2024:1431, ECLI:NL:GHARL:2024:1293, ECLI:NL:GHAMS:2023:3102, ECLI:NL:GHAMS:2023:2708, ECLI:NL:GHSHE:2023:2087, ECLI:NL:GHSHE:2023:187, ECLI:NL:GHDHA:2023:18, ECLI:NL:GHARL:2022:10863, ECLI:NL:GHARL:2022:7675, ECLI:NL:GHARL:2022:7301	ECLI:NL:GHARL:2024:2235, ECLI:NL:GHARL:2023:5384, ECLI:NL:GHAMS:2023:2708, ECLI:NL:GHSHE:2023:187, ECLI:NL:GHARL:2022:9750
15	Geef mij de meest recente uitspraken van de Commissie van Beroep over een verzekering, niet zijnde een beleggingsverzekering.	Kifid 2025-0023, Kifid 2025-0011, Kifid 2025-0008, Kifid 2024-0085, Kifid 2024-0083, Kifid 2024-0060, Kifid 2024-0058, Kifid 2024-0055, Kifid 2024-0054, Kifid 2024-0041	Kifid 2025-0011, Kifid 2025-0009, Kifid 2024-0083, Kifid 2024-0066, Kifid 2024-0041

continued on next page

continued from previous page

ID	Query	Responses	Golden Case Responses
16	Geef mij conclusies van de advocaat generaal van de Hoge Raad die zien op de zorgplicht van een financieel adviseur.	ECLI:NL:PHR:2019:473, ECLI:NL:PHR:2017:894, ECLI:NL:PHR:2020:1069, ECLI:NL:PHR:2021:161, ECLI:NL:PHR:2020:471, ECLI:NL:PHR:2022:703, ECLI:NL:PHR:2021:1252, ECLI:NL:PHR:2025:516, ECLI:NL:PHR:2025:514, ECLI:NL:PHR:2019:384	ECLI:NL:PHR:2022:703, ECLI:NL:PHR:2022:327, ECLI:NL:PHR:2024:258, ECLI:NL:PHR:2020:688, ECLI:NL:PHR:2014:531
17	Geef mij uitspraken van de rechtbank waarin ambtshalve wordt getoetst aan de kredietwaardigheidstoets	ECLI:NL:RBGEL:2025:123, ECLI:NL:RBGEL:2025:122, ECLI:NL:RBGEL:2025:2626, ECLI:NL:RBGEL:2025:3201, ECLI:NL:RBROT:2024:7351, ECLI:NL:RBGEL:2024:4221, ECLI:NL:RBOVE:2024:5154, ECLI:NL:RBOVE:2024:5155, ECLI:NL:RBMNE:2024:7629, ECLI:NL:RBMNE:2024:6395	ECLI:NL:RBGEL:2025:123, ECLI:NL:RBGEL:2025:122, ECLI:NL:RBGEL:2025:2626, ECLI:NL:RBGEL:2025:3201, ECLI:NL:RBROT:2024:7351
18	Geef mij uitspraken waarin de aansprakelijkheid van de consument wordt beperkt zoals bedoeld in artikel 7:529 lid 2 BW	Kifid 2025-0382, Kifid 2024-1126, Kifid 2025-0018, Kifid 2023-0779, Kifid 2014-074, ECLI:NL:GHSHE:2017:1885, ECLI:NL:RBROT:2015:9378, ECLI:NL:GHAMS:2017:1960, ECLI:NL:HR:2021:749, Kifid 2025-0402	Kifid 2025-0382, Kifid 2024-1126, Kifid 2025-0018, Kifid 2023-0779, Kifid 2014-074
19	Geef me uitspraken vanaf 2023 over overkreditering	ECLI:NL:RBDHA:2024:18875, ECLI:NL:HR:2023:778, ECLI:NL:HR:2022:1945, Kifid 2025-0258, Kifid 2024-0873, Kifid 2023-0691, Kifid 2025-0416, Kifid 2023-0545	ECLI:NL:RBROT:2023:7844, ECLI:NL:RBMNE:2022:6606, ECLI:NL:GHARL:2023:5384, Kifid 2025-0051, Kifid 2024-0234A
20	Geef me uitspraken over een schending van de zorgplicht door een hypotheekadviseur wegens termijnoverschrijding	ECLI:NL:RBGEL:2020:7223, ECLI:NL:RBZWB:2021:5238, ECLI:NL:RBZWB:2024:7286, ECLI:NL:HR:2018:2298, ECLI:NL:RBNHO:2025:16, Kifid 2019-880, Kifid 2018-426, Kifid 2017-629, Kifid 2017-766, Kifid 2017-800	ECLI:NL:RBAMS:2025:2804, Kifid 2024-0600, Kifid 2020-938, Kifid 2019-388, Kifid 2024-0053
21	Geef me uitspraken van Kifid waarin een verstekuitspraak is gedaan (uitspraak zonder dat de FD verweer heeft gevoerd)	Kifid 2025-0433, Kifid 2025-0362, Kifid 2025-0214, Kifid 2025-0240, Kifid 2025-0234, Kifid 2024-0941, Kifid 2022-0860	Kifid 2025-0433, Kifid 2025-0362, Kifid 2025-0214, Kifid 2025-0240, Kifid 2025-0234