LEMUR: A Corpus for Robust Fine-Tuning of Multilingual Law Embedding Models for Retrieval

Narges Baba Ahmadi, Jan Strich, Martin Semmann, Chris Biemann
Hub of Computing and Data Science (HCDS)
University of Hamburg, Germany

Correspondence: `{first_name}.{last_name}@uni-hamburg.de`

Abstract

Large language models (LLMs) are increasingly used to access legal information. Yet, their deployment in multilingual legal settings is constrained by unreliable retrieval and the lack of domain-adapted, open-embedding models. In particular, existing multilingual legal corpora are not designed for semantic retrieval, and PDF-based legislative sources introduce substantial noise due to imperfect text extraction. To address these challenges, we introduce LEMUR, a large-scale multilingual corpus of EU environmental legislation constructed from 24,953 official EUR-Lex PDF documents covering 25 languages. We quantify the fidelity of PDF-to-text conversion by measuring lexical consistency against authoritative HTML versions using the Lexical Content Score (LCS). Building on LEMUR, we fine-tune three state-of-the-art multilingual embedding models using contrastive objectives in both monolingual and bilingual settings, reflecting realistic legal-retrieval scenarios. Experiments across low- and high-resource languages demonstrate that legal-domain fine-tuning consistently improves Top-k retrieval accuracy relative to strong baselines, with particularly pronounced gains for low-resource languages. Cross-lingual evaluations show that these improvements transfer to unseen languages, indicating that fine-tuning primarily enhances language-independent, content-level legal representations rather than language-specific cues. We publish code[1] and data[2].

1 Introduction

LLMs are transforming legal work and research by making access to legal knowledge, automated document review, and case law summarization significantly faster and easier (Zheng et al., 2021). However, the deployment of these models in legal practice is often hindered by "hallucinations" and a lack of grounding in authoritative legal sources (Reuter et al., 2025; Magesh et al., 2025). To mitigate these risks, Retrieval-Augmented Generation (RAG) has become the de facto standard architecture, ensuring that model outputs are anchored in verifiable primary documents (Lewis et al., 2020).

While RAG relies on the combination of an LLM for generation and embedding models for retrieval, its success is fundamentally dependent on the retrieval setup and the embedding model used for the vector database (Gao et al., 2023). While these models are typically general-purpose, fine-tuning them on domain-specific data consistently yields superior performance compared to existing specialized models (Tang and Yang, 2025), particularly in law, where text often contains archaic terminology, complex syntactic structures, or polysemy (Ariai et al., 2025). Nevertheless, these models are untrained in the legal domain and primarily monolingual (Chalkidis et al., 2020, 2022) or proprietary (Voyage AI, 2024).

While multilingual datasets in law already exist, the data are typically formatted for pretraining (Henderson et al., 2022; Niklaus et al., 2024) or for classification of legal documents (Chalkidis et al., 2021), leaving a void for high-quality benchmarks dedicated to cross-lingual semantic retrieval. Furthermore, legal corpora are often stored in PDF format, which can introduce inaccuracies when converted to text for effective search due to multi-column layouts and nested tables. This 'extraction gap' affects data integrity in RAG systems, as downstream embedding models are forced to process corrupted or misaligned tokens. To address these gaps, our contributions are:

- **Multilingual Dataset (LEMUR):** We introduce a **L**aw **E**uropean **MU**ltilingual **R**etrieval corpus (**LEMUR**), which consists of **25k** EU legal PDFs in 25 languages, designed for training embedding models on legal text.

[1] GitHub Repository
[2] Hugging Face Dataset

- **The Lexical Content Score (LCS):** We systematically analyze PDF-to-text conversion quality by measuring content consistency across **twenty-five** languages.

- **Legal Embedding Fine-Tuning:** We fine-tune **three SOTA** embedding models on **five languages** for a legal document retrieval task and evaluate them in monolingual, bilingual, and cross-lingual settings.

2 Related Work

Multilingual Legal Corpora. Research on multilingual legal corpora has produced both supervised benchmarks (Chalkidis et al., 2020, 2022; Zheng et al., 2021; Ma et al., 2021) and large-scale pretraining resources (Niklaus et al., 2024; El-Haj and Ezzini, 2024). Chalkidis et al. (2021) introduces MULTIEURLEX, a multilingual multi-label dataset of EU legislation in 23 languages for legal document classification, while Chalkidis et al. (2022) proposes LEXGLUE, a suite of English legal NLU benchmarks that has become a standard evaluation protocol for legal language models. Beyond EU legislation, Zheng et al. (2021) presents CaseHOLD, a multiple-choice benchmark comprising more than 53,000 U.S. case-law holdings, and Ma et al. (2021) introduces LeCaRD, a large-scale case-retrieval dataset for the Chinese criminal law system with expert-designed relevance criteria.

Our work contributes to this line of research by constructing a new multilingual EU law dataset directly from official legislative PDFs and targeting downstream embedding-model fine-tuning across multiple European languages, thereby bridging large-scale pretraining corpora and task-specific benchmarks in an EU legislative setting.

Embedding Models and Legal Retrieval. Recent work shows that structure-aware models such as SAILER (Li et al., 2023) and DELTA (Li et al., 2025) capture section-level or structural dependencies to improve legal case retrieval, while SM-BERT-CR (Vuong et al., 2022) and REAKASE-8B (Tang et al., 2025) incorporate supporting-relation modeling and reasoning-driven representations. For multilingual and cross-lingual settings, LEXCLIPR (Upadhya and T.y.s.s, 2025) enables paragraph-level retrieval across ECtHR judgments, showing that off-the-shelf multilingual encoders struggle without domain-adaptive training.

Domain-specific pretraining consistently improves legal NLP tasks. Limsopatham (2021) shows that in-domain pretraining and long-document handling benefit legal classification, while Darji et al. (2023) demonstrates gains from adapting BERT to legal NER over BiLSTM–CRF baselines. More broadly, Tang and Yang (2025) and BloombergGPT (Wu et al., 2023) provide evidence that domain-adapted embeddings remain essential despite strong general-purpose LLMs.

Although these studies focused mainly on monolingual legal data or non-legal domains, they have not systematically studied cross-lingual retrieval for EU legislation. We address this gap by introducing a multilingual EU law corpus and evaluating fine-tuned embedding models for monolingual and cross-language retrieval on EUR-Lex texts.

3 LEMUR

We construct LEMUR from official documents published on EUR-Lex[3]. Section 3.1 details how the source documents are identified, selected, and collected from the EUR-Lex repository. Section 3.2 then explains the process of converting the original PDF files into a structured and machine-readable text format, and Section 3.3 describes the subsequent preprocessing steps, including the construction of high-quality query–document pairs used in our experiments. An overview of the data preparation process from EUR-Lex PDFs to structured JSONL is shown in Figure 3.

3.1 Document Collection

To build a focused corpus, we gathered all legal acts listed under *Category 15* (Environment, consumers and health protection), *Subcategory 10* (Environment), across all available publication years. This yielded **1,174** distinct legal acts from **1961–2025**. Because each act is available in 25 official EU languages, the collection comprises a total of **24,953** PDF documents and results in **461k** pages. Figure 2 summarizes the number of records per language. Coverage is highest for languages with long-standing EU membership (e.g., German, Dutch, English, Italian), and lower for countries with more recent membership (e.g., Croatia or Bulgaria).

3.2 PDF–to–Text Conversion

In the original source of EUR-Lex the dataset is available in PDF and HTML format. Previous

[3] https://eur-lex.europa.eu/homepage.html

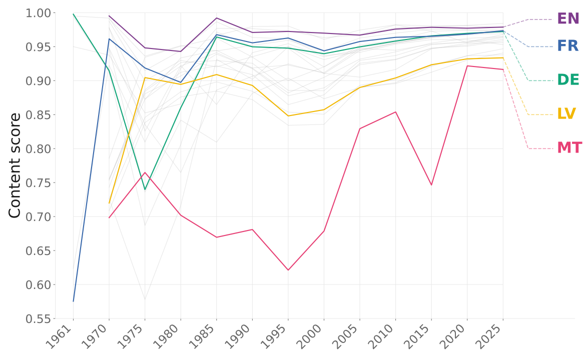


Figure 1: Average Content Score similarity per year (5-year bins) for the five languages used in our experiments

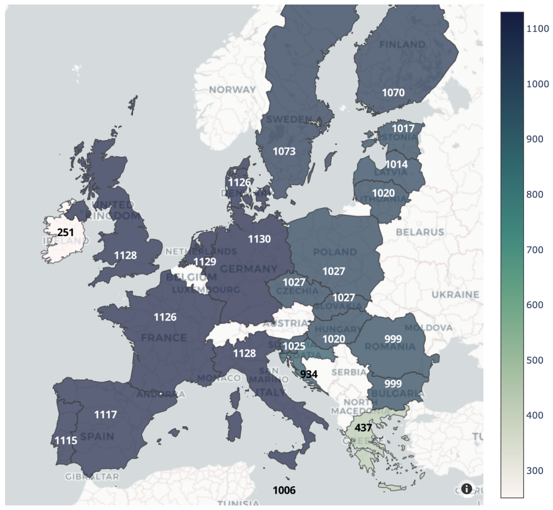


Figure 2: Number of documents per country in LEMUR.

datasets (Chalkidis et al., 2021) used the HTML version, but we found that tables were not converted correctly. Therefore, we tested multiple PDF–to–text services (Docling (Livathinos et al., 2025), Unstructured (Unstructured Team, 2023), PyMuPDF (Developers, 2021)) but found that the best results were obtained by converting all PDFs into structured JSONL files using **olmOCR** (Poznanski et al., 2025). On average, documents contain approximately **19** pages, with approximately **403** tokens per page, yielding roughly **7,781** tokens per document. These values indicate that LEMUR consists primarily of long-form legislative text, making it well-suited for evaluating embedding models on long-document and multilingual retrieval tasks.

To verify the quality of the PDF–to–text conversion, we compare each converted document against the corresponding HTML version available on EUR-Lex. While HTML files provide a clean textual baseline, they often linearise tables in ways that differ from the official PDF layout. In contrast, the JSONL files extracted with OLMOCR preserve table structure more consistently and also in markdown format, which is essential for downstream retrieval tasks that rely on faithful representation of legislative formatting. For this reason, the JSONL representation is used as the primary data source in LEMUR, while the HTML version serves solely as a reference for evaluation. We present LCS for all approaches in Appendix B.

Lexical Content Similarity (LCS). To evaluate the PDF-to-text conversion, we compute a content similarity score between each converted document and its corresponding HTML version. Before that, the HTML text is normalized to remove superficial differences that could affect lexical comparison. This includes removing all styling attributes (e.g., class, id, style) from HTML tags, stripping leading and trailing whitespace, converting to lowercase, normalizing numeric formatting (e.g., \$ 100 becomes \$100), and collapsing repeated punctuation (e.g., ... is replaced with .).

After normalization, we represent both as bag-of-words vectors (Qader et al., 2019), **v**H and **v**PDF, in a shared vocabulary of size n, where each entry corresponds to the frequency of a unique word. The content similarity score is then defined as the cosine similarity between these vectors, as

$$\mathrm{LCS}(h_{\mathrm{H}}, h_{\mathrm{PDF}}) = \frac{\sum_i v_{\mathrm{H},i} \cdot v_{\mathrm{PDF},i}}{\sqrt{\sum_i v_{\mathrm{H},i}^2}\,\sqrt{\sum_i v_{\mathrm{PDF},i}^2}} \quad (1)$$

where $v_{\mathrm{H},i}$ and $v_{\mathrm{PDF},i}$ denote the counts of the i-th word in the HTML and PDF texts, respectively. By applying these preprocessing steps and computing cosine similarity, the content score measures the actual lexical similarity between documents.

Conversion Results. Figure 1 illustrates the average Content Score similarity between the converted JSONL documents and their original HTML counterparts across all languages in LEMUR, stratified by year. For our primary analysis, we evaluate five languages with varying degrees of representation: high-resource languages (English (EN), German (DE), and French (FR)) and low-resource languages (Latvian (LV) and Maltese (MT)). This selection allows us to assess whether the model's performance generalizes from well-represented languages in the pretraining corpus to those that are comparatively underrepresented.

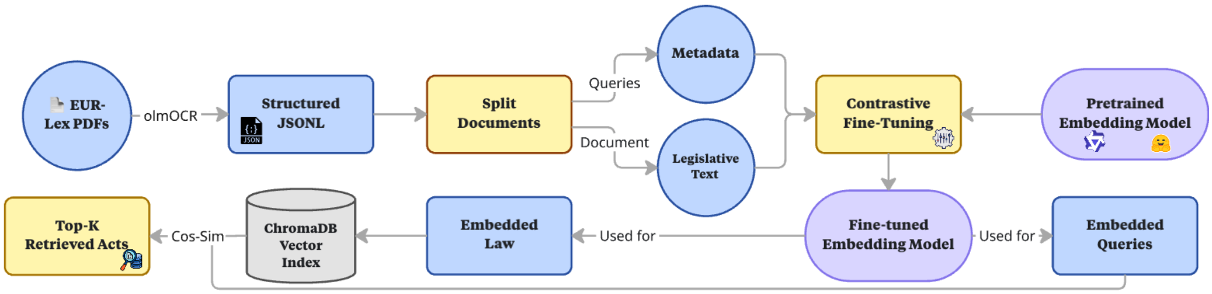


Figure 3: End-to-end pipeline for data preparation, contrastive fine-tuning, and retrieval. EUR-Lex PDFs are processed into structured JSONL, split into queries (metadata) and documents (legislative text), and used to fine-tune embedding models. The resulting embeddings are indexed for Top-k retrieval of legislative acts.

Our results indicate that for high-resource languages, the OLM-OCR model (Poznanski et al., 2025) achieves a similarity score exceeding 95%. However, we observe performance degradation for older documents, likely due to less standardized formatting compared with modern web documents. We hypothesize that the OLM-OCR training distribution is more closely aligned with contemporary layout standards. While performance is lower for low-resource languages, averaging approximately 90% for Latvian and 80% for Maltese, the similarity scores remain sufficiently high to justify using these converted documents for fine-tuning embedding models. We also present avg. LCS for all other languages in Appendix A as well as an overview of the distribution of publications per year in Appendix C.

3.3 Data Preprocessing

Figure 3 shows the pipeline for the data preprocessing to transform the documents to query–document pairs. After the transformation to structured JSONL, each legislative document in LEMUR begins with a short introductory block that we refer to as *metadata*. This block typically includes the act type (e.g., Commission Decision), the date, a brief description of the subject matter, references to the underlying legal basis, and standard publication notes, such as notification numbers, statements regarding the authentic language version, and indications of whether the text is relevant to the EEA.

We split each document into two parts: the introductory metadata block at the beginning of the document, which serves as the query, and the remaining substantive text of the legal act, which constitutes the document to be retrieved. This setup reflects realistic legal search behavior: a user begins with a short, structured description that provides only partial information about the act, whereas the retriever must identify the full legislative text. We use metadata as queries and the remaining text as a corpus, producing a large set of retrieval-ready pairs for both monolingual and cross-lingual evaluation. Examples are provided in Appendix D.

4 Method

This section gives an overview of the training procedure of the embedding models on LEMUR and the construction of our retrieval pipeline. Subsection 4.1 outlines the retrieval-oriented data representation, while Subsection 4.2 presents our monolingual fine-tuning procedure based on a contrastive learning approach (Hadsell et al., 2006; Henderson et al., 2017). Subsection 4.3 extends this approach to bilingual multi-positive training. Subsection 4.4 details the construction of the vector-database component used for retrieval.

4.1 Retrieval-Oriented Training Pairs

As described in Section 3, every document in LEMUR is split into a short metadata block and the remaining substantive legislative text. We directly adopt that structure for retrieval.

Accordingly, each legislative act yields a single query–document pair without requiring additional query construction or rewriting. This setup reflects realistic legal search behavior, in which users often begin with brief structured information. Each data entry contains the complete page content, with a clear separation between the metadata block and the remainder of the legislative text. The data is split into **60% training**, **20% validation**, and **20% test** sets, independently for each language or language pair, such that the same underlying legislative acts are assigned to the same split across languages, with each split containing the corresponding translations of those acts.

4.2 Monolingual Contrastive Fine-Tuning

We first adapt embedding models to the EUR-Lex retrieval setting in a **monolingual** fashion, fine-tuning one model per language. We experiment with the publicly available embedding models `Qwen3-0.6B` and `Qwen3-4B` (Yang et al., 2025), as well as `E5-Multilingual` (Wang et al., 2024), all obtained from the MTEB leaderboard[4]. These models were selected to cover a range of sizes and to have been pretrained on multilingual data and legal-domain tasks. They are also widely used in production and scored high on the MTEB leaderboard (3M downloads per month, Dec 25)[5]. For each model and language, a dedicated embedding model is fine-tuned using metadata as queries and the corresponding legislative text as the positive document. Fine-tuning uses a contrastive *Multiple Negatives Ranking* (MNR) objective with in-batch negatives (Henderson et al., 2017).

Objective Function. Given a batch of query–document pairs $\{(q_i, d_i)\}_{i=1}^{B}$, each (q_i, d_i) is treated as a positive pair, while all other documents in the batch act as negatives. Let $f(\cdot)$ denote the encoder producing L_2-normalized embeddings, and let $s_{ij} = f(q_i)^\top f(d_j)/T$ denote the temperature-scaled cosine similarity. We optimize the symmetric MNR loss:

$$\mathcal{L} = -\frac{1}{2B} \sum_{i=1}^{B} \left(\log \frac{e^{s_{ii}}}{\sum_j e^{s_{ij}}} + \log \frac{e^{s_{ii}}}{\sum_j e^{s_{ji}}} \right) \tag{2}$$

Training Setup. We train for up to 30 epochs, with early stopping based on the validation loss. Most models support a maximum sequence length of 2,048 tokens; the only exception is `E5-Multilingual`, which is restricted to 512 tokens. Training uses `bfloat16` precision, gradient checkpointing where supported, and a linear warm-up schedule. Training is performed on NVIDIA RTX A6000 and NVIDIA A100 (80GB) GPUs, with the larger `Qwen3-4B` model trained on A100 due to its higher memory requirements. In terms of training cost, fine-tuning `E5` typically completes within approximately 20–30 minutes per language, `Qwen3-0.6B` requires on the order of 2–4 hours, and `Qwen3-4B` requires roughly 6–8 hours per language, depending on the dataset size.

4.3 Bilingual Multi-Positive Contrastive Fine-Tuning

To exploit the availability of parallel legislative acts across languages, we extend the monolingual setup to a **bilingual multi-positive** scenario. In this setting, one metadata query is paired with *multiple* language versions of the same legislative act, and all corresponding documents are treated as positives during training. This enables the model to learn jointly from aligned legal content across languages.

Objective Function. We use a **grouped multi-positive** extension of the symmetric MNR loss, following Zhao et al. (2024). For each query embedding q_i, all document embeddings corresponding to aligned versions of the same legislative act are treated as positives, while all other documents in the batch serve as negatives. This objective encourages each query to be simultaneously close to multiple positive documents, promoting cross-lingual semantic alignment.

Similarity is computed using L_2-normalized embeddings and a temperature-scaled dot product. We optimize the following symmetric grouped multi-positive MNR objective:

$$\mathcal{L} = -\frac{1}{2B} \sum_{i=1}^{B} \Bigg[\log \frac{\sum_{j \in \mathcal{P}(i)} e^{s_{ij}}}{\sum_j e^{s_{ij}}} + \log \frac{e^{s_{ii}}}{\sum_j e^{s_{ji}}} \Bigg] \tag{3}$$

where $\mathcal{P}(i)$ denotes the set of positive documents for query q_i within the batch.

4.4 VectorDB Construction for Retrieval

To test the performance of the embedding models, we simulated a retrieval by constructing a vector store using ChromaDB (Chroma Team, 2025), a lightweight vector database optimized for similarity search. For each language, we created a collection for both the base and fine-tuned embedding models. Very long documents are truncated using a sequence of decreasing token caps to ensure compatibility with model limits. Across languages and models, approximately 8–15% of documents require truncation; for these documents, roughly 40–50% of their original tokens are removed. All stored vectors are L_2-normalized, and cosine similarity is used during retrieval.

At inference time, the metadata again serves as the query. It is embedded using the same model

[4] https://huggingface.co/spaces/mteb
[5] https://huggingface.co/intfloat/multilingual-e5-large

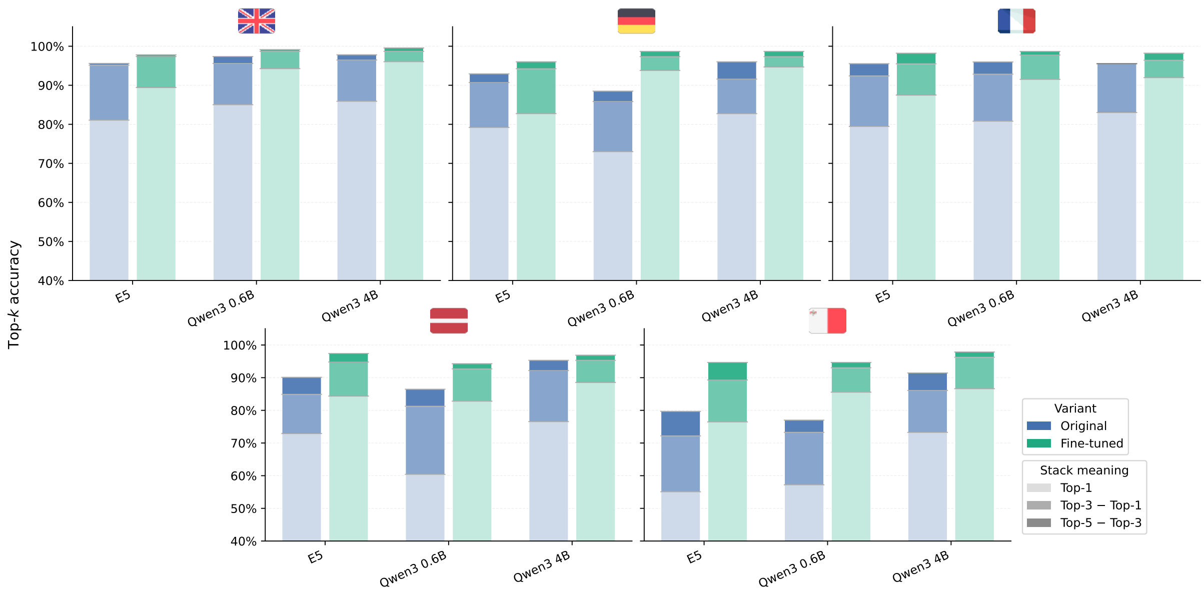


Figure 4: Monolingual fine-tuning of three embedding models (E5, Qwen-0.6B & Qwen-4B) on five languages (EN, DE, FR, LV, MT). Performance is measured using Acc@k for 1/3/5 on test queries evaluated against the test document collection, represented as stacked bars, and compared between the base model and the fine-tuned variant.

that indexed the documents, and nearest-neighbor search is performed in ChromaDB using cosine similarity to retrieve the most semantically similar legislative texts. This retrieval component constitutes the retrieval pipeline used in our experiments.

5 Evaluation and Results

This section outlines the three main experiments we conducted to demonstrate that multilingual embedding models can be trained on law data. All experiments follow the same pipeline shown in Figure 3, differing only in the fine-tuning configuration and language setup. Firstly, we used monolingual contrastive fine-tuning to train on five individual languages (EN, DE, FR, LV & MT) as described in Subsection 5.3. Secondly, we conducted a bilingual fine-tuning experiment to study the interaction between high-resource and low-resource languages. In this setting, a model was trained jointly on pairs of languages to analyze how the inclusion of a high-resource language influences representation learning for a low-resource language, and conversely, whether low-resource data affects performance in a high-resource setting, as described in Subsection 5.4. We conclude our analysis by evaluating our fine-tuned models cross-lingually across multiple languages to test whether performance is driven by content rather than language, and to investigate content generalization, as shown in Figure 5.

5.1 Retrieval Task and Evaluation Settings

We evaluate **metadata-to-document retrieval** performance as initialized in Subsection 4.4. For each legal act, the introductory metadata block serves as the query, while the remaining substantive text (with metadata removed) forms the retrieval target. A query is considered correct if its corresponding ground-truth document is retrieved within the top-k results. To assess retrieval performance under different corpus conditions, we consider two complementary evaluation settings. In **Full-dataset search**, each test query is evaluated against a collection containing all documents (training, validation, and test) in the relevant language(s). In contrast, **Test-only search** restricts retrieval to the subset of held-out test documents.

The size of the test set varies across languages. This variation arises because some languages were introduced into EU legislation at later stages, resulting in fewer available legal acts for earlier years, and because a small number of documents were excluded due to data corruption or incomplete text extraction. The exact number of test queries per language is reported in Appendix E.

5.2 Evaluation Metrics

Let $\mathcal{Q}$ be the set of test queries, and let $\mathrm{rank}(q)$ denote the rank position of the ground-truth document returned for query q (with $\mathrm{rank}(q) = \infty$ if not retrieved). We compute Top-k accuracy as:

$$\text{Acc@}k \;=\; \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathbb{I}\big[\text{rank}(q) \le k\big] \qquad (4)$$

where $\mathbb{I}[\cdot]$ is the indicator function. We report Acc@1, Acc@3, and Acc@5.

5.3 Monolingual Fine-Tuning

In the monolingual setting, each model is fine-tuned on a dedicated language and evaluated on retrieval in that same language. We chose three high-resource languages (EN, DE, FR) and two low-resource languages from the dataset corpus to test our hypotheses. Figure 4 summarizes Top-k retrieval accuracy across all five languages for test queries evaluated against the test document collection, highlighting the impact of fine-tuning on retrieval quality. Across all evaluated languages, fine-tuning consistently improves retrieval performance compared to the corresponding pre-trained models, with gains observable at Top-1, Top-3, and Top-5. While high-resource languages showed consistently better performance even on the baseline, gains were observed across all languages. On the other hand, for low-resource languages, baseline performance was much lower, but fine-tuning brought it to levels comparable to those of high-resource languages. While absolute accuracy varies by language and backbone, the effect direction is consistent: monolingual contrastive adaptation yields a more reliable ranking of the correct legislative act among the top retrieved results. This indicates that fine-tuning effectively aligns short metadata-style queries with their associated legal texts and that this benefit generalizes across multiple European languages. We also report the results for all other languages in Appendix H.

5.4 Bilingual Fine-Tuning

We evaluate bilingual fine-tuning by training on a high-resource language (English) jointly with a low-resource language (Latvian), treating aligned versions of the same legal act as positives. Table 1 shows the results for the three models, showing baseline results, trained on English-only, Latvian-only, and EN_LV together.

The results are mixed across models. For `E5`, fine-tuning across multiple languages has an additive effect, and retrieval performance improves when using both languages. This result is not consistent with the `Qwen` models. We find that, for both

Model	Train	EN Eval		LV Eval	
		Top-1	Top-5	Top-1	Top-5
E5	ORIG	81.06	95.59	72.91	90.10
	EN	89.43	**97.80**	82.29	94.27
	LV	87.22	96.03	**84.37**	97.39
	EN–LV	**90.30**	97.35	83.85	**97.91**
Qwen3-0.6B	ORIG	85.02	97.36	60.41	86.45
	EN	**94.27**	**99.12**	74.47	92.18
	LV	91.18	98.67	**82.82**	94.27
	EN–LV	88.54	97.35	77.08	**97.39**
Qwen3-4B	ORIG	85.90	97.80	76.56	95.31
	EN	**96.04**	**99.56**	87.50	**97.39**
	LV	95.59	95.55	**88.54**	96.87
	EN–LV	90.74	98.67	75.52	96.35

Table 1: Top-1 and Top-5 performance of three models trained on English (EN), Latvian (LV), and combined EN–LV data, evaluated on EN and LV datasets.

models, training on the dedicated language yields better results than training on both languages together. The performance using both languages for fine-tuning is, in most cases, better than without training at all, but shows no additive effect.

In addition to these results, we find that bilingual fine-tuning does not improve retrieval performance on English compared with English-only training. Across all models, adding Latvian data neither enhances nor substantially degrades English Top-1 or Top-5 accuracy. This asymmetry suggests that bilingual training primarily benefits lower-resource languages by leveraging additional high-resource supervision, while preserving strong performance on high-resource languages without introducing negative transfer.

5.5 Cross-Lingual Transfer Results

For further testing, regardless of whether the models learn language-independent general content, we evaluated the fine-tuned models on additional language evaluation datasets. Therefore, models are fine-tuned on a source language and assessed on a different target language without further training. During evaluation, both queries and documents are in the target language, but embeddings are produced using the source-language fine-tuned model. This setting isolates the extent to which legal-domain knowledge learned in one language transfers to unseen languages. We conducted this experiment for each model and present the results in Figure 5 for `Qwen-0.6B` and for the others in Appendix F.

The results for the `Qwen3 0.6B` model again show differences between high- and low-resource languages. Across the high-resource languages (EN, DE, FR), the fine-tuned models generalize to

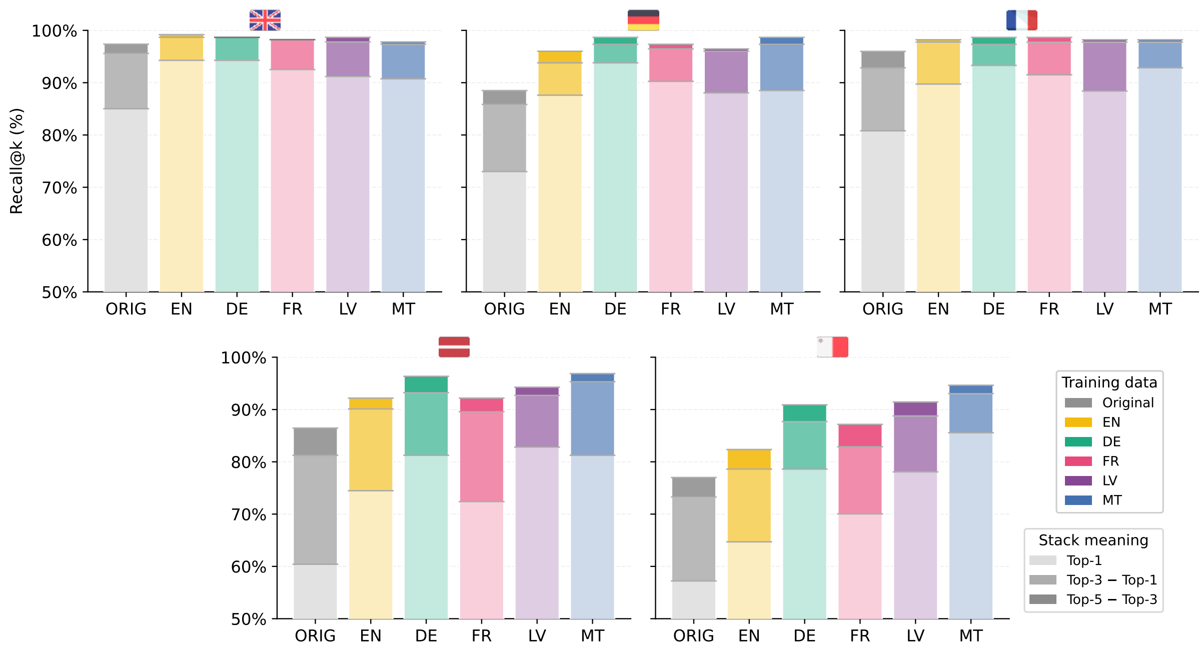


Figure 5: Cross-lingual fine-tuning for Qwen3 0.6B on five languages (EN, DE, FR, LV, MT). Performance is measured using Acc@k for 1/3/5, with results presented as stacked bars, and compared between the base model and the fine-tuned variant.

other languages. For each language, we observe that Top-1 performance increases by at least 10% relative to the baseline. Top-5 performance is consistent across all three languages, with scores above 98%, indicating that the task can be fully solved in other languages as well.

For low-resource languages, we find that base performance is lower, but fine-tuning on other languages still yields higher results on those languages for Top-1 and Top-5. Improvements indicate that fine-tuning does not merely adapt the model to a specific language but instead enriches it with transferable legal-domain representations.

5.6 Main Takeaways

Across all experiments, fine-tuning embedding models on legal-domain data consistently improves metadata-to-document retrieval performance. Monolingual contrastive fine-tuning leads to higher Top-k accuracy across languages and model sizes, indicating that domain-specific supervision helps models better capture the relationship between short metadata queries and their corresponding legislative texts.

Bilingual and cross-lingual evaluations further show that the improvements introduced by fine-tuning are not confined to the training language. Joint training with a high-resource language improves retrieval robustness for a lower-resource language. In contrast, cross-lingual evaluation shows that fine-tuned models generalize better than their original counterparts to unseen languages. Together, these observations suggest that fine-tuning primarily enhances content-level legal representations rather than relying on language-specific signals.

6 Conclusion

In this paper, we introduced LEMUR, a large-scale multilingual corpus of EU environmental legislation derived from official EUR-Lex PDFs. We proposed a unified framework for training and evaluating multilingual legal embedding models. To ensure data reliability, we introduced the Lexical Content Score (LCS), a systematic measure of PDF-to-text conversion quality. Using LEMUR, we fine-tuned three state-of-the-art embedding models on five languages from the corpus. We evaluated them on metadata-to-document retrieval, reflecting realistic legal search scenarios.

Our results show that legal-domain contrastive fine-tuning consistently improves retrieval performance across languages and model sizes. Bilingual training further demonstrates that incorporating a high-resource language benefits retrieval in a low-resource setting without degrading high-resource performance. At the same time, cross-lingual evaluation confirms that these gains generalize beyond the training language. Together, these findings indicate that fine-tuning primarily enhances content-level legal representations rather than language-specific patterns.

Future work will expand LEMUR to additional legal domains and languages and reduce the remaining PDF-to-text noise.

Limitations

Limited topical coverage within EUR-Lex. LEMUR is restricted to *Category 15* and *Subcategory 10* (Environment). While this yields a focused benchmark, it limits topical diversity and may reduce generalizability to other EUR-Lex categories with different legal styles and terminology. Future work could extend collection and fine-tuning to additional categories and subcategories.

Limited bilingual fine-tuning coverage. Bilingual multi-positive fine-tuning is evaluated only on one language pair (EN–LV). Although this setting provides initial insights into bilingual training behavior, it does not explore the full range of possible language combinations available in LEMUR. Extending experiments to additional language pairs and resource configurations remains an important direction for future work.

Noise from PDF-to-text conversion. Although conversion quality is validated against HTML, the average lexical similarity across languages is about 0.94, indicating remaining extraction noise. Such noise can affect both fine-tuning and retrieval performance, particularly for older documents and lower-resource languages. Exploring alternative conversion pipelines and layout-aware post-processing could further improve text fidelity.

Acknowledgement

This work is supported by the Genial4KMU project, Universität Hamburg, funded by BMBF (grant no. 16IS24044B).

References

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2025. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. *ACM Comput. Surv.*, 58(6). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - A Multi-Lingual And Multi-Label Legal Document Classification Dataset For Zero-Shot Cross-Lingual Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets Straight Out Of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. 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)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Chroma Team. 2025. Chroma: Open-Source Search And Retrieval Database For AI Applications. Https://github.com/chroma-core/chroma.

Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2023. German BERT Model for Legal Named Entity Recognition. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, pages 723–728. SCITEPRESS - Science and Technology Publications.

PyMuPDF Developers. 2021. PyMuPDF: Python Bindings for the MuPDF Library. Https://pymupdf.readthedocs.io.

Mo El-Haj and Saad Ezzini. 2024. The Multilingual Corpus of World's Constitutions (MCWC). In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation @ LREC-COLING 2024*, pages 57–66, Torino, Italia. ELRA and ICCL.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.10997.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yunhsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *arXiv preprint*. ArXiv:1705.00652.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 29217–29234. Curran Associates, Inc.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, virtual.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 1035–1044, New York, NY, USA. Association for Computing Machinery.

Haitao Li, Qingyao Ai, Xinyan Han, Jia Chen, Qian Dong, and Yiqun Liu. 2025. DELTA: Pre-train a Discriminative Encoder for Legal Case Retrieval via Structural Word Alignment. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25, Philadelphia, PA, USA. AAAI Press.

Nut Limsopatham. 2021. Effectively Leveraging BERT for Legal Document Classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 210–216, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter W. J. Staar. 2025. Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion. *arXiv preprint*. ArXiv:2501.17887.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. LeCaRD: A Legal Case Retrieval Dataset for Chinese Law System. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2342–2348, New York, NY, USA. Association for Computing Machinery.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. MultiLegalPile: A 689GB Multilingual Legal Corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand. Association for Computational Linguistics.

Jake Poznanski, Aman Rangapur, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models. *arXiv preprint*. ArXiv:2502.18443.

Wisam A. Qader, Musa M. Ameen, and Bilal I. Ahmed. 2019. An Overview of Bag of Words; Importance, Implementation, Applications, and Challenges. In *2019 International Engineering Conference (IEC)*, pages 200–204.

Markus Reuter, Tobias Lingenberg, Rūta Liepiņa, Francesca Lagioia, Marco Lippi, Giovanni Sartor, Andrea Passerini, and Burcu Sayin. 2025. Towards Reliable Retrieval in RAG Systems for Large Legal Datasets. *arXiv preprint*. ArXiv:2510.06999.

Yanran Tang, Ruihong Qiu, Xue Li, and Zi Huang. 2025. ReaKase-8B: Legal Case Retrieval via Knowledge and Reasoning Representations with LLMs. *arXiv preprint*. ArXiv:2510.26178.

Yixuan Tang and Yi Yang. 2025. Do We Need Domain-Specific Embedding Models? An Empirical Investigation. *arXiv preprint*. ArXiv:2409.18511.

Unstructured Team. 2023. Unstructured: Open-Source Preprocessing Library. Https://github.com/Unstructured-IO/unstructured.

Rohit Upadhya and Santosh T.y.s.s. 2025. LexCLiPR: Cross-Lingual Paragraph Retrieval from Legal Judgments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13971–13993, Vienna, Austria. Association for Computational Linguistics.

Voyage AI. 2024. Voyage Law 2 - Embedding Model. Https://blog.voyageai.com/2024/04/15/domain-specific-embeddings-and-retrieval-legal-edition-voyage-law-2/.

Yen Thi-Hai Vuong, Quan Minh Bui, Ha-Thanh Nguyen, Thi-Thu-Trang Nguyen, Vu Tran, Xuan-Hieu Phan, Ken Satoh, and Le-Minh Nguyen. 2022. SM-BERT-CR: A Deep Learning Approach for Case Law Retrieval with Supporting Model. *Artif. Intell. Law*, 31(3):601–628.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint*. ArXiv:2402.05672.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv preprint*. ArXiv:2303.17564.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388.

Kaiyan Zhao, Qiyu Wu, Xin-Qiang Cai, and Yoshimasa Tsuruoka. 2024. Leveraging Multi-lingual Positive Instances in Contrastive Learning to Improve Sentence Embedding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 976–991, St. Julian's, Malta. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning For Law And The CaseHOLD Dataset Of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pages 159–168, New York, NY, USA. Association for Computing Machinery.

A Average Content Score for each language

This table presents the average content score for each language.

Language	Avg. Content Score
English (EN)	0.9740
Spanish (ES)	0.9734
Dutch (NL)	0.9673
Bulgarian (BG)	0.9671
French (FR)	0.9608
Romanian (RO)	0.9598
Irish (GA)	0.9588
Portuguese (PT)	0.9539
Hungarian (HU)	0.9533
Swedish (SV)	0.9520
German (DE)	0.9487
Italian (IT)	0.9463
Croatian (HR)	0.9456
Slovenian (SL)	0.9371
Polish (PL)	0.9376
Czech (CS)	0.9323
Slovak (SK)	0.9294
Greek (EL)	0.9135
Latvian (LV)	0.9078
Lithuanian (LT)	0.9067
Finnish (FI)	0.9065
Estonian (ET)	0.8991
Maltese (MT)	0.8027

Table 2: Average lexical content similarity between JSONL and HTML documents across languages.

B Content Score Comparison Across PDF-to-Text Conversion Methods

This appendix provides a comparison of PDF-to-text conversion quality across languages and conversion pipelines. Figure 6 reports the average Content Score for each language in LEMUR, computed separately for the three conversion methods used in our study: OLMOCR, PyMuPDF, and Unstructured. Scores are averaged over all documents available for a given language and method.

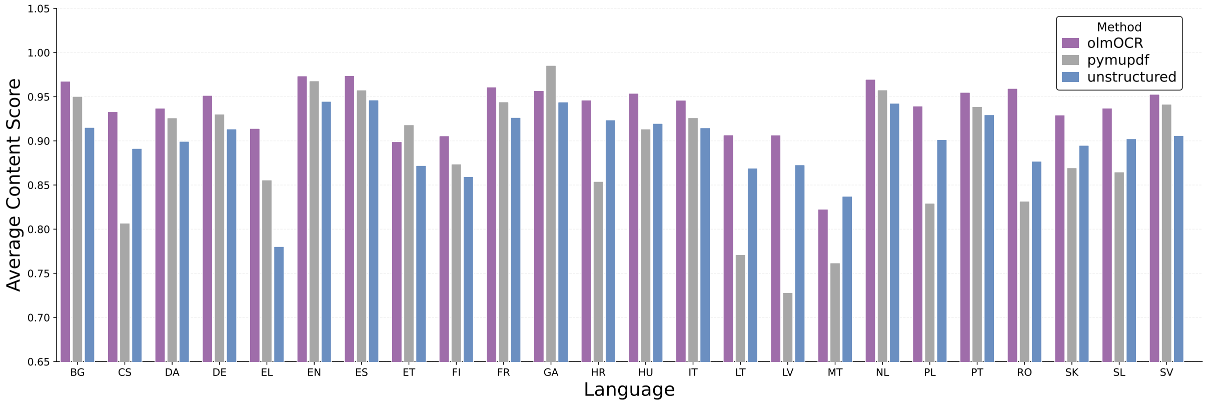


Figure 6: Average Content Score per language for three PDF-to-text conversion methods. Scores are averaged over all documents available for each language.

C Content Score by Year and Dataset Coverage

This appendix reports how PDF-to-text conversion quality varies over time and how document availability is distributed across publication years. Figure 7 plots (i) the average Content Score aggregated per year (left axis) and (ii) the corresponding percentage of files per year (right axis).

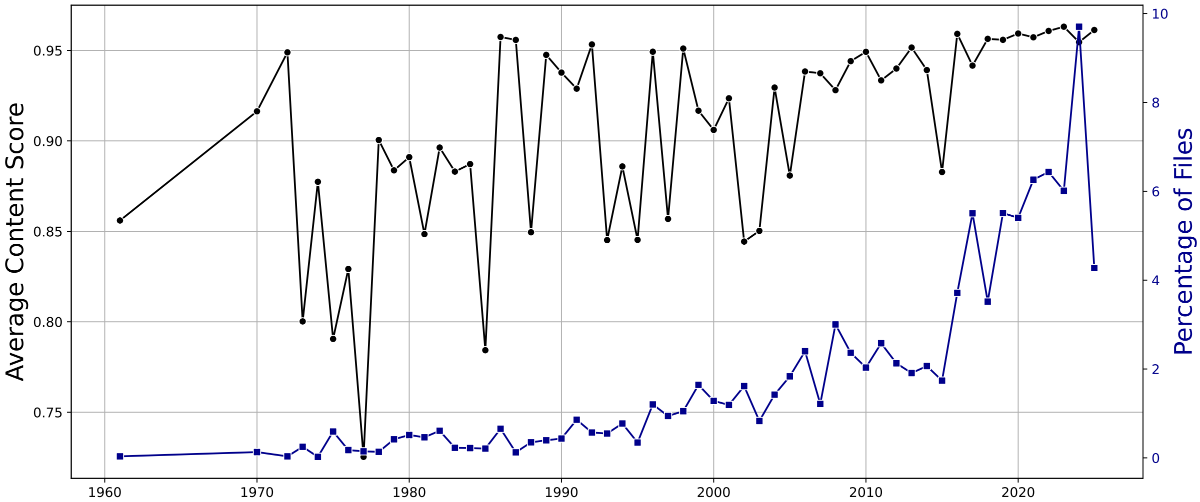
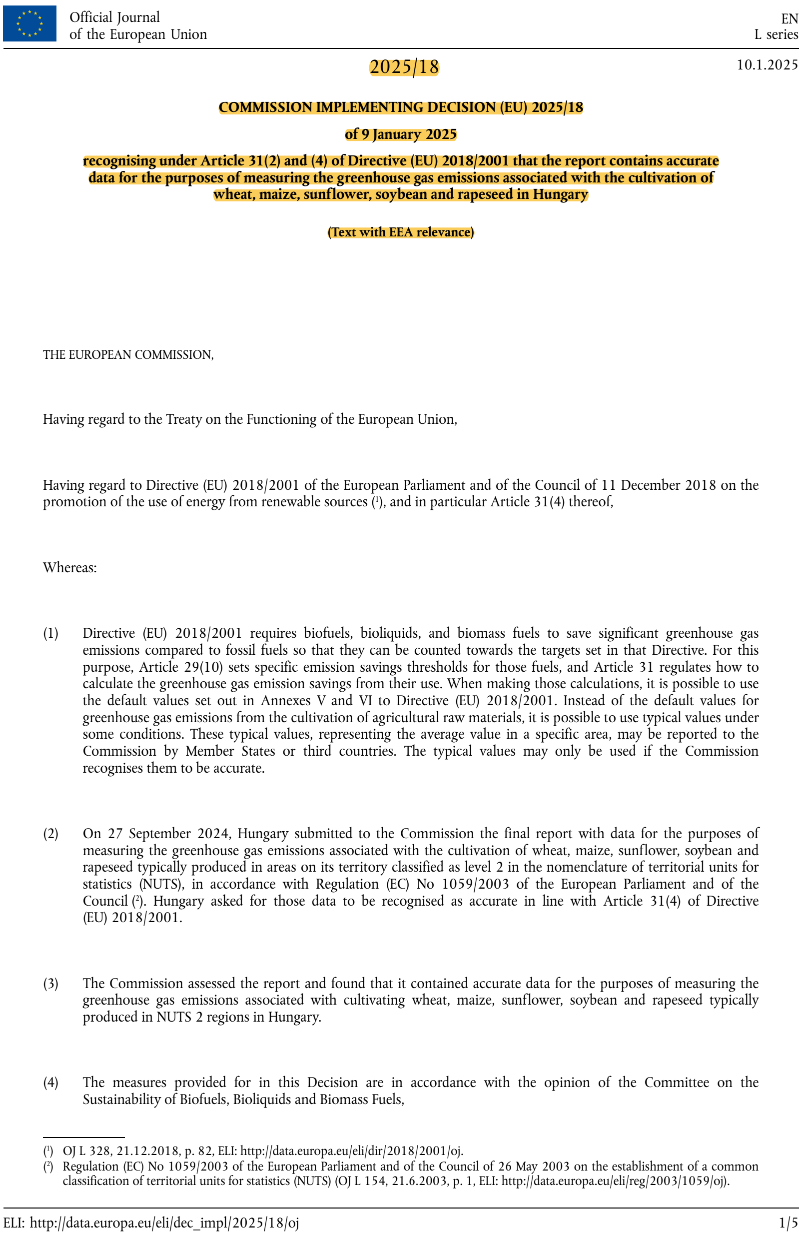


Figure 7: Average Content Score (left axis) and percentage of files (right axis) by publication year.

D Example Metadata–Document Pair

This appendix illustrates the metadata–document structure used throughout LEMUR. For each legislative act, the introductory metadata block is extracted and used as the retrieval query, while the remaining substantive legislative text constitutes the retrieval target. Figure 8 shows a concrete example of this split for a single EU legislative document.



Official Journal of the European Union | EN L series

2025/18 | 10.1.2025

COMMISSION IMPLEMENTING DECISION (EU) 2025/18

of 9 January 2025

recognising under Article 31(2) and (4) of Directive (EU) 2018/2001 that the report contains accurate data for the purposes of measuring the greenhouse gas emissions associated with the cultivation of wheat, maize, sunflower, soybean and rapeseed in Hungary

(Text with EEA relevance)

THE EUROPEAN COMMISSION,

Having regard to the Treaty on the Functioning of the European Union,

Having regard to Directive (EU) 2018/2001 of the European Parliament and of the Council of 11 December 2018 on the promotion of the use of energy from renewable sources [1], and in particular Article 31(4) thereof,

Whereas:

(1) Directive (EU) 2018/2001 requires biofuels, bioliquids, and biomass fuels to save significant greenhouse gas emissions compared to fossil fuels so that they can be counted towards the targets set in that Directive. For this purpose, Article 29(10) sets specific emission savings thresholds for those fuels, and Article 31 regulates how to calculate the greenhouse gas emission savings from their use. When making those calculations, it is possible to use the default values set out in Annexes V and VI to Directive (EU) 2018/2001. Instead of the default values for greenhouse gas emissions from the cultivation of agricultural raw materials, it is possible to use typical values under some conditions. These typical values, representing the average value in a specific area, may be reported to the Commission by Member States or third countries. The typical values may only be used if the Commission recognises them to be accurate.

(2) On 27 September 2024, Hungary submitted to the Commission the final report with data for the purposes of measuring the greenhouse gas emissions associated with the cultivation of wheat, maize, sunflower, soybean and rapeseed typically produced in areas on its territory classified as level 2 in the nomenclature of territorial units for statistics (NUTS), in accordance with Regulation (EC) No 1059/2003 of the European Parliament and of the Council [2]. Hungary asked for those data to be recognised as accurate in line with Article 31(4) of Directive (EU) 2018/2001.

(3) The Commission assessed the report and found that it contained accurate data for the purposes of measuring the greenhouse gas emissions associated with cultivating wheat, maize, sunflower, soybean and rapeseed typically produced in NUTS 2 regions in Hungary.

(4) The measures provided for in this Decision are in accordance with the opinion of the Committee on the Sustainability of Biofuels, Bioliquids and Biomass Fuels,

[1] OJ L 328, 21.12.2018, p. 82, ELI: http://data.europa.eu/eli/dir/2018/2001/oj.
[2] Regulation (EC) No 1059/2003 of the European Parliament and of the Council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS) (OJ L 154, 21.6.2003, p. 1, ELI: http://data.europa.eu/eli/reg/2003/1059/oj).

ELI: http://data.europa.eu/eli/dec_impl/2025/18/oj | 1/5

Figure 8: Example of a metadata–document pair in LEMUR.

E Test Set Size for Each Language

This table reports the number of test queries available for each language used in the retrieval evaluation.

Language	**Test Set Size**
English (EN)	227
German (DE)	226
French (FR)	224
Latvian (LV)	192
Maltese (MT)	187

Table 3: Number of test queries per language used in the retrieval evaluation.

F Cross-Lingual Results for E5 and Qwen-4B

This appendix presents additional cross-lingual retrieval results for the `E5-Multilingual` and `Qwen3-4B` models. Figures 9 and 10 report Acc@k ($k \in \{1, 3, 5\}$) for five target languages when models are fine-tuned on a single source language and evaluated cross-lingually without further adaptation.

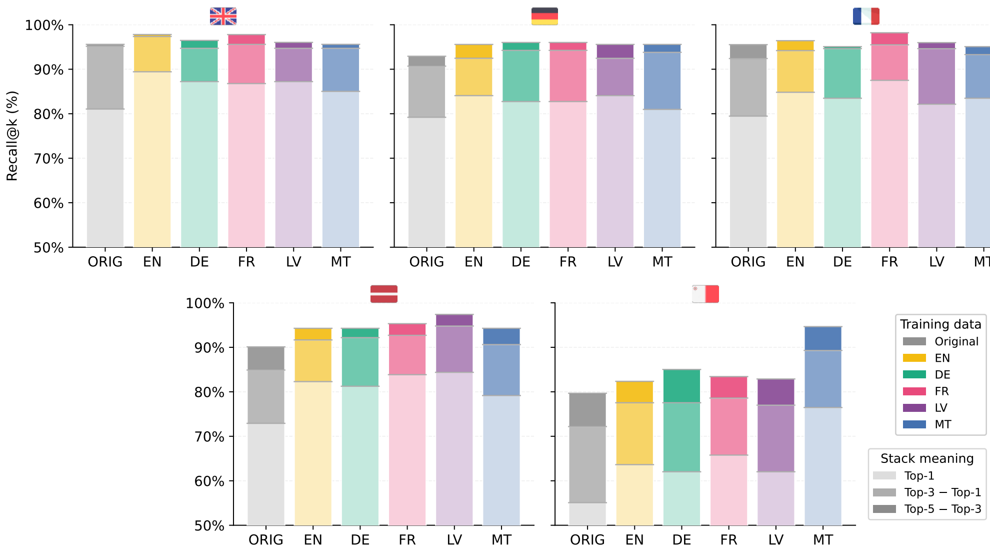


Figure 9: Cross-lingual fine-tuning for E5 on five languages (EN, DE, FR, LV, MT). Performance is measured using Acc@k for 1/3/5, with results presented as stacked bars, and compared between the base model and the fine-tuned variant.

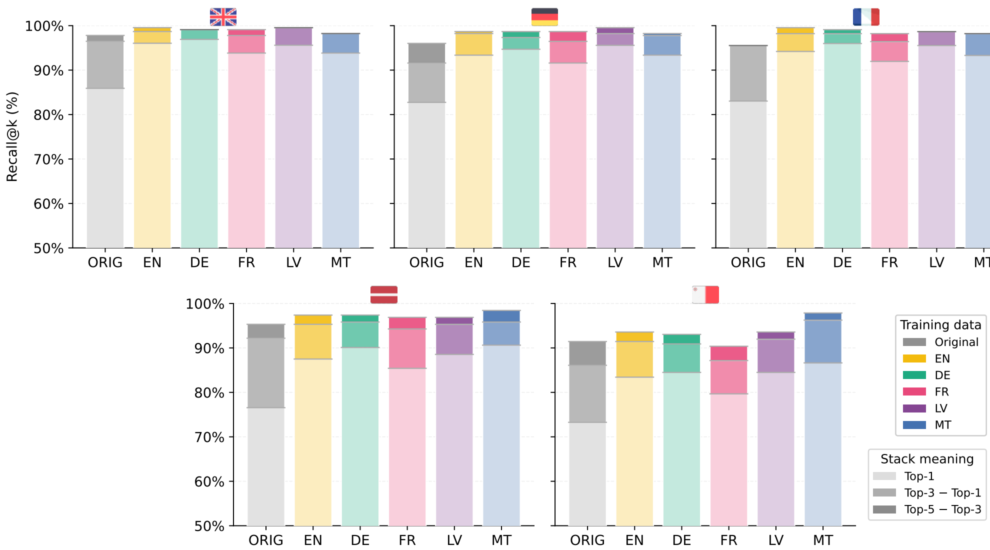


Figure 10: Cross-lingual fine-tuning for Qwen3-4B on five languages (EN, DE, FR, LV, MT). Performance is measured using Acc@k for 1/3/5, with results presented as stacked bars, and compared between the base model and the fine-tuned variant.

G Monolingual Retrieval Performance with Test Queries over the Full Collection

Figure 11 shows monolingual retrieval performance when test queries are evaluated against the full document collection, including training, validation, and test documents. Results compare pretrained and fine-tuned models across five languages and three embedding backbones.

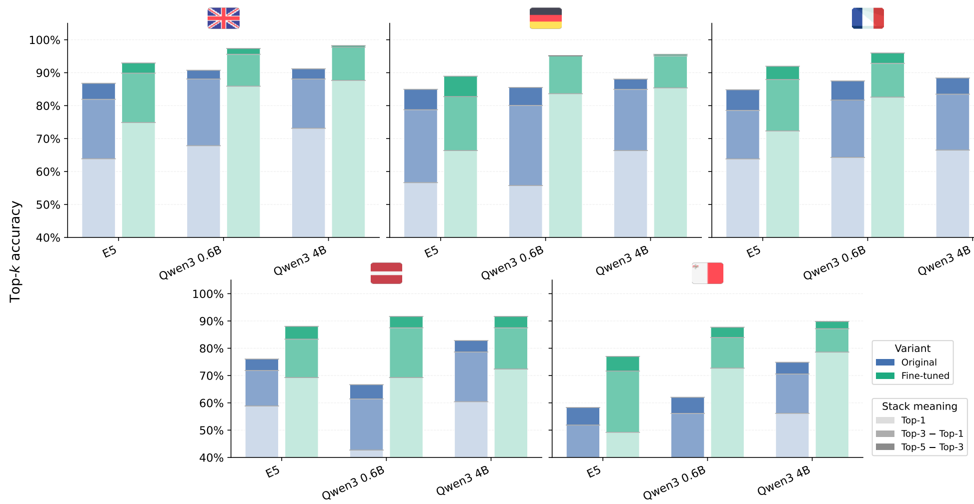


Figure 11: Monolingual fine-tuning of three embedding models (E5, Qwen-0.6B & Qwen-4B) on five languages (EN, DE, FR, LV, MT). Performance is measured using Acc@k for 1/3/5 on test queries, evaluated against the full document collection, and is represented as stacked bars, with comparisons between the base model and the fine-tuned variant.

H Retrieval Results by Language

This appendix reports monolingual retrieval results across eighteen languages. Table 4 shows fine-tuned and original model performance for Acc@*1* and Acc@*5* when test queries are evaluated against **test documents only**, while Table 5 shows results when test queries are evaluated against **all documents**.

Language	Fine-tuned Top 1	Fine-tuned Top 5	Original Top 1	Original Top 5
Spanish (ES)	83.03	96.87	77.23	92.18
Dutch (NL)	84.95	96.01	79.64	95.57
Bulgarian (BG)	80.31	96.27	73.40	88.29
Romanian (RO)	86.55	98.38	77.95	93.54
Irish (GA)	69.76	97.67	39.53	67.44
Portuguese (PT)	81.61	95.96	77.57	91.92
Hungarian (HU)	83.85	93.75	72.39	90.62
Swedish (SV)	89.20	99.53	81.22	95.30
Italian (IT)	85.77	95.55	75.55	90.66
Croatian (HR)	82.08	93.06	76.30	86.70
Slovenian (SL)	83.58	93.84	77.43	91.28
Polish (PL)	85.56	97.93	77.83	95.87
Czech (CS)	86.08	96.90	78.35	94.84
Slovak (SK)	85.64	96.92	77.43	94.35
Greek (EL)	77.00	95.00	74.00	90.00
Lithuanian (LT)	82.98	93.81	65.46	82.47
Finnish (FI)	80.46	93.02	69.76	87.44
Estonian (ET)	89.06	97.39	78.12	93.75

Table 4: Retrieval results for test queries evaluated against test documents only.

Language	Fine-tuned Top 1	Fine-tuned Top 5	Original Top 1	Original Top 5
Spanish (ES)	69.64	90.17	57.58	81.69
Dutch (NL)	66.37	89.82	55.30	79.64
Bulgarian (BG)	62.23	84.04	55.85	77.65
Romanian (RO)	66.12	89.78	62.36	85.48
Irish (GA)	34.88	76.74	23.25	48.83
Portuguese (PT)	65.47	87.89	54.26	80.71
Hungarian (HU)	63.02	87.50	50.52	78.64
Swedish (SV)	69.01	90.61	61.03	85.91
Italian (IT)	71.55	89.77	56.88	83.55
Croatian (HR)	63.00	85.54	57.22	78.03
Slovenian (SL)	68.71	88.20	56.41	77.94
Polish (PL)	69.58	89.69	58.76	85.05
Czech (CS)	72.68	90.72	61.34	85.05
Slovak (SK)	70.76	87.17	60.51	84.61
Greek (EL)	60.00	85.50	58.00	79.00
Lithuanian (LT)	61.34	81.95	50.00	75.77
Finnish (FI)	64.65	85.11	53.48	73.95
Estonian (ET)	71.35	92.70	60.41	85.41

Table 5: Retrieval results for test queries evaluated against all documents.