Ablating Retrieval Modules for Temporal Conflict Resolution in Legal RAG

Dmytro Asieiev[1⋆], Emilija Bareikaitė[1], Jadrine Kaburu[1], Jonas Urnėžius[1], Aušra Šubonienė[1], and Aistis Raudys[2]

[1] AAI Labs
http://www.aai-labs.com
dmytro@aai-labs.com
[2] Institute of Informatics, Vilnius University

Abstract. Large language models (LLMs) are increasingly applied in domains such as legal, medical, and financial systems, yet they are prone to hallucinations, producing fluent but factually incorrect statements. This issue is particularly problematic in legal contexts, where strict requirements for factual accuracy and reliable citations exist. Retrieval-augmented generation (RAG) has been proposed as a solution to mitigate hallucinations by grounding model outputs in external documents; however, factual errors still occur when retrieved sources are conflicting, outdated, or irrelevant.

This paper investigates how different retrieval strategies influence factual correctness in legal question answering systems based on RAG. Ten variants of a legal RAG pipeline were evaluated using an ablation study on a dataset of 200 EurLex queries. The experiments compare dense and hybrid retrieval, metadata filtering, query expansion, and re-ranking techniques, and assess their impact using metrics such as Recall@5, non-answer rate, citation consistency, outdated retrieval rate, and cost per query.

The results show that hybrid retrieval substantially improves recall compared to dense retrieval, while metadata filtering reduces outdated results but can significantly lower recall when applied aggressively. Query expansion and re-ranking provide only limited improvements when retrieval quality is poor. Overall, the findings demonstrate that retrieval quality is the primary factor influencing the performance of legal RAG systems, highlighting the importance of optimizing retrieval strategies rather than relying solely on complex generation or reasoning mechanisms.

Keywords: retrieval-augmented generation · legal question answering · temporal conflict resolution · information retrieval

1 Introduction

In recent years, large language models (LLMs) have gained increasing attention for applications in various spheres, for instance, in enterprise, legal, medical and

⋆ Corresponding author.

financial systems [1]. However, LLMs hallucinate, meaning they produce fluent, but factually incorrect text, which is especially alarming in legal systems which impose strict requirements. There are various possible solutions for this serious issue, one of which being implementing retrieval-augmented generation in the models [5, 3]. However, using RAG does not guarantee that there will be no factual incorrectness in the generated text.

Empirical studies have demonstrated that the majority of errors for factual mistakes in text generation using RAG come from the retrieval process [11, 8]. In RAG, the retrieval component assumes that the retrieved documents can be added together into one best answer, however, in practice, the retrieved information conflicts and is inconsistent. This leads to generating factually incorrect text, which is unacceptable in legal corpora, as it leads to serious damage and consequences.

Most research on this topic is about finding ways on improving successful information retrieval using various techniques (re-rankers and chain-of-thought reasoning) [4, 10, 9]. However these studies mainly improve how we select context and reason. They do not pay enough attention to identifying and resolving contradictions when different sources have conflicting information, which is often the case in legal context.

Our paper aims to study how RAG systems work in the legal field when there are conflicting statements due to changes over time when regulations or decisions are updated or replaced. We created a set of questions to test the system and see how different ways of finding information affect the answers given and if wrong information shows up due to conflicts. The main contributions of our paper are:

1. An evaluation on how different retrieval strategies affect factual correctness in legal question answering and how they interact with temporal data filtering,
2. An analysis of failures caused by conflicting legal documents,
3. Practical design advice for legal RAG systems based on the performed analysis.

Hypotheses: We test three predictions about legal RAG performance:

1. **H1:** Metadata filtering + hybrid retrieval reduces temporal conflicts (CRR/ORR) >50% vs naive RAG—simple metadata beats complex reasoning.
2. **H2:** Retrieval quality dominates generation errors.
3. **H3:** Hybrid retrieval provides highest ROI: maximum Recall@5 gains at minimum cost increase.

2 Retrieval Strategies

Retrieval strategies determine how relevant documents are selected from the knowledge base before being provided to the language model. In this section, we present several retrieval mechanisms commonly used in RAG systems, including sparse, dense, and hybrid retrieval approaches. Additionally, we examine

complementary techniques such as query rewriting, metadata filtering, and reranking that can improve retrieval quality and reduce temporal conflicts in legal documents.

2.1 Sparse Retrieval

Sparse retrieval algorithms are based on keyword matching between queries and documents. Traditional sparse retrieval algorithms like BM25 estimate document relevance based on the number of keyword matches. The top documents are then sent to the LLM.

2.2 Dense Retrieval

Dense retrieval uses neural embedding models to convert user queries and available documents into vectors. After such transformation, texts with similar semantic meaning are positioned closely together. We use FAISS for efficient similarity search. In our implementation, document chunks are embedded and stored in the FAISS index during the indexing phase. At retrieval time, user query is embedded using the same method and matched against the vectors to return the top-k most relevant documents.

2.3 Hybrid Retrieval

Hybrid retrieval combines the two previous approaches and executes both of them in parallel. It then merges the results using a fusion algorithm. This method combines keyword matching with semantic similarity search.

We perform hybrid retrieval by concurrently performing dense vector search with FAISS and sparse keyword search with the BM25 algorithm. To combine the two, we use a weighted sum fusion technique that normalizes results based on defined weights. In this case, we assign equal weights of 50% to dense and sparse search algorithms.

2.4 Metadata Filtering

Metadata filtering can improve RAG systems by refining search queries. Leveraging this technique, RAG system adds selected metadata to a search query (e.g., date, status, etc.) filtering out irrelevant documents early in the search process. Thus, it is an irreplaceable tool in large and complicated datasets.

In our experiments, two metadata filters are evaluated. The first is a temporal filter that restricts retrieval to documents published after a specified year (e.g., date $>$ 2010). The second is a legal status filter that limits results to documents marked as *In Force*, ensuring that only currently valid legal acts are considered during retrieval.

2.5 Re-Ranking

Re-ranking in RAG system means reordering a retrieved set of documents according to their relevance to the original query. This step helps to improve the quality of information that the LLM receives to generate a response.

We use the Cross-Encoder model from the Sentence Transformers library. Unlike other retrieval models, which compute embeddings for queries and documents separately, the Cross-Encoder computes the embedding for the query-document pair in a single forward pass. This allows the model to capture more complex semantic relationships that are not captured by embedding-based similarity metrics.

The re-ranking procedure has three steps. Firstly, each retrieved document pairs with the original query and passes through the Cross-Encoder, which produces a refined relevance score. Secondly, candidate documents sort in descending order by these scores. Lastly, the reordered list and updated relevance values replace the initial retrieval ranking.

3 Experimental Scope & Data Preparation

This section describes the experimental scope of the study and the preparation of the data used in our experiments. First, we explain how the dataset was constructed and how the final subset of legal acts, queries, and ground-truth annotations were created. Next, we present the chunking strategy and the embedding and indexing procedures used to build the retrieval infrastructure. Finally, we describe the experimental setup by distinguishing between fixed and variable components of the RAG pipeline.

3.1 Dataset Construction

For this paper, the EurLex dataset was used [6]. The dataset contains 142036 legal European Union acts passed between 1952–2019. Each law has a CELEX, various legal metadata, such as act type, date, status, attributes for amendment relationships and the full law's text.

The first stages of the dataset construction involved downloading the full dataset, preprocessing it and performing an initial inspection. The preprocessing process included normalizing CELEX identifiers for each law, converting the date format and parsing amendment chains. As for the initial inspection, its purpose was to investigate five chosen acts and see where conflicts between them arise. The analysis demonstrated that CELEX identifiers with numeric suffixes are not versions or amendments of the same law – thus to find amendments or conflicts between legal decisions, we had to focus on explicit amendment relationships and other temporal legal metadata.

After completing working on the full EurLex dataset, we moved on to forming the final data subset. The subset was created using metadata filters and a selection process involving different steps. In the end, the final dataset was

constructed, which contains 555 legal acts, where 231 acts are acts with amendments, 229 acts that are expired and 326 active acts.

The last step of the dataset construction process was creating the query and ground-truth sets. For the query set, 100 legal acts were randomly selected from the previously created subset of 555 documents while making sure that each chosen legal decision comes only from one of the four buckets: acts with amendments, expired acts, active acts and legal acts in general. Then for each law a factual and temporal question was written - this formed the basis of the query set. As for the ground-truth dataset, we took the created questions from the query set and assigned to each the corresponding law's CELEX, the expected date range, and the difficulty level (easy, medium or hard).

3.2 Chunking Strategy

The chunking strategy for this paper involved chunking only the created EurLex subset of 555 legal documents. The text was split by sentences into chunks of 512 tokens, with a 32-token overlap between chunks. Each chunk captures all relevant information associated with the legal acts – CELEX, the act's status, publication date, etc. Chunks are stored in dense FAISS embeddings and additionally in a sparse BM25 index, where hybrid retrieval is chosen (see section 3.4).

3.3 Embedding and Index Construction

For this paper we have two choices for retrieval type - dense or hybrid retrieval. For dense retrieval, each chunk is embedded using the OpenAI's text-embedding-3-small model, which produces 1536-dimensional vectors. These vectors are then stored in a FAISS IndexFlatL2 index, which supports efficient nearest-neighbor search over high-dimensional embeddings [2]. We save the constructed index to disk and we have the choice to rebuilt it from the same dataset if needed. In the cases where hybrid retrieval is used, we combine the FAISS index with a sparse BM25 index built over the same chunks. BM25 uses token-level term matching and supports optional top-k filtering, while FAISS provides semantic similarity [7].

3.4 Fixed vs. Variable Experimental Components

Since our paper performs different experiments in order to investigate how the RAG system works when conflicts in documents are present, this setup leads to some components being fixed and some – variable. Fixed elements remain constant across all performed experiments: the dataset, overall RAG pipeline architecture created using LlamaIndex, chunking parameters, index structures, the embedding model and the language model. While the variable components are those which differ depending on the different defined experiment variants. Variable components include retrieval type (dense or hybrid), top-k retrieval, metadata filtering (constraints on document date), query expansion, and re-ranking using a cross-encoder model. Query expansion has two possible choices

in our setup – manual rewriting or HyDe-generated reformulations. The detailed setup for all experiment variants can be found in Table 1.

Table 1. Experiment matrix

ID	Retrieval	Top-k	Query Expansion	Filters	Re-ranking
V0	Dense	5	None	None	No
V1	Dense	10	None	None	No
V2	Hybrid	5	None	None	No
V3	Dense	5	None	Date (>2010)	No
V4	Hybrid	5	None	Status (In Force)	No
V5	Dense	5	Manual rewrite	None	No
V6	Dense	5	HyDE	None	No
V7	Dense	5	None	None	Yes
V8	Hybrid	5	None	None	Yes
V9	Hybrid	5	HyDE	Date (>2010)	Yes

4 Evaluation Methodology

In this paper, for effective evaluation of how RAG deals with legal data, where conflicts or law amendments are present, we focus on both the retrieval and generation phases of the pipeline. We aim to identify retrieval strategies that reduce outdated answers, ensure that answers are complete and avoid introducing new reasoning failures. In order to achieve this goal, both retrieval and generation-level metrics are tracked and later analyzed. In addition, we discuss cost and latency considerations in order to understand the computational efficiency of the evaluated pipelines. Finally, we outline the reproducibility guarantees applied in our experiments to ensure that results can be consistently reproduced.

4.1 Retrieval-Level Metrics

The retrieval-level metrics focus on the documents returned by the RAG system before the answer generation process. Since our work focuses on version conflicts and temporal misalignments, the chosen metrics indicate which regulatory versions are retrieved and what temporal information the model can use when generating the outcome. The metrics we decided to use in our paper are conflict retrieval rate (CRR), outdated retrieval rate (ORR), conflict density (CD) and latest version recall (LVR).

CRR measures the proportion of retrieved CELEXs that participate in at least one amendment relationship within the retrieved set. CRR is calculated by the following formula:

$$CRR = \frac{C}{R} \tag{1}$$

where C is the subset of retrieved regulations that participate in at least one amendment relationship with another regulation in the retrieved set R. Higher CRR means that more of the retrieved documents are linked through amendments, so the model is more likely to encounter different versions of related regulations. On the other hand, lower CRR means the retrieved set focuses on fewer related versions, making the context easier to interpret.

CD shows how many of the retrieved regulations are connected to each other through amendments, compared to the overall size of the retrieved set. To measure the CD metric we use the formula:

$$CD = \frac{P}{R} \tag{2}$$

where P is the set of retrieved regulations pairs that are connected by amendments. A high CD value means that many of the retrieved regulations are linked to each other by amendments compared to the overall number of retrieved documents. Meanwhile, a low CD value shows that few amendment relationships exist among retrieved documents.

ORR measures the proportion of retrieved CELEX identifiers that are considered outdated. A CELEX is considered outdated if its status metadata marks it as such or if its valid time period ended before the reference date. The ORR is computed with the formula below:

$$ORR = \frac{O}{R} \tag{3}$$

where O is the set of retrieved outdated CELEX identifiers. A high ORR value leads to the fact that a large share of retrieved documents are no longer valid at the reference date. In contrast, a low ORR means that most of the retrieved documents are still valid.

LVR measures whether the latest version of a regulation is present in the retrieved set. We calculate this metric in two ways: the first situation assumes that the relevant CELEX is known. Given this case, we calculate LVR by the following formula:

$$LVR = \begin{cases} 1, & \text{if latest}(c^*) \in R \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where c^* is the expected CELEX, latest(c^*) is the latest version in the knowledge base and R is the set of retrieved CELEX identifiers. If LVR equals to 1 that means that the system retrieved the most recent version, while 0 indicates that the latest regulation was not retrieved and generating a factually accurate answer is unlikely.

The second situation assumes that the relevant CELEX is not given, then to measure LVR we use the formula:

$$LVR = \frac{|\{\, c \in R \mid \text{latest}(c) \in R \,\}|}{|R|} \tag{5}$$

If this formula is used then a high LVR value means that most retrieved CELEX's include their latest versions and a low LVR indicates that the retrieved documents often lack their most recent version.

4.2 Generation-Level Metrics

Generation-level metrics measure how the model behaves when given the retrieved context. They are analyzed to understand how the model reasons and generates the final output when conflict is present in the given information. The generation-level metrics we chose to track in our paper are answer temporal alignment (ATA), citation consistency (CC), hedging behavior (HB) and answer completeness (AC).

ATA classifies which regulation version the answer reflects. We calculate this metric by first identifying all the regulations mentioned in the generated answer and then comparing each one to the most recent version in our knowledge base. Based on the results, we classify the generated output as latest, outdated, mixed or unclear. A latest output demonstrates that the model used up-to-date information to generate its response. On the other hand, answers which are labeled outdated or mixed indicate that the model relied on old or conflicting information.

CC measures whether the citations in the final answer are up-to-date and correctly match the documents that were retrieved. We calculate CC by first listing all regulations cited in the answer, then comparing these citations to the set of retrieved documents to see if they were actually provided and checking if multiple versions of the same regulation are cited. Lastly, the answer is classified into four of the available classes: clean, conflicting (multiple versions of the same regulation are cited), outdated-only and no-citation. An output labeled as clean means all citations are current and appear in the retrieved documents. Conflicting, outdated and ungrounded outputs suggest the answer could be misleading.

HB measures the certainty of the model's produced answer. To calculate this metric we search for words or phrases that indicate uncertainty, such as, "might," "could," "typically", or conditional statements ("if", "depends on"). Then using the extracted information, the outputs are classified as confident, hedged and conditional (valid only under certain circumstances). AC measures how fully the answer addresses the question, regardless of whether it is correct. We measure this metric by counting the words in the answer and compare to predefined minimum thresholds. Next, we look at the important terms in the question and see how many of them are mentioned in the answer. Lastly, based on the previous steps, the produced output is classified as non-answer, partial and complete. Complete outputs fully answer the query, partial outputs – address only some parts and non-answers are insufficient.

5 Results

This section presents the experimental evaluation of the proposed legal RAG system variants. First, we analyze the impact of individual system components

through a series of ablation experiments that isolate retrieval strategies, filtering mechanisms, query expansion, and re-ranking techniques. We then perform a failure analysis to better understand the sources of incorrect outputs and distinguish between retrieval-bound and generation-bound errors. Together, these analyses provide insight into how different design choices influence the factual correctness and reliability of legal RAG systems.

5.1 Component Ablation Experiments

Ten different variants of the study, from V0 to V9, were used, each modifying a single component to isolate its effect. Each variant was evaluated on a similar corpus of 200 EurLex legal queries (except V5: 198 queries), and the metrics measured were Recall@5, non-answer rate, average coverage, citation consistency, ungrounded citation rate, outdated retrieval rate, and cost per query.

Additionally it was seen that switching from dense-only retrieval (V0) to hybrid retrieval (V2), which combines dense embeddings with BM25, resulted in a substantial performance increase. Recall@5 improved from 1.0% to 82.5%, the non-answer rate dropped to 0.5%, and coverage increased to 90.8%, while ungrounded citation rates remained similar (see Table 2, Figure 1).

Table 2. Retrieval strategy ablation (dense vs hybrid)

Metric	V0	V2	Delta
Recall@5	1.0%	82.5%	+81.5 pp
Non-answer rate	6.5%	0.5%	-6.0 pp
Avg. coverage	81.4%	90.8%	+9.4 pp
Ungrounded citation rate	13.5%	14.0%	+0.5 pp
Cost per query (USD)	0.0232	0.0255	+10%

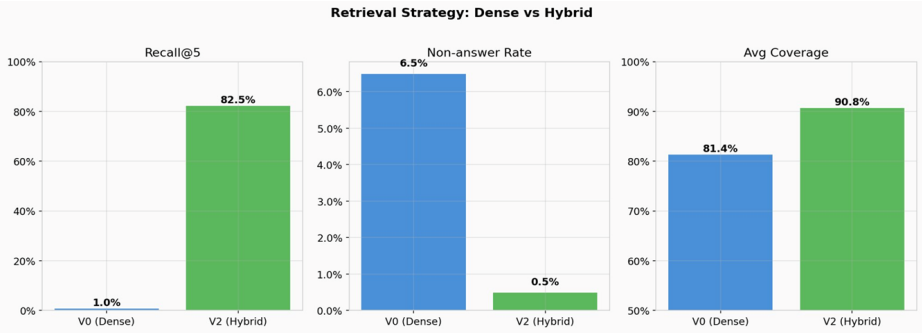

Fig. 1. Retrieval strategy comparison

Applying metadata filters demonstrates significant trade-offs between recall, outdated retrievals, and clean citations. Dense retrieval with a date filter (V3) reduced outdated retrievals but nearly eliminated recall (0.5%) and doubled the ungrounded citation rate to 39.5%, with no clean citations. Hybrid retrieval with a status filter (V4) eliminated outdated retrievals entirely and resulted in the highest number of clean citations (107 out of 200), but reduced Recall@5 to 43.5% due to the exclusion of ground-truth documents outside the filter window (see Table 3, Figure 2).

Table 3. Metadata filtering ablation

Metric	V0	V3	V2	V4
Recall@5	1.0%	0.5%	82.5%	43.5%
Non-answer rate	6.5%	3.5%	0.5%	4.5%
Outdated retrieval rate	60.4%	32.0%	56.1%	0.0%
Ungrounded citation rate	13.5%	39.5%	14.0%	18.5%
Clean citation count	16/200	0/200	58/200	107/200
Cost per query (USD)	0.0232	0.0075	0.0255	0.0133

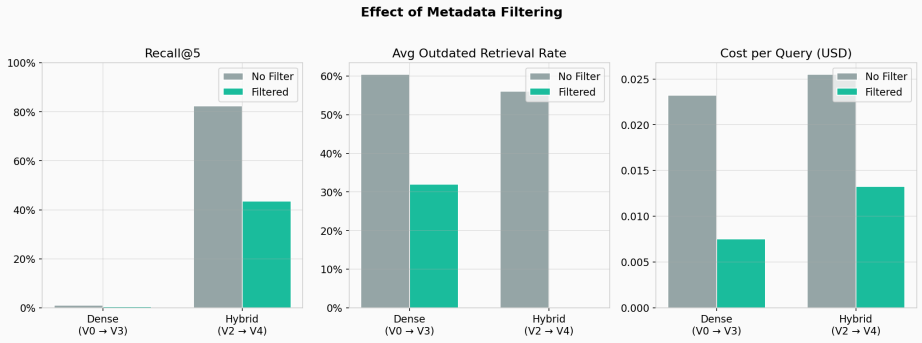

Fig. 2. Impacts of metadata filtering

The complete pipeline (V9) achieved a Recall@5 of 34.0%, reduced cost per query by 28%, but increased the ungrounded citation rate to 25.0% (see Table 4, Figure **??**). Overall, the experimentation proved that hybrid retrieval combined with re-ranking provides the most effective balance between recall, cost, and citation quality.

Table 4. Combined pipeline vs naive baseline (V9 vs V0)

Metric	V0	V9	Delta
Recall@5	1.0%	34.0%	+33.0 pp
Non-answer rate	6.5%	2.5%	-4.0 pp
Avg. coverage	81.4%	87.8%	+6.4 pp
Ungrounded citation rate	13.5%	25.0%	+11.5 pp
Outdated retrieval rate	60.4%	49.2%	-11.2 pp
Cost per query (USD)	$0.0232	$0.0168	-28%
HyDE generation cost (total)	None	$0.9952	–

5.2 Failure Analysis

This section is dedicated for investigating the seen failures in our performed experiments across different RAG variants. We classify the observed failures as retrieval-bound and generation-bound. Retrieval-bound errors occur when the system fails to find the correct or latest law, while generation-bound failures happen when the correct law is retrieved, but the answer is incomplete, temporally misaligned, or contains citation mistakes. If neither of these errors are detected for an output, the answer can be classified as "success" or "partial success". Successful answers require retrieving the latest law with fully correct, complete, and properly cited laws, while outputs that are mostly correct but not perfect are labeled as partial success. The detailed requirements for these four answer classes can be found in Table 5.

Table 5. Requirements for failure type classification

Answer Type	Requirements
Retrieval-bound failure	$LVR = 0$ or ($ORR = 1$ and $ATA \in \{\text{Ambiguous/Unclear}, \text{Outdated}\}$)
Generation-bound failure	$LVR = 1$ and ($AC =$ Non-answer or $ATA \neq$ Latest or $CC \neq$ Clean)
Success	$LVR = 1$ and $ATA =$ Latest and $CC =$ Clean and $AC =$ Complete
Partial success	Any answer not meeting the strict success criteria but not classified as retrieval- or generation-bound failure.

Looking at the answer type distributions per variants (see Figure 3 and Table 6), we can see that the worst performing variants are V0, V1, V3, V5, V6 and V7, as they do not produce any successful or partially successful answers and all their outputs are classified as retrieval errors. The rest variants show more promising results, with V4 demonstrating the best result (successful answers outnumber the rest of the categories). Additionally, it can be seen that there are no cases where generation-bound failures exceed retrieval-bound errors. The

mean CRR and ORR values seen in the table reinforce these findings. Variants dominated by retrieval failures tend to show relatively high ORR values, which indicate frequent reliance on outdated information. In contrast, V4, the strongest variant, has both mean CRR and mean ORR equal to zero, suggesting that it effectively avoids conflicting and outdated retrieval.

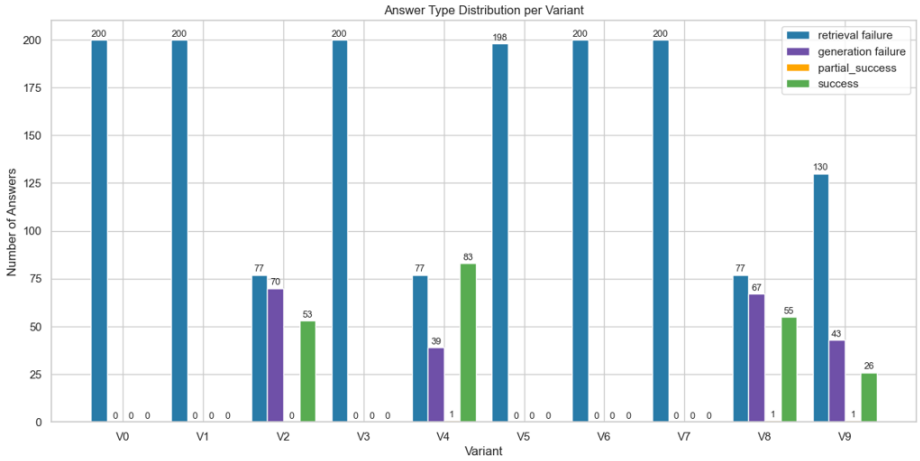

Fig. 3. Answer type distribution per variant

Table 6. Failure results summary per variant

Variant	Failure %	Retrieval-bound %	Generation-bound %	Mean CRR	Mean ORR
V0	100.0	100.0	0.0	0.00	0.60
V1	100.0	100.0	0.0	0.02	0.56
V2	73.5	38.5	35.0	0.18	0.56
V3	100.0	100.0	0.0	0.00	0.32
V4	58.0	38.5	19.5	0.00	0.00
V5	100.0	100.0	0.0	0.01	0.55
V6	100.0	100.0	0.0	0.02	0.66
V7	100.0	100.0	0.0	0.00	0.60
V8	72.0	38.5	33.5	0.18	0.56
V9	86.5	65.0	21.5	0.18	0.49

6 Conclusion

In this paper, an exhaustive evaluation of ten different variants of the legal RAG system was successfully performed to analyze the efficiency of different retrieval strategies and the accuracy of their generated answers. The effectiveness is assessed using latest version recall, conflict retrieval rate, outdated retrieval rate, answer temporal alignment, citation consistency, answer completeness, and the cost of the operation, among other factors. The RAG system's variants are evaluated using 200 created queries from EurLex subset.

These ablation experiments offered useful insights into the performance of the legal RAG system:

- Increasing the query window size by doubling the value of k, changing it from 5 to 10, slightly increased non-answer rates, but the recall rates remained the same, and the costs almost doubled.
- Hybrid retrieval had the best overall results, with the Recall@5 value increasing from 1.0% to 82.5%, while also reducing non-answer rates and improving answer coverage at a relatively low cost.
- Metadata filters were effective in reducing the number of outdated results, although overly selective filters reduced the overall recall and non-answer rates, particularly with historically framed queries. Notably, some configurations (e.g., V4) demonstrated that lower recall can still lead to better temporal correctness and citation quality.
- Query expansion, both manual and using HyDE, had little effect on overall recall, although the HyDE results improved citation rates with dense-only approaches.
- Re-ranking with cross-encoders showed some positive results, improving non-answer and citation rates, but these results are not sufficient to mask the failure of the first and most important stage of the overall system.
- Using all of the components (in this study, version 9), the Recall@5 value was increased from 1.0% to 34.0%, with reduced costs per query, although recall rates were reduced due to strict application of temporal filters compared to the application of hybrid retrieval without aggressive filters.

Furthermore, the failure analysis shows that most errors (78%) are retrieval-bound, mainly due to not retrieving the latest version of a regulation (LVR = 0). This suggests that retrieving correct and up-to-date documents is more important than recall alone. In addition, higher recall can still include outdated results (high ORR), so it should be considered together with temporal and citation-based metrics.

In summary, the results show that having correct metadata brings the most to the table in terms of mitigating retrieval failures and temporal conflicts. However, we understand that this is not always possible given large unstructured datasets, which are commonly used in knowledge bases for RAG systems.

- For these cases, we recommend sticking to hybrid as the default approach, as it provides the best trade-off between recall, temporal correctness, and overall answer reliability.

– We also advise against using re-ranking and HyDe techniques until the retrieval-level issues are addressed, as they improve citation and non-answer metrics but do not improve retrieval correctness.

References

1. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33. pp. 1877–1901 (2020), https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
2. Faiss Contributors: Faiss documentation. https://faiss.ai/ (2026), documentation site, accessed 2026-04-10
3. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey. CoRR **abs/2312.10997** (2023). https://doi.org/10.48550/arXiv.2312.10997, https://arxiv.org/abs/2312.10997
4. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.550, https://aclanthology.org/2020.emnlp-main.550/
5. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems 33. pp. 9459–9474 (2020), https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
6. puskas78: EUR-Lex dataset. https://www.kaggle.com/datasets/puskas78/eurlex-dataset (2025), kaggle dataset page, accessed 2026-04-10
7. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009). https://doi.org/10.1561/1500000019, https://www.nowpublishers.com/article/Details/INR-019
8. Vach, M., Gliem, M., Weiss, D., Ivan, V.L., Hauke, F., Boschenriedter, C., Rubbert, C., Caspers, J.: Evaluating retrieval augmented generation-enhanced large language models for question answering on german neurovascular guidelines. Clinical Neuroradiology (2025). https://doi.org/10.1007/s00062-025-01562-z, https://link.springer.com/article/10.1007/s00062-025-01562-z, published online 2 September 2025
9. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems 35. pp. 24824–24837 (2022),

https://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
10. Xu, J., Luo, S., Chen, X., Huang, H., Hou, H., Song, L.: RALL-Rec: Improving retrieval augmented large language model recommendation with representation learning. In: Companion Proceedings of the ACM Web Conference 2025. pp. 1436–1440. Association for Computing Machinery (2025). https://doi.org/10.1145/3701716.3715508, https://dl.acm.org/doi/10.1145/3701716.3715508
11. Zhang, W., Zhang, J.: Hallucination mitigation for retrieval-augmented large language models: A review. Mathematics **13**(5), 856 (2025). https://doi.org/10.3390/math13050856, https://www.mdpi.com/2227-7390/13/5/856