From Statement of Facts to Statutory Provisions – Efficient Retrieval of Relevant Legislation

Aleksander Smywiński-Pohl[1 [0000−0001−6684−0748]],

Magdalena Król[1 [0000−0003−0392−0921]], and Piotr Mirosław[1]

AGH University of Krakow, Kraków, Poland

Abstract. Legal question-answering systems play a crucial role in enhancing access to justice by providing both citizens and legal professionals with accurate interpretations of the law. However, existing AI-based legal models struggle with processing layperson inputs, mapping them to formal legal language, and ensuring robustness across different languages and jurisdictions. This study explores the use of retrieval-augmented generation (RAG) systems to address Polish legal queries, with a focus on improving information retrieval components. We evaluate several pretrained retrieval models across four legal datasets to identify the most effective architectures. Subsequently, we fine-tune the best-performing models to determine which dataset types yield the greatest improvements when addressing lay-language questions. Our results demonstrate that fine-tuning on provision-based datasets significantly enhances retrieval accuracy and contextual relevance. Conversely, datasets with high lexical overlap between questions and provisions offer limited benefit when models are applied to layperson inputs—challenging the common practice of using large language models to generate training questions. In response, we propose a novel dataset construction method based on legal judgments, which performs nearly as well as manually annotated datasets containing layperson queries.

Keywords: legal question-answering, retrieval-augmented generation, information retrieval, fine-tuning, legal AI

1 Introduction

Legal questions play a pivotal role in the legal system by enabling both citizens and legal professionals to comprehend and interpret legal provisions. They facilitate access to legal information and promote accurate interpretation, thus assisting in the resolution of legal disputes. By bridging real-world problems with legal terminology, datasets containing layman questions contribute to the democratization of legal assistance. Traditionally, the ability to answer legal questions and provide accurate interpretations has been the domain of lawyers and judges. However, the advent of intelligent systems capable of addressing and resolving legal inquiries has become increasingly prominent.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

Despite the availability of numerous tools that are able to retrieve legal provisions and generate content based on them, a significant gap persists in the accessibility and reliability of tools designed to assist laypeople navigate legal regulations. Specifically, there is a need for systems that can assign an appropriate legal interpretation to an individual description of facts, particularly in languages other than English and within legal frameworks distinct from the American legal system.

Lay language differs significantly from formal legal terminology, often being ambiguous and subjective. For example, a person might say,"Someone hit me in the face," whereas legal classification could align with "A violation of bodily integrity occurred" (Polish Penal Code, Article 217 § 1) or "Battery" (Polish Penal Code, Article 158 § 1). A legal AI system should recognize such descriptions and refine them through user interaction, leading to accurate legal classification.

Key factors include alignment with legal provisions, the currency of legal norms, and a verifiable legal basis. Large Language Models (LLMs) are increasingly used for question answering, yet "existing research indicates that this approach may be misleading due to the potential for inaccuracies and the dynamic nature of legal norms" [8], [2]. Given the evolving legal landscape, a "robust, reliable system grounded in a solid retrieval framework, such as the Retrieval-Augmented Generation (RAG) System," is essential [33], [17].

2 Related Work

Retrieval systems are an efficient way to retain information with substantialy reduced risk of hallucinations [26]. Retrieval systems in law are becoming more and more popular [25], and the need for accurate retrieval is widely recognized as essential [27], [9]. Recent advancements such as Snowflake Arctic have significantly boosted retrieval capabilities, particularly in handling complex queries and maintaining high accuracy in large-scale retrieval tasks [18]. These advances [19] demonstrate the increasing capability of retrieval models to handle complex queries while maintaining high performance and reliability.

Polish retrieval has advanced significantly with fine-tuned models like `stella-pl` [7], `mmlw-retrieval-roberta-large` [7], and `ipipan/silver-retriever-base-v1.1` [24], demonstrating state-of-the-art performance in Polish information retrieval. The Polish Information Retrieval Benchmark (PIRB) provides a robust evaluation framework, assessing dense and sparse retrieval methods through knowledge distillation and hybrid approaches [7].

In legal retrieval, systems like LawPal employ retrieval-augmented generation (RAG) with FAISS-based vector search, improving the accuracy and accessibility of legal information [10]. Fine-tuning large language models on specialized datasets further enhances performance, with research showing that scaling task diversity, increasing model size, and incorporating chain-of-thought (CoT) reasoning significantly improve generalization [5,32]. Studies on PaLM [28] and T5 [21] confirm these benefits across various NLP benchmarks [5]. These advance-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

ments have shown promise when applied specifically to complex legal queries and contexts.

Applying fine-tuning techniques to legal NLP models can enhance their ability to interpret complex legal queries, align layperson descriptions with appropriate legal provisions, and provide more accurate legal recommendations. By incorporating diverse legal datasets and structured instructional training, fine-tuned models can offer more precise and context-aware legal interpretations, ultimately advancing the state of AI-driven legal assistance. An interesting question is whether retrieval performance varies significantly between dataset types. A large gap would indicate that layperson questions require specialized handling and challenge the assumption that automatically generated QA datasets, which often share vocabulary with the target text, are sufficient [3,11].

3 Research Questions

The goal of this research is primarily geared towards improving access to justice by employing information retrieval system. We want to achieve this goal by evaluating and potentially improving the performance of such system with respect to questions imposed by laymen. To achieve this goal we have formulated the following research questions.

3.1 RQ1: Does Question Answering for Laymen Differ from Professionals?

We investigate whether retrieval model performance differs between datasets with questions from legal professionals and those from laymen. Using four datasets (detailed in Section 4), we assess their effectiveness in retrieving the most relevant legal provision. If no difference exists, existing retrieval models should perform well across all datasets.

Two datasets (*LQUAD-PL*, *Simple Legal QA*) were created similarly to automated QA generation, with questions derived directly from legal provisions, reflecting professional legal language. The other two (*Lemkin questions*, *Rulings questions*) were generated more naturally—either posed by laymen or extracted from court judgments.

The key question is whether retrieval performance varies significantly between these dataset types. A large gap would indicate that laymen's questions require specialized handling and challenge the assumption that automatically generated QA datasets, which often share vocabulary with the target text, are sufficient.

3.2 RQ2: Which Models Are Best Suited for Improving Access to Justice in Polish?

We are testing a large number of models of moderate size (no larger than 1.5B parameters) which occupy leading positions on English, Multilingual and Polish

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

information retrieval benchmarks. The second question is about the differences between the models on the different datasets. Will only one model appear the best on all of them? What is the impact of language specific fine-tuning? Is there a huge discrepancy between English-only, multilingual and Polish models? These questions are all related to this general RQ.

3.3 RQ3: Can the Access to Justice be Improved via Fine-tuning?

In the last question we want to find out to what extent we can improve the results obtained in RQ2 via fine-tuning. We are mostly concerned with the performance on the *Lemkin questions* dataset, since these are questions imposed by laymen. When building a legal information systems for real people, we can expect that the questions will resemble them the most. So we are most eager to improve the retrieval performance specifically on it. We will test different scenarios where we use the data from the remaining datasets (except *Simple Legal Questions*, since it is too small), to find out if there is any help from them. As a result, during fine-tuning, we evaluate performance on the validation subset of *Lemkin questions* to select the model best aligned with this dataset—ultimately prioritizing the model most effective for improving access to justice.

As a result, during fine-tuning we will measure the performance on the validation subset of *Lemkin questions*, to pick the model best suited for this dataset, so a model best suited for improving access to justice.

4 Approach to Legal Information Retrieval

In this section we present the datasets and the models used in our research. In the first part we present four legal datasets targeting Polish language, of which three were not yet revealed publicly. In the second part we discuss the information retrieval models that are tested on these datasets.

4.1 Datasets

Lemkin Legal Questions – Laymen Questions Lemkin Legal Questions (short: *Lemkin Questions*) is a dataset of questions and manually annotated provisions collected during the development of *Lemkin – Intelligent Legal Information System*. The questions were collected from real users during the public availability of the system at `https://lemkin.pl`. The dataset contains: 333 questions in the development subset, 2471 in the train subset and 327 in the test subset. The questions are very diverse in nature; they include very short and pretty long questions (sometimes resembling factual descriptions introduced in the second dataset), they include spelling errors, unusual spelling (like replacing spaces with dots), etc.

The system presents legal provisions in response to questions, using only content from Polish statutory law. Each bill was divided by articles, and relevant passages were manually annotated by law students. For each question, a query

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

was sent to ElasticSearch (using BM25 retrieval model [23]), and the top-10 results were annotated by two annotators. Conflicts were resolved by a superannotator. The dataset contains over 30,000 question-passage pairs with scores of 0 (irrelevant) and 1 (relevant), based on 8383 annotated provisions.

Court Rulings The second corpus is a novel dataset, automatically extracted from court rulings (short: *Rulings Questions*). These are not questions in the linguistic sense, but descriptions of the important facts of the case that appear at the beginning of each judgment (factual situation). This description usually contains sentences that are easier to understand by a layperson, than the rest of a judgment, since the rapporteur judge usually describes the facts of the case using the statements of the eye witnesses, excerpts of the documents and similar sources. Still, it is prepared by a legal professional, so it will not contain colloquial speech, imprecise terms, and other linguistic phenomena typical for an everyday language.

The dataset creation process included the following assumptions:

- Each ruling contains a factual description section, which presents the case in language closely resembling layman speech.
- Each ruling references the legal provisions upon which the case is interpreted.

As a result the dataset was created by pairing the factual description with the most relevant legal provision appearing in the judgment (the provision that is referenced the largest number of times).

Polish court rulings are accessible through various platforms, such as SAOS [12]. The API provided by the service's creators allows for downloading of large volumes of cases using HTTP requests. The platform catalogs judgments from 1986 until the end of 2023, totaling over 480,000 documents as of August 12, 2024.

To extract the factual descriptions, we developed a model using annotations from the Lemkin project. Initially, we treated this task as a token classification problem, but such an approach struggled with identifying section endings. We then switched to sentence classification, using Stanza [20] to split text into sentences. This approach improved performance, achieving nearly 70% F1-score by identifying only section beginnings, as final sentences varied too much for reliable classification. Finally we have decided to extract the factual description as a consecutive sequence of sentences, from the first detected sentence up to 7 sentences, if no new beginning was detected within the next 6 sentences. The approach was applied to approx. 10 thousand documents, even though a much larger corpus could be easily extractred.

For legal provision retrieval, we structured the data as triples:

1. a *factual description* from the document as a **query**,
2. the provision with the *largest number of references* in the document as a **positive example**,
3. the provision with *zero references* in the document *most similar* to the positive example as a **negative example**.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

The dataset comprised 2545 provisions and was split into training (5733 queries), development (1967 queries), and test (2004 queries) sets.

Legal Trainee Exam Questions – LQUAD-PL The tests for the legal trainees in Poland are available publicly. Each year they have almost the same structure: each group of trainees (attorneys, notaries and bailiffs) receives a set of 150 questions[1] with three choices, with only one valid answer. The answer sheet contains the indication of the valid answer together with its legal base. Both questions and answers are distributed as PDF files and require substantial processing to be used to train machine learning models.

The *LQAuAD-PL* dataset was created by converting PDFs into text, pairing questions with relevant legal provisions, and defining train, test, and validation subsets. Since the original questions are formed so the possible answers are probable continuations of them, they were rewritten to form normal linguistic questions. The dataset includes 3653 provisions, with 2965 questions in the training set, 318 in validation, and 370 in testing. The subset division was designed to assess the system's generalization, ensuring that questions, provisions, and answers from different legal domains were placed in separate subsets.

Simple Legal Questions This dataset was created in the style of SQuAD [22], but again for the legal domain. Students of Computer Science were given a corpus of Polish bills in textual format, divided into passages containing legal provisions (similarly to the other datasets) and were asked to create a question based on the content of the passage, so that the question could be answered based on it. Due to design of this process, there is a high lexical overlap between the created question and the answer. There are 1436 questions in the dataset and there are almost 26 thousand passages in the corpus. Since the number of questions is small, there is only the test subset available (i.e. the full dataset is the test dataset). This dataset is publicly available and was the legal subsets of the PolEval 2022/23 information retrieval competition [13].

Summary Table 1 contains the statistics regarding the different datasets we employed in our research. Comparing these datasets reveals the following observations. First of all *Lemkin questions* and *LQuAD-PL* have very similar number of development and test queries (around 300); *Rulings questions* and *Simple Legal Questions* have similar number of queries in the test subset (2000 – 1400); *Rulings questions* and *LQuAD-PL* have similar number of passages (2500 – 3600). We can conclude that these datasets are pretty diversified regarding the distributions of the passages and queries, so they are well suited for verifying different aspects of the information retrieval in the context of Polish law.

[1] With the exception of 2007, when they received 250 questions.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

Dataset	Provisions	Train	Dev	Test
Lemkin	8383	2471	333	327
Rulings	2545	5733	1976	2004
LQuAD-PL	3654	2965	318	370
Simple	$\sim$ 26000	–	–	1436

Table 1. The distributions of provisions (passages) and queries in the training, development and test subsets of the legal datasets.

4.2 Information Retrieval Models

Vanilla Models To answer the first and the second research questions, we have tested numerous retrieval models. We wanted to know what performance can be achieved when you apply these models into scenario with laymen questions. So the performance we obtain with these models requires very little research effort (no fine-tuning) and due to the strong capabilities of the models is very popular in commercial settings. These models were taken from the Huggingface MTEB leaderboard[2] and from the Polish Information Retrieval Leaderboard[3]. We have included not only the top-performing models, since there is a huge amount of different models and it is pretty easy to test them. We have restricted the comparison to models with up to 1.5B parameters and only tested the dense embedding feature of the models, to make the comparison and the implementation simpler.

The list below includes all the models, we have tested. They are named after the names available on HuggingFace with a citation to a relevant paper (if available). For models without papers, we just provide the link to its HuggingFace model card. In each case we put only one citation for a group of models, sharing the same paper – it applies to all the models, up to the one with the citation:

1. stella-pl
2. stella-pl-retrieval
3. mmlw-e5-large
4. mmlw-e5-small
5. mmlw-retrieval-roberta-base
6. mmlw-retrieval-roberta-large
7. mmlw-retrieval-e5-base
8. mmlw-retrieval-e5-large
9. mmlw-roberta-base
10. mmlw-roberta-large
11. polish-splade [7]
12. snowflake-arctic-embed-l-v2.0
13. snowflake-arctic-embed-l
14. snowflake-arctic-embed-m-v1.5
15. snowflake-arctic-embed-m-v2.0
16. snowflake-arctic-embed-s
17. snowflake-arctic-embed-xs [18]
18. silver-retriever-base-v1 [24]
19. e5-base
20. e5-large
21. e5-small-v2 [30]
22. multilingual-e5-large
23. multilingual-e5-large-instruct [31]
24. stella-en-1.5B
25. stella-en-400M-v5 [34]
26. KartonBERT-USE-base-v1[4]

[2] `https://huggingface.co/spaces/mteb/leaderboard`
[3] `https://huggingface.co/spaces/sdadas/pirb`
[4] `https://huggingface.co/OrlikB/KartonBERT-USE-base-v1`
[5] `https://huggingface.co/OrdalieTech/Solon-embeddings-large-0.1`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

27. bilingual-embedding-large [4,6,15,29]
28. Solon-embeddings-large-0.1[5]
29. arabic-english-sts-matryoshka-v2.0 [15]
30. gte-base-en-v1.5
31. gte-modernbert-base
32. gte-large-en-v1.5 [35,14]
33. nomic-embed-text-v1.5 [16]
34. MedEmbed-small-v0.1 [1]
35. all-MiniLM-L12-v2[6]
36. all-MiniLM-L6-v2[7]

Model Fine-tuning The third research question concerns the fine-tuning of the models. In this scenario, we are mostly interested in approaches that do not require a lot of manual annotation. The dataset created for the Lemkin project has a high value, be it was a lengthy and costly process. So we seek methods for improving the information retrieval performance with respect to the laymen questions, that will minimize the manual annotation effort.

So for the fine-tuning we will primarily compare an approach when we fine-tune the best performing vanilla models on the training part of the Lemkin project, with an approach when the other datasets are utilized. The first approach will serve as a baseline in a broad sense, meaning that we don't expect the trainings on the other datasets to beat that baseline.

We specifically compare the dataset with the exam questions, which works as a proxy to methods based on automatic question generation, with an approach based on the factual descriptions of facts, taken from the judgments. The second approach, although requires more work than simple question generation with the help of LLM, assuming access to a large number of judgments, is easily adaptable to the other languages and legislations. This is the primary reason we want to explore this approach.

So we have the following setups:

1. *Lemkin questions* – baseline, manual annotation of real questions,
2. *LQuAD-PL* – proxy for a method with automatically generated questions, high lexical overlap between a question and a passage,
3. *Rulings questions* – approach when factual descriptions from judgments are used to create query – passage pairs,
4. *LQuAD-PL + Rulings questions* – combination of the above approaches.

For the fine-tuning we use the following setup – learning rate: 1e-5, training epochs: 50, warmup ratio: 0.1, batch sampler: no duplicates, evaluation metric: NDCG@10 on validation subset of Lemkin dataset (used to select the best model for a given training setup), loss: multiple negatives ranking loss, batch size: 4.

[6] `https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2`
[7] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

NDCG	**@5**	**@10**	**@50**	**@100**
Lemkin questions				
stella-pl	39.0%	**51.1%**	**66.1%**	69.6%
snowflake-arctic-embed-l-v2.0	**40.7%**	49.6%	65.5%	**71.3%**
silver-retriever-base-v1	37.6%	48.8%	65.1%	69.9%
multilingual-e5-large	36.7%	47.9%	64.4%	69.9%
stella-pl-retrieval	34.4%	47.8%	63.0%	65.9%
Rulings Questions				
snowflake-arctic-embed-l-v2.0	**23.9%**	**26.7%**	**31.4%**	**32.9%**
snowflake-arctic-embed-m-v2.0	21.1%	23.8%	29.4%	30.6%
Solon-embeddings-large-0.1	19.7%	22.0%	26.9%	29.0%
multilingual-e5-large-instruct	19.4%	22.0%	28.1%	29.4%
stella-pl-retrieval	19.2%	21.7%	27.8%	29.4%
LQuAD-PL				
snowflake-arctic-embed-m-v2.0	**95.0%**	**96.2%**	97.0%	97.3%
snowflake-arctic-embed-l-v2.0	94.8%	**96.2%**	97.0%	97.6%
silver-retriever-base-v1	93.4%	**96.2%**	97.3%	97.6%
stella-pl-retrieval	94.8%	95.9%	**97.6%**	**98.1%**
multilingual-e5-large	93.6%	95.9%	97.0%	97.8%
Simple Legal Questions				
snowflake-arctic-embed-m-v2.0	**92.5%**	**96.5%**	97.9%	98.3%
snowflake-arctic-embed-l-v2.0	**92.5%**	96.2%	**98.6%**	**98.7%**
multilingual-e5-large	92.2%	96.2%	98.3%	98.5%
Solon-embeddings-large-0.1	92.2%	94.6%	98.3%	98.5%
mmlw-retrieval-roberta-large	90.9%	94.6%	97.6%	98.3%

Table 2. Top-5 NDCG@k results for *Lemkin questions, Rulings questions, LQuAD-PL* and *Simple Legal Questions* datasets sorted by NDCG@10 (and NDCG@5 in case of a tie).

5 Results

5.1 RQ1: Dataset Comparison

The results of the experiment with respect to the first question are given in Table 2. The results are sorted by NDCG@10 (and NDCG@5 in case of a tie) and we give only 5 best results, to save the space.

There are huge differences among the evaluated datasets. For *Simple Legal Questions* and *LQuAD-PL* the models obtain almost perfect NDCG even for k=5. It means that these datasets are very easy for the SOTA models and don't pose any real challenge currently. It might stem from the fact that retrieval models are pre-trained largely on lexical overlapping datasets and such tasks are trivial for them to solve.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

Yet when we look at the two other datasets we see they are much more challenging. For *Lemkin questions* the best performing models achieve only 50% NDCG@10 – a huge discrepancy with respect to the other datasets, even though all of them include the same type of questions and documents – i.e. questions about some legal matters and legal provisions as the passages. This result shows that it is **much harder to provide answers when a layman asks a legal question**. The result also indicates that when the questions are directly based on the documents, the results obtained are too optimistic. This should be a warning sign for all teams assuming that a generated QA dataset indicates the performance to be expected when it is deployed, especially when the task concerns a highly specialized language on the one hand and casual users on the other.

For *Rulings questions* we observe an even worse result – the best models achieves around 27% for NDCG@10, showing that the task is harder for this setup. This low result might stem from the fact that **this kind of data is rarely present in the training sets of the available models**. We haven't found any mentions of this type of data being included either in pre-training or fine-tuning datasets. Yet, it seems that the legal applications could benefit from such data, since the correlation between the descriptions of the facts and the legal provision applying to these facts seems to be very natural. In the following experiments we will check how much we can improve that result via fine-tuning.

6 RQ2: Model Comparison

Regarding the models that perform the best on the legal datasets, we have the following observations. It seems that the best model in this scenario is Snowflake Arctic Embedd v. 2 in the large version. It appears in the top-5 for all the datasets, it is the best for *Rulings questions* and takes the second position for the remaining 3 datasets (when we consider NDCG@10).

The second-strongest model is the medium version of Snowflake Arctic Embedd v. 2. It is the best model for 2 of the datasets (*LQuAD-PL* and *Simple Legal Questions*), the second for *Rulings questions* but does not appear in the top-5 for *Lemkin questions* (it was the 7th model in that setup). Although this model takes the first place for two of the dataset, this result is not that strong since these are the easy datasets with high lexical overlap.

The third strong model is Multilingual e5 in the large version, appearing among the top-5 models in 3 out of 4 datasets (it occupied the 15th position for the *Rulings questions* dataset). Polish version of the Stella model (retrieval variant) also appears in the top-5 for 3 datasets (appearing at the 6th position for the 4th dataset), so it can be counted equally strong. There is also Silver Retriever, a model targeting Polish, which is among the top-5 models in 2 cases, similarily to the Solon model.

The interesting outcome from this comparison is that the models trained specifically for Polish appeare worse than the best performing multilingual models (specifically the Snowflake Arctic Embedd v. 2). It appears that this family

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

of models seems to be the best choice for scenarios when we want to perform retrieval without doing fine-tuning and we don't have an extensive testing dataset to validate across different scenarios. The only Polish model that won the competition on one of the dataset is Stella PL (not the retrieval variant), but it does not appear in the top-5 results for the remaining datasets.

Looking from a different angle, it appears that the *Lemkin questions* dataset is in fact different from the other datasets, since there are three Polish models in the top-5 results. The reason behind that result might be the fact that this dataset includes questions imposed by laymen, so we can expect inaccuracies, colloquial phrases, etc. On the other hand, Polish models fall short on the other datasets – it seems the training setup does not include enough simple examples, when there is a large lexical overlap between the query and the passage (it's not the case for the Silver Retriever since it occupies the first position for the LQuAD-PL dataset, *ex-aequo* with the Snowflake models).

7 RQ3: Fine-tuning

For the fine-tuning we have selected: *Snowflake Artcitc Embedd* in the large and medium variants, *Solon* and *Silver Retriever*. We also tried to fine-tune the Polish Stella model, but without success. The results of fine-tuning of the models are given in Table 3. We report NDCG scores on the testing subset of the *Lemkin questions* sorted by NDCG@10.

Across all models, fine-tuning on the Lemkin dataset itself yielded the highest NDCG scores. The best-performing model, *Snowflake Arctic embed Large v2.0* fine-tuned on Lemkin, achieved 62.0% for NDCG@10, which is not surprising, since that model was performing the best without fine-tuning. The other models' performance when fine-tuning on this dataset is not much worse: *Solon* achieves 59.1%, *medium size of Snowflake Arctic Embedd* achieves 58.0% and *Silver Retriever* achieves 56.7% NDCG@10, so all these models work reasonably well when they are fine-tuned directly on manually annotated laymen questions. Thanks to fine-tuning we were able to improve the scores by more than 20 percentage points (pp.) for NDCG@10, which is definitely a huge gain. Still, when we look at the outcomes for the *LQuAD-PL* and *Simple Legal Questions* datasets with vanilla models, we observe that even the fine-tuned models perform much worse on this dataset. This result indicates that in reality the performance that can be obtained with laymen questions (the real-world scenario) will be much worse than the numbers observed for datasets with high lexical overlap (a very popular in-vitro scenario).

When we look at the second result for each fine-tuning setup, we observe that for 3 out of 4 setups, the best results can be obtained when training jointly on *Rulings questions* and *LQuAD-PL* questions. The best result with that setup was achieved by Snowflake Arctic Embedd in the large variant (55.5% NDCG@10), the remaining models obtained around 52% NDCG@10 score. This result is particularly encouraging since it shows that the fine-tuning on these datasets, rather than the manually annotated dataset, will also give strong results (for the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

NDCG	@5		@10		@50		@100	
Snowflake Arctic Embedd large v.2								
Lemkin	**59.2%**	(**+18.5**)	**62.0%**	(**+12.4**)	**66.4%**	(**+0.9**)	67.0%	(-4.3)
Rulings + LQuAD-PL	50.1%	(+9.4)	55.5%	(+5.9)	60.9%	(-4.6)	62.2%	(-9.1)
LQuAD-PL	45.9%	(+5.2)	51.3%	(+1.7)	56.3%	(-9.2)	57.3%	(-14.0)
Rulings	34.7%	(-6.0)	39.4%	(-10.2)	45.7%	(-19.8)	47.1%	(-24.2)
Snowflake Arctic Embedd medium v.2								
Lemkin	55.1%	(**+17.2**)	58.0%	(**+11.0**)	63.6%	(**+3.8**)	64.4%	(-5.0)
Rulings + LQuAD-PL	46.8%	(+8.9)	52.4%	(+5.4)	56.3%	(-3.5)	58.1%	(-11.3)
Rulings	46.9%	(+9.0)	51.1%	(+4.1)	55.9%	(-3.9)	57.6%	(-11.8)
LQuAD-PL	42.8%	(+4.9)	48.2%	(-1.2)	52.6%	(-7.2)	54.2%	(-15.2)
Solon								
Lemkin	54.5%	(**+21.7**)	59.1%	(**+16.5**)	63.3%	(-0.3)	63.7%	(-5.1)
Rulings + LQuAD-PL	49.1%	(+16.3)	51.9%	(+9.3)	57.9%	(-5.7)	59.3%	(-9.5)
Rulings	47.7%	(+14.9)	51.3%	(+8.7)	57.7%	(-5.9)	58.9%	(-9.9)
LQuAD-PL	44.8%	(+12.0)	49.1%	(+6.5)	54.9%	(-8.7)	56.1%	(-12.7)
Silver Retriever base v. 1								
Lemkin	52.6%	(**+15.0**)	56.7%	(**+7.9**)	60.9%	(-4.2)	61.9%	(-8.0)
Rulings	48.7%	(+11.1)	52.4%	(+3.6)	58.2%	(-6.9)	59.2%	(-10.7)
Rulings + LQuAD-PL	47.7%	(+10.1)	52.1%	(+3.3)	57.6%	(-7.5)	58.9%	(-11.0)
LQuAD-PL	45.5%	(+7.9)	49.6%	(+0.8)	55.3%	(-9.8)	56.6%	(-13.3)

Table 3. Results of fine-tuning 4 retrieval models on various combinations of datasets. The measured performance is the NDCG score on the testing subset of *Lemkin questions* sorted by NDCG@10.

best setup 7 pp. worse than the best model fine-tuned on laymen questions – 7 pp. seems a huge gap, but we should also acknowledge that without fine-tuning we get only 50% NDCG@10 for the best model, so there is a huge performance gain). Only for Silver Retriever the second-best results was obtained when training only on the *Rulings questions* dataset, but the difference with the combined setup is very small (only 0.3 pp.).

For all setups but Snowflake Arctic Embedd large, the worst results were achieved with *LQuAD-PL* dataset. It's interesting to observe that for 3 of the setups the difference when training on the *Lemkin questions* and *LQuAD-PL* is around 10 pp. (Snowflake large – 10.7 pp., Snowflake medium – 9.8 pp., Solon – 10 pp. and Silver Retriever – 7.1 pp.). Since we know that *LQuAD-PL* works as a proxy for automatically generated questions, this is a very important result, showing that the performance gap is pretty huge, when training only on such a dataset. We gained approx. 10 pp. from the best vanilla model, but were always at least 10 pp. away from the best setup. This result clearly indicates that trying to find a dataset tailored to our scenario is a much better option than simple question generation.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

There are three additional observations to be made. First of all, the resutls when fine tuning Snowflake Arctic Embedd large on the *Rulings questions* solely is much worse than in the other setups. We have not found any reason to explain that phenomenon. The second observation relates to the fact that when fine-tuning the model on the factual descriptions, we obtain much better results than the vanilla models obtain on that dataset, without fine-tuning (less than 30% for the best model). We concluded that this kind of dataset is in fact very valuable for fine-tuning retrieval models and should be included, especially if the model is later applied in the legal domain. There is also an important observation that when we fine tune the models, the performance for NDCG@50 and NDCG@100 gets worse (this does not hold for NDCG@50 when training on the Lemkin dataset). So we have to be very careful during the fine-tuning procedure and we have to take into account to final deployment of the model – if a huge context window works for our LLM, we should at least optimize for NDCG@50 or @100, otherwise the fine-tuned model will work worse than the vanilla models.

8 Conclusions

In this study, we investigated the challenges of interpreting legal questions, particularly those posed by laymen, and evaluated various retrieval models on multiple legal datasets. Our results highlight significant differences between datasets, with *Simple Legal Questions* and *LQuAD-PL* proving too easy for state-of-the-art models, while *Lemkin Questions* (50% NDCG@10) and *Rulings Questions* (27% NDCG@10) pose greater difficulties. Fine-tuning on *Lemkin Questions* yielded the highest performance (62% NDCG@10), improving results by over 20 percentage points, though fine-tuned models still struggle with laymen queries. A key finding is that training on datasets tailored to legal scenarios is far more effective than relying on automatically generated questions.

In our future work, we plan to explore additional fine-tuning strategies and incorporating more diverse legal texts to improve retrieval accuracy. Given the strong correlation between factual descriptions and legal provisions, further refinement of training data could enhance model performance in real-world applications. We also aim to analyze the impact of fine-tuning on generalization across different legal domains, ensuring that retrieval models remain robust and reliable when applied to various legal contexts. Our study underscores the importance of dataset selection in legal AI development and lays the groundwork for improving automated legal information systems.

9 Acknowledgments

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017168.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

We would like to thank AGH University and the Inicjatywa Doskonałości – Uczelnia Badawcza programme for funding the grant "LQuAD – Zastosowanie transferu uczenia w dziedzinie odpowiadania na pytania w domenie prawa."

References

1. Balachandran, A. Medembed: Medical-focused embedding models, 2024.
2. Blair-Stanek, A., and Durme, B. V. Llms provide unstable answers to legal questions. *Arxiv Preprint* (2025).
3. Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., and Aletras, N. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Dublin, Ireland, May 2022), S. Muresan, P. Nakov, and A. Villavicencio, Eds., Association for Computational Linguistics, pp. 4310–4330.
4. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216* (2024).
5. Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).
7. Dadas, S., Perełkiewicz, M., and Poświata, R. PIRB: A comprehensive benchmark of polish dense and hybrid text retrieval methods.
8. Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. Large legal fictions: Profiling legal hallucinations in large language models. *Arxiv Preprint* (2024).
9. de Oliveira Lima, J. A. Unlocking legal knowledge with multi-layered embedding-based retrieval. *CoRR abs/2411.07739* (2024).
10. Dnyanesh Panchal, Aaryan Gole, V. N. R. J. Lawpal : A retrieval augmented generation based system for enhanced legal accessibility in india. *Arxiv Preprint* (2025).
11. Hendrycks, D., Burns, C., Chen, A., and Ball, S. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268* (2021).
12. System analizy orzeczeń sądowych SAOS, https://www.saos.org.pl/.
13. Kobyliński, Ł., Ogrodniczuk, M., Rybak, P., Przybyła, P., Pezik, P., Mikołajczyk, A., Janowski, W., Marcińczuk, M., and Smywiński-Pohl, A. Poleval 2022/23 challenge tasks and results. In *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)* (2023), IEEE, pp. 1243–1250.
14. Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. Towards general text embeddings with multi-stage contrastive learning, 2023.
15. Nils Reimers, I. G. Sentence-bert: Sentence embeddings using siamese bert-networks. *https://arxiv.org/abs/1908.10084* (2019).
16. Nussbaum, Z., Morris, J. X., Duderstadt, B., and Mulyar, A. Nomic embed: Training a reproducible long context text embedder, 2024.
17. Pawitsapak Akarajaradwong, Pirat Pothavorn, C. C. P. T. T. N. S. N. Nitibench: A comprehensive studies of llm frameworks capabilities for thai legal question answering. *Arxiv Preprint* (2025).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

18. Puxuan Yu, Luke Merrick, G. N., and Campos, D. Arctic-embed 2.0: Multilingual retrieval without compromise. *Arxiv Preview* (2024).
19. Puxuan Yu, Luke Merrick, G. N., and Campos, D. Arctic-embed 2.0: Multilingual retrieval without compromise. *Arxiv Preview* (2024).
20. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2020).
21. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res. 21*, 1 (Jan. 2020).
22. Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint arXiv:1606.05250* (2016).
23. Robertson, S. E., and Walker, S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Berlin, Heidelberg, 1994), SIGIR '94, Springer-Verlag, p. 232–241.
24. Rybak, P., and Ogrodniczuk, M. Silver retriever: Advancing neural passage retrieval for Polish question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (Torino, Italia, May 2024), N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., ELRA and ICCL, pp. 14826–14831.
25. Sansone, C., and Sperlí, G. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems 106* (2022), 101967.
26. Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., and Xu, H. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv preprint arXiv:2306.13906* (2023).
27. Smywiński-Pohl, A., and Libal, T. Enhancing legal argument retrieval with optimized language model techniques. In *JSAI International Symposium on Artificial Intelligence* (2024), Springer, pp. 93–108.
28. Solaiman, I., and Dennison, C. Process for adapting language models to society (palms) with values-targeted datasets, 06 2021.
29. Thakur, N., Reimers, N., Daxenberger, J., and Gurevych, I. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv e-prints* (2020), arXiv–2010.
30. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* (2022).
31. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672* (2024).
32. Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).
33. Yao, R., Wu, Y., Wang, C., Xiong, J., Wang, F., and Liu, X. Elevating legal llm responses: Harnessing trainable logical structures and semantic knowledge with legal reasoning. *Arxiv Preprint* (2025).
34. Zhang, D., Li, J., Zeng, Z., and Wang, F. Jasper and stella: distillation of sota embedding models, 2025.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20

35. Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., and Zhang, M. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval, 2024.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_20