LexCLiPR: Cross-Lingual Paragraph Retrieval from Legal Judgments

Rohit Upadhya*, **Santosh T.Y.S.S***
School of Computation, Information, and Technology
Technical University of Munich, Germany

Abstract

Efficient retrieval of pinpointed information from case law is crucial for legal professionals but challenging due to the length and complexity of legal judgments. Existing works mostly often focus on retrieving entire cases rather than precise, paragraph-level information. Moreover, multilingual legal practice necessitates cross-lingual retrieval, most works have been limited to monolingual settings. To address these gaps, we introduce LexCLiPR, a cross-lingual dataset for paragraph-level retrieval from European Court of Human Rights (ECtHR) judgments, leveraging multilingual case law guides and distant supervision to curate our dataset. We evaluate retrieval models in a zero-shot setting, revealing the limitations of pre-trained multilingual models for cross-lingual tasks in low-resource languages and the importance of retrieval based post-training strategies. In fine-tuning settings, we observe that two-tower models excel in cross-lingual retrieval, while siamese architectures are better suited for monolingual tasks. Fine-tuning multilingual models on native language queries improves performance but struggles to generalize to unseen legal concepts, highlighting the need for robust strategies to address topical distribution shifts in the legal queries. [1].

1 Introduction

Searching for relevant information in case law is a fundamental yet time-consuming task for legal professionals, as the case law serves as a primary source of legal precedent, containing interpretations and applications of statutes that shape future rulings and guide legal reasoning. These judgments are typically dense and lengthy, filled with nuanced language and complex arguments, making it challenging to locate the exact information needed. Research indicates that lawyers spend approximately 15 hours per week reviewing case law, (Lastres, 2015), which accounts for almost 30% of their annual working hours (Poje, 2014), drawing valuable time away from critical tasks. This underscores the need for advanced retrieval systems capable of pinpointing relevant information efficiently so it could significantly reduce research time, allowing legal professionals to focus on deeper analysis, ultimately enhancing their productivity.

Legal information retrieval poses unique challenges that go beyond those of traditional IR due to the complexity and specificity inherent in legal texts. Unlike general documents, legal judgments contain highly structured reasoning, specialized terminology, and intricate argumentation, often tied to jurisdiction-specific doctrines and interpretative frameworks. Moreover, each paragraph's relevance depend on the contextual understanding of specific legal issues or doctrines, demanding retrievers to go beyond keyword matching to provide meaningful results. Legal queries are typically complex, often requiring systems to interpret layered, implicit legal questions that span multiple legal aspects and contexts. Furthermore, case law is dynamic, with laws evolving and interpretations shifting over time, leading to an ever evolving array of legal concepts, making it essential for retrievers to comprehend new queries and determine relevance.

While much of the research in legal retrieval has focused primarily on retrieving entire cases based on complete-case queries aimed at identifying similar precedents (Santosh et al., 2024b; Ma et al., 2021; Mandal et al., 2017; Goebel et al., 2023; Joshi et al., 2023), there has been a recent shift towards finer-grained retrieval tasks such as paragraph-based retrieval from lengthy legal documents, where either the entire case (Rabelo et al., 2022) or specific, targeted legal queries serve as the query input (Santosh et al., 2024c). This paragraph-

[1]The dataset and code is available at `https://github.com/rohit-upadhya/lexclipr`

*These authors contributed equally to this work.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13971–13993
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

level retrieval addresses the need for retrieving precise information from within a judgment, aligning more closely with the detailed, context-specific information that legal professionals often require. Despite these advances, most studies in legal retrieval have been monolingual, where queries and document corpora share the same language.

As legal practice globalizes, retrieval systems must support diverse linguistic needs, especially in multi-jurisdictional settings. Supra-national courts like the ECHR, CJEU, ICC, and AfCHPR, as well as national courts in multilingual countries like India, serve regions where professionals may submit queries in their native languages, even if judgments are in one of the court's primary languages. These demands underscore the need for retrieval models that bridge language barriers and map queries to relevant content accurately across languages. Addressing cross-lingual queries requires advanced retrieval systems that navigate both linguistic and legal complexities, ensuring accessible legal knowledge for diverse communities.

To investigate the ability of current retrieval models to identify relevant paragraphs in cross-lingual way, a high-quality labeled dataset is imperative. In this study, we employ distant supervision to construct LexCLiPR, a dataset tailored for query-based relevant paragraph extraction from European Court of Human Rights (ECtHR) judgments, which address alleged violations of rights protected under the European Convention on Human Rights. We leverage the ECtHR's Knowledge Sharing platform, which provides case law guides in multiple languages, to curate a cross-lingual dataset spanning seven languages. Using the section headers from these guides as queries, we draw on the discussions under each section, which contain paragraph-level citations to English ECtHR judgments, as our relevance signal. Furthermore, we design dataset splits to evaluate the generalizability of retrieval systems on new legal concepts (not seen during training), providing insights into how well these models adapt to the dynamic nature of law.

We evaluate the performance of current multilingual models on our cross-lingual retrieval task in a zero-shot setting. Our findings indicate that: (i) general pre-trained models (e.g., mBERT) and those further pre-trained on legal corpora (e.g., mLegalBERT) underperform compared to models further fine-tuned on general retrieval datasets (e.g:, mDPR i.e., mBERT fine-tuned on mMARCO). This highlights the importance of retrieval-specific fine-tuning. (ii) Multilingual models perform better with English-translated queries than with native-language queries, highlighting significant challenges in cross-lingual semantic alignment, particularly for low-resource languages. (iii) Monolingual models such as BERT and DPR consistently outperform multilingual models, even with translated queries, underscoring the trade-off between broader multilingual coverage and language-specific depth in multilingual models.

Further, we fine-tune monolingual models on English-translated data and multilingual models on both original and translated queries using siamese and two-tower architectures. Our key observations include: (i) Two-tower models generally excel in cross-lingual retrieval, while siamese architectures are more effective for monolingual retrieval. (ii) mDPR fine-tuned and tested on native-language queries outperforms its performance on translated queries or monolingual models, suggesting that fine-tuning can mitigate language disparities in multilingual models. (iii) However, this advantage diminishes on an unseen query split, indicating the need for more robust strategies to generalize across unseen topic distributions while maintaining language alignment.

2 Related Work

Legal IR Efficiently retrieving critical legal information is essential for lawyers, spanning tasks such as locating relevant legislation either through specific searches or by providing factual descriptions to identify pertinent statutes (Wang et al., 2018; Paul et al., 2022; Louis and Spanakis, 2021; Santosh et al., 2024d). These tasks extend to retrieving similar past cases (Rabelo et al., 2022; Mandal et al., 2017), civil codes (Kim et al., 2016, 2014), litigation documents for tasks like technology-assisted review (Cormack et al., 2010; Baron et al., 2006), patents (Piroi et al., 2013), and within a firm's internal support system (Moens, 2001). Our research centers specifically on legal case retrieval. While most existing legal retrieval systems focus on retrieving entire cases (Sansone and Sperlí, 2022) based on various query types—such as whole cases (Rabelo et al., 2022; Ma et al., 2021; Mandal et al., 2017; Joshi et al., 2023; Santosh et al., 2024b) or legal-specific queries (Locke et al., 2017; Locke and Zuccon, 2018; Koniaris et al., 2016)—our approach retrieves relevant paragraphs at a finer level of granularity, allowing practitioners to access targeted information. At the paragraph level, tasks like the legal case entailment task in COLIEE involve finding paragraphs that align with a new case's decision (Rabelo et al., 2022), using the entire case as a query, unlike the short queries we use, inspired from recent work of Santosh et al. (2024c). This paragraph-level retrieval is critical for building legal question-answering systems (Khazaeli et al., 2021; Sovrano et al., 2021; Verma et al., 2020) and query-focused summarization systems (Santosh et al., 2024a). While prior retrieval studies predominantly focus on monolingual retrieval, where query and documents share the same language, our work advances cross-lingual retrieval for legal documents, addressing multilingual challenges in legal information access.

Cross Lingual IR Cross-Lingual Information Retrieval (CLIR) involves retrieving documents in which search queries and target documents are in different languages (Hull and Grefenstette, 1996). Traditionally, translation-based approaches tackle CLIR by translating either the query or document into a common language, leveraging external machine translation systems or bilingual dictionaries to then use monolingual retrieval methods (McCarley, 1999; Oard, 1998; Zhou et al., 2012; Nair et al., 2020). Recently, neural end-to-end CLIR approaches have emerged, utilizing cross-lingual word embeddings (Vulic and Moens, 2015; Zhang et al., 2019; Litschko et al., 2018). With advances in unsupervised language modeling, models like Multilingual BERT (Devlin, 2018), XLM-R (Conneau, 2019), and Multilingual T5 (Xue, 2020) have been leveraged to extract cross-lingual representations. Transfer learning techniques applied to these cross-lingual embeddings help mitigate the scarcity of non-English data (Van Nguyen et al., 2021; Shi and Lin, 2019; Nair et al., 2020; Schuster et al., 2018). In contrast to the extensive monolingual IR resources, cross-lingual IR datasets are limited. The first CLIR collection emerged with manually translated English queries into German (Salton, 1970). Over time, community-driven evaluations through TREC (Davis and Dunning, 1995; Schäuble and Sheridan, 1998; Voorhees and Harman, 2000), CLEF (Peters, 2019), NCTIR (Kando et al., 1999), and FIRE (Majumder et al., 2010) further enriched CLIR resources. Later, automated pipelines for dataset creation and large multilingual corpora such as Common Crawl enabled the development of datasets like HC4 (Lawrie et al., 2022), HC3 (Lawrie et al., 2023), WikiCLIR (Schamoni et al., 2014), CLIR-Matrix (Sun and Duh, 2020), Large Scale CLIR (Sasaki et al., 2018), and AfriCLIRMatrix (Ogundepo et al., 2022). In this study, we introduce LexCLiPR, a new dataset tailored to advancing CLIR research specifically for legal text collections.

Tasks on ECHR corpora Prior works on the ECtHR corpus have explored diverse tasks, including judgment prediction (Aletras et al., 2016; Chalkidis et al., 2019, 2021; Santosh et al., 2022, 2023a,b, 2024f,e), argument mining (Mochales and Moens, 2008; Mochales and Ieven, 2009; Habernal et al., 2023; Poudyal et al., 2019, 2020; Held and Habernal, 2023), legal reasoning (Chlapanis et al., 2024), event extraction (Filtz et al., 2020; Navas-Loro and Rodriguez-Doncel, 2022), vulnerability classification (Xu et al., 2023), summarization (Santosh et al., 2024a, 2025a), prior case retrieval (Santosh et al., 2024b, 2025b), and relevant paragraph retrieval (Santosh et al., 2024c). While many of these studies focus on judgment documents, recent datasets, like the one by Santosh et al. (2024c) for paragraph retrieval, draw from case law guides curated by the ECtHR registry. Notably, all these works emphasize English. In contrast, we leverage multilingual case law guides to develop our CLIR dataset, contributing a valuable resource to the research community.

3 Dataset

Our task of relevant paragraph extraction from legal judgements is defined as follows: Given a query Q and a judgement document J composed of n paragraphs $P_J = \{p_1, p_2, \ldots, p_n\}$, the objective is to identify the subset of paragraphs $P_J^+ \in P_J$ which are relevant to the query.

3.1 Dataset Curation Pipeline

To create LexCLiPR, we leverage case-law guides from the ECtHR Knowledge Sharing Platform[*], a resource managed by the Court's registry that tracks case law evolution across individual convention articles (e.g., Article 9 - Freedom of Thought, Conscience, and Religion) and transversal themes (e.g., Terrorism, Mass Protests, LGBTI Rights). Below, we outline our pipeline for transforming these case law guides into a structured dataset containing query collections and relevant paragraphs within each referenced ECHR judgement. These

[*] https://www.echr.coe.int/knowledge-sharing

guides, available in multiple languages[*], enable the curation of a cross-lingual dataset. Here, the queries derived from these guides are presented in various languages, while the referenced ECHR judgements remain in English.

Judgements Collection We rely on the recent ECHR case collection from Santosh et al. (2024b), sourced from HUDOC[*], the ECtHR's public database. This collection consists of judgments in English, which we segment into individual paragraphs according to the paragraph numbers located at the beginning of each paragraph, providing unique identifiers for cross-referencing. Following Santosh et al. (2024c), we apply hand-crafted heuristics to manage challenges such as inconsistent HTML structures, nested sub-paragraphs, and spurious numbering introduced by verbatim quoting text from other documents to reference them.

Query Collection The case law guides outline key legal concepts under each article or theme, presenting them in a hierarchical structure, with sub-concepts further detailing each concept. A representative index structure of a case law guide is illustrated in Figure 1. For example, this is the hierarchical path of concepts within the Turkish theme guide of Terör (Terrorism) $\rightarrow$ İstihbarat Aşamasından Eylem Aşamasına Geçiş (Moving on from the surveillance stage to the active phase) $\rightarrow$ Devlet görevlileri tarafından ölümcül güce başvurulması (Use of lethal force by agents of the State) *to* ... $\rightarrow$ Devlet görevlilerinin seçimi ve eğitimi (Training and selection of State agents). We extract this table of content hierarchy from the PDF guides. Then we construct query by concatenating these multiple concepts along the path (from the article or theme title to the leaf node in the PDF structure) by using a delimiter, to maintain clarity by providing context. This approach generates structured queries that closely mirror the types of concept lists legal professionals typically search for and use for indexing cases in legal analytics databases.

Relevant Paragraphs in Judgements Each legal concept in the guides is discussed in detail, with references provided to relevant paragraphs within specific ECtHR judgments. An example of this concept description, with cross-references to relevant paragraphs in specific judgments, is illustrated from Turkish guide in App. Fig. 2. We gather all paragraph references within a specific judgment under each concept and label them as relevant to the associated query within that judgment. It is important to note that not all judgments are referenced in these guides, as the focus is primarily on key cases contributing to substantial developments in the law. Thus, for our dataset, we pair each query exclusively with judgments explicitly referenced in the guides and derive their corresponding relevant paragraphs within those specific judgements. While our methodology could, in theory, be applied across all judgments in the corpus, we intentionally limit each query to only the specific judgments mentioned under the legal concept in the case law guides. This approach to query-judgment pairing ensures high-quality relevance, reducing false negatives in the evaluation setup and enhancing the dataset's reliability.

Finally, we filter the query-judgment pairs to exclude any that lack paragraph-level references. Subsequently, we map each remaining query-judgment pair back to our judgment collection, ensuring that we exclude any references to non-English documents that fall outside our English-only judgment dataset collection.

3.2 Dataset Splits & Analysis

We curate the LexCLiPR dataset across seven languages—English, French, Italian, Romanian, Russian, Turkish, and Ukrainian—utilizing their respective case law guides. This results in a total of 27718 query-judgment pairs, with 7313 unique queries. The distribution of query-judgment pairs across each language is presented in Table 1, showing the highest count in English (7874 query-judgment pairs) and the lowest in Russian (1222 query-judgment pairs). The number of paragraphs in each judgment ranges from 28 to 942, with a mean of 122.46 (Fig. 3a). The percentage of relevant paragraphs in each query-judgment pair varies from 0.11% to 19.64% of the total number of paragraphs in that judgment, with a mean around 2.36%, as depicted in Fig. 4a. The average lengths of queries and paragraphs are 54.19 tokens and 140.04 tokens, respectively, illustrated in Figures 6a and 5a. Detailed language-specific distribution plots can be found in Appendix A.2.

We partition the article/theme case law guides from each language into two distinct splits, ensuring that the first split for each language contains a minimum of 5% of the total query-judgment pairs. This first split consists of query-judgment pairs used exclusively for testing, referred to as "Unseen

[*]Not all guides are available in every language

[*]http://hudoc.echr.coe.int/

	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.
Unique queries	1889	1269	971	1334	207	1223	420
Q-J Pairs	7874	4097	3292	4313	1222	5023	1897
Avg #Para. per judgement	123.24	124.39	127.01	129.46	111.82	116.84	112.76
Avg. % rel para. per Q-J pair	2.41	2.38	2.46	2.34	2.16	2.25	2.35
Unique queries in unseen test	54	43	21	46	48	92	39
Unseen test Q-J pairs	473	346	207	415	192	317	145
Train Q-J Pairs	5887	2965	2450	3085	818	3746	1392
Validation Q-J Pairs	757	393	317	405	106	480	179
Seen Test Q-J Pairs	757	393	318	408	106	480	181

Table 1: Statistics of LexCLiPR dataset. rel., para. denote relevant and paragraphs respectively.

Legal Queries." It evaluates the model's performance on unfamiliar legal concepts that it has not encountered during training. The second split referred to as 'Seen Legal Queries' is further divided into training, validation, and test sets. This test set assesses the model's understanding of familiar legal concepts—those seen during training—when applied to new judgments in the test set. To prevent any leakage of unseen test concepts during training, we ensure that none of the case law guides in the unseen split for any language overlap with the seen split of other languages. This guarantees that all unseen concepts remain entirely unfamiliar, even across languages. This approach enforces stricter unfamiliarity, particularly in the context of using translation-based methods for cross-lingual modeling. Details of the case law guides included in the unseen and seen splits across each language are presented in Appendix 4. Statistics on the number of query-judgment pairs and unique queries in the unseen and seen (training, validation, and test) splits are provided in Table 1.

4 Retrieval Methods

We outline the frameworks utilized in our task. Our approach involves calculating a relevance score for each paragraph in the judgment relative to the query, followed by selecting the top-k most relevant paragraphs based on these scores.

Lexical Retrieval We employ BM25 (Robertson et al., 1995), a bag-of-words model that assesses the relevance of paragraphs to queries by analyzing the presence of query terms within the paragraphs. Although this method cannot be directly applied in our cross-lingual setting due to the mismatch between the languages of the queries and documents, it can be applied in combination with a translation model that converts the queries into English, the language of our document corpus.

Dense Retrieval We employ neural bi-encoders to encode queries and paragraphs into low-dimensional representations capturing their semantic content. The final relevance score is computed using the dot product of the representations from the encoders as $rel(q, p) = E_q(q).E_p(p)$ where E_q and E_p represent query and paragraph encoder respectively. The training objective of retrievers is to learn representations such that relevant pairs of queries and paragraphs exhibit higher similarity than irrelevant ones. To mitigate the computational burden due to the abundance of irrelevant paragraphs, we utilize negative sampling. Let $\{<q_i, p_i^+, p_{i,1}^-, \dots p_{i,n}^->\}_{i=1}^m$ be the training data consisting of m instances with each instance consisting of one query q_i and one relevant passage p_i^+, along with n irrelevant (negative) passages $p_{i,j}^-$. Note these negative paragraphs for a query are sampled from the same document as positive, in our task setup. We optimize the negative log-likelihood loss as follows:

$$L = -log(\frac{\exp(rel(q_i, p_i^+))}{\exp(rel(q, p_i^+)) + \sum_{j=1}^{n} \exp(rel(q, p_{i,j}^-))}) \quad (1)$$

Following Karpukhin et al. 2020, we consider negative samples drawn from randomly selected irrelevant paragraphs to the query, from the same judgment as the positive sample. This approach, Dense Passage Retrieval (DPR) can be implemented in two ways: (i) Siamese (Reimers, 2019; Xiong et al., 2020), which uses a single model to map both the query and document into a shared dense vector space, and (ii) Two-tower (Karpukhin et al., 2020), which employs two independent models to encode the query and document separately into distinct embedding spaces.

5 Experiments

Metrics In accordance with Santosh et al. (2024c), we evaluate performance using Recall@k% (R@K%), which measures the proportion of relevant paragraphs within the top-k% of all paragraphs in a judgment. We report the mean Recall@k% across all instances for $k = \{2, 5, 10\}$. Utilizing k as a percentage rather than an absolute value accommodates the varying number of paragraphs across different judgments. Higher recall scores indicate better performance.

5.1 Zero-shot

Models We assess zero-shot performance using both multilingual and monolingual models. Our multilingual models include (A, B) mBERT (Devlin, 2018), (E, F) mDPR (Zhang et al., 2021, 2022), which is a multilingual adaptation of DPR (Karpukhin et al., 2020) with BERT replaced by mBERT and further fine-tuned on the English MS MARCO dataset (Bajaj et al., 2016), (H, I) mLegal-BERT (Niklaus et al., 2023) which is continually pre-trained XLM-R model on multilingual legal corpus. All these multilingual models evaluated by using queries in their original languages (A, E, H) as well as English-translated queries (B. F. I) to simulate monolingual retrieval. Our monolingual models include (C) BERT (Devlin, 2018) and (G) DPR (Karpukhin et al., 2020), which is trained on the English Natural Questions dataset (Lee et al., 2019; Kwiatkowski et al., 2019); these models are tested using English-translated queries. We also incorporate a lexical retrieval method, BM25 (D), with English-translated queries for comparison. To handle query translation from the original language to English, we employ the NLLB model (Costa-jussà et al., 2022), supporting over 200 languages. Note that mDPR follows a shared Siamese architecture, while DPR is a two-tower architecture.

5.1.1 Results

We report Recall@5% results for seen split queries in Table 2, with the unseen split results in Appendix Table 9. Note that for zero-shot experiments, both splits are treated as unseen, as the models are not fine-tuned on any task-specific data. Additional metrics can be found in Appendix Tables 5-10. In a monolingual retrieval setup, models need to generate well-aligned semantic embeddings within a single language to achieve strong performance. In contrast, for cross-lingual retrieval, models must not only produce well-aligned embeddings within each language but also maintain alignment across languages to effectively handle cross-lingual tasks. Key findings are summarized as follows: (i) Multilingual Models with Native Language Queries (A, E, H): Among multilingual models, mDPR (E) outperforms mBERT (A) and mLegal (H) across most languages, highlighting the benefits of retrieval-specific fine-tuning. This advantage is particularly strong in English, where mDPR benefits from continued fine-tuning on the English MSMARCO retrieval dataset. However, mBERT and mLegal perform better in Turkish and Ukrainian, suggesting that mDPR's English-centric fine-tuning may degrade performance in low-resource languages. These findings emphasize the need for continual training strategies that enhance multilingual capabilities without weakening performance in low-resource languages. (ii) Multilingual Models with Translated Queries (B, F, I): Using English-translated queries improves performance across most languages for multilingual models. While mDPR performs better with English-translated queries (F) than with native ones (E)—which is expected given its fine-tuning on English retrieval corpora—it is notable that mBERT and mLegal (A, H) also perform worse with native queries, despite their multilingual pre-training. This suggests that multilingual models still prioritize English-centric embeddings, likely due to the dominance of English data during pre-training and also highlight significant challenges in cross-lingual semantic alignment, especially for low-resource languages, highlighting the need for improved cross-lingual training objectives to better align multilingual embedding spaces. (iii) Monolingual Models with Translated Queries (C, G): Monolingual models outperform multilingual models even when using translated queries (B, F, I), highlighting the trade-off in multilingual pre-training, which sacrifices depth in individual language representation for broader coverage. Among monolingual models, DPR outperforms BERT, consistent with earlier observations where mBERT and mLegal lag behind mDPR. This reinforces the limitations of standard MLM-based pre-training for retrieval and the need for retrieval-specific fine-tuning strategies. While DPR and BM25 each excel on different query sets (Tab. 2,9), this variability suggests that a hybrid retrieval approach—combining lexical (BM25) and dense (DPR) matching techniques—could effectively handle diverse query types. (iv) Overall,

	Model	Train Data	Model Config	Test Data	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.	Avg.
				Zero-shot								
A	mBERT			Ori.	19.91	17.48	21.62	18.79	16.94	16.46	15.17	17.72
B				Trans.	19.91	19.16	19.26	21.24	19.58	21.54	14.96	19.72
C	BERT			Trans.	23.72	21.65	23.06	28.03	25.18	25.16	**26.41**	24.74
D	BM25			Trans.	25.04	23.17	19.64	25.32	21.12	17.44	17.29	21.29
E	mDPR			Ori.	**29.83**	22.45	20.68	22.75	**29.79**	14.84	17.57	22.56
F				Trans.	**29.83**	**27.55**	**24.70**	27.71	25.22	23.82	18.45	25.33
G	DPR			Trans.	28.85	25.27	23.78	**28.55**	29.47	**26.57**	22.28	**26.40**
H	mLegal			Ori.	18.10	17.16	17.76	21.77	18.39	20.35	21.55	19.30
I				Trans.	18.10	17.98	18.48	21.28	22.15	19.35	17.01	19.19
				Fine-tuning								
a	mBERT	Ori.	Siam.	Ori.	**44.02**	41.67	40.17	41.17	36.79	34.88	38.44	39.59
b				Trans.	**44.02**	42.74	41.27	**48.08**	50.14	42.37	41.14	44.25
c		Trans.		Ori.	40.74	41.53	42.83	39.19	41.64	43.68	41.10	41.53
d				Trans.	40.74	44.42	41.95	43.41	53.20	44.39	46.33	44.92
e		Ori.	Two-tow	Ori.	37.59	39.78	40.03	44.99	51.16	39.23	42.06	42.12
f				Trans.	37.59	40.31	39.58	44.75	49.86	39.86	40.77	41.82
g		Trans.		Ori.	38.22	39.37	40.18	40.48	50.56	39.67	39.78	41.18
h				Trans.	38.22	40.27	37.26	40.75	50.65	41.03	40.47	41.24
i	BERT	Trans.	Siam.	Trans.	40.83	44.25	41.17	44.85	**55.76**	44.11	44.45	45.06
j		Trans.	Two-tow	Trans.	40.46	43.73	**45.30**	45.09	49.26	44.07	44.86	44.68
k	mDPR	Ori.	Siam.	Ori.	42.15	**45.08**	44.95	45.19	48.93	44.16	**48.91**	**45.62**
m				Trans.	42.15	44.74	43.62	41.63	50.93	**47.48**	43.67	44.89
n		Trans.		Ori.	40.63	39.20	42.54	42.64	29.58	42.95	36.45	39.14
o				Trans.	40.63	41.06	42.19	45.54	48.59	41.62	43.79	43.35
p		Ori.	Two-tow	Ori.	40.08	42.89	43.18	46.29	53.37	43.26	41.48	44.36
q				Trans.	40.08	42.92	43.74	46.15	55.11	42.89	41.40	44.61
r		Trans.		Ori.	40.56	40.15	43.86	41.46	43.15	39.99	43.07	41.75
s				Trans.	40.56	40.20	43.56	40.97	44.92	38.64	41.45	41.47
t	DPR	Trans.	Two-tow	Trans.	41.41	43.53	43.99	42.03	51.96	41.28	47.68	44.55
u	mLegal	Ori.	Siam.	Ori.	32.65	37.63	40.16	37.94	45.62	36.55	38.75	38.47
v				Trans.	32.65	34.39	31.83	26.82	31.81	34.77	34.80	32.44
w		Trans.		Ori.	37.18	36.10	39.46	40.64	37.76	05.91	34.49	33.08
x				Trans.	37.18	36.33	39.38	38.36	38.11	36.52	35.09	37.28
y		Ori.	Two-tow	Ori.	39.16	40.49	43.50	41.60	49.86	39.24	41.77	42.23
z				Trans.	39.16	40.13	43.08	42.00	47.66	40.07	42.05	42.02
aa		Trans.		Ori.	37.11	35.95	40.35	38.23	44.87	35.97	39.68	38.88
ab				Trans.	37.11	37.28	43.46	39.56	44.16	39.88	41.71	40.45

Table 2: Recall@5% performance on seen legal queries test split.

monolingual retrieval outperforms cross-lingual retrieval, underscoring the challenges of multilingual representations. Given their complementary strengths, future research should explore ensemble of monolingual and cross-lingual retrieval methods to achieve more robust performance.

5.2 Fine-Tuning

Models We fine-tune three multilingual models—mDPR, mBERT, and mLegal—on our training dataset using queries from all language splits, exploring both Siamese and two-tower architectural framework for each of the model. We experiment

	Model	Train Data	Model Config	Test Data	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.	Avg.
				Fine-tuning								
a	mBERT	Ori.	Siam.	Ori.	27.68	27.89	31.07	29.61	34.78	29.64	35.63	30.90
b				Trans.	27.68	28.75	33.16	30.41	36.02	33.39	37.53	32.42
e		Ori.	Two-tow	Ori.	22.34	22.85	25.86	23.17	26.15	25.12	31.26	25.25
f				Trans.	22.34	22.83	26.12	24.22	27.43	24.67	31.95	25.65
i	BERT	Trans.	Siam.	Trans.	**31.01**	30.14	34.61	**34.86**	**38.77**	32.07	37.07	**34.08**
j		Trans.	Two-tow	Trans.	30.49	**32.44**	**34.98**	32.78	33.90	32.30	**40.63**	33.93
k	mDPR	Ori.	Siam.	Ori.	25.83	28.04	29.79	27.46	33.81	30.08	38.05	30.44
m				Trans.	25.83	31.48	30.17	26.20	38.64	30.29	33.16	30.82
p		Ori.	Two-tow	Ori.	28.34	28.40	30.99	27.97	31.39	33.50	32.99	30.51
q				Trans.	28.34	29.22	30.35	29.44	31.69	**33.56**	34.71	31.04
t	DPR	Trans.	Two-tow	Trans.	28.15	29.33	31.25	30.38	34.92	30.68	34.54	31.32
u	mLegal	Ori.	Siam.	Ori.	22.51	27.79	30.27	28.35	30.00	22.54	36.38	28.26
v				Trans.	22.51	23.39	20.34	22.42	18.11	22.42	20.98	21.45
y		Ori.	Two-tow	Ori.	28.64	28.92	31.72	31.05	29.59	33.43	32.41	30.82
z				Trans.	28.64	29.40	31.95	31.50	29.21	31.67	32.41	30.68

Table 3: Recall@5% performance on unseen legal queries test split.

with two training setups: one using queries in their original languages and another using queries translated into English. After fine-tuning, we evaluate the models on both original-language queries and English-translated queries. Row a in Table 2 and 3 represents mBERT with Siamese architecture trained using queries in their original languages and is then evaluated on queries in the original languages. Row f represents mBERT with Two-Tower architecture, trained using queries in original language and tested on English-translated queries. For monolingual models—BERT and DPR—we fine-tune on the English-translated queries in the training set and also evaluated them on English-translated queries in the test set. For BERT, we experiment with both the siamese and the two-tower, while for DPR, we adhere to its base architecture of two-tower structure. Implementation details are presented in App. B.

5.2.1 Results

We report Recall@5% performance for the seen legal query split in Table 2, with detailed results available in Appendix 5-7. Our key findings are as follows: (i) Multilingual Models Fine-Tuned and Tested with Native Language Queries (a, e, k, p, u, y): Two-tower architectures outperform Siamese models by better aligning semantic embeddings across languages through separate encoders. In contrast, the Siamese setup struggles with cross-lingual alignment due to shared encoding constraints and the dominance of English document data during training. However, mDPR in the Siamese setting performs slightly better than in the two-tower mode, suggesting that retrieval-specific fine-tuning can partially compensate for architectural limitations. Across all models, mDPR remains the strongest performer. (ii) Multilingual Models Fine-Tuned with Native Queries and Tested on English Translated Queries (b, f, m, q, v, z): Models in the two-tower setup (f, q, z) underperform compared to their counterparts evaluated on native language queries (e, p, y). This suggests that fine-tuning on native languages is more effective at overcoming language disparities, particularly for low-resource languages than relying on an English translation-based testing framework. However, in the Siamese setting, performance varies: mBERT performs better on translated queries (a vs. b), while mLegal performs better on native queries (u vs. v).

(iii) Multilingual Models Fine-Tuned and Tested with English-Translated Queries (d, h, o, s, x, ab): The Siamese architecture performs well in this setup, as retrieval is monolingual in nature. Comparing these models to their counterparts fine-tuned and tested on native languages (point (i) above) suggests that fine-tuning on native language data effectively bridges the cross-lingual gap in domain-specific text. This contrasts with zero-shot retrieval, where models perform well with translated queries than in native language. (iv) Multilingual Models Fine-Tuned on English-Translated Queries and Tested with Native language Queries (c, g, n, r, w, aa): As expected, these models perform best with English queries but also achieve comparable results with native language queries. This suggests that pre-training has provided cross-lingual transfer capabilities, allowing legal-domain knowledge learned in English during fine-tuning to transfer to other languages. This transferability is particularly valuable for low-resource languages, where fine-tuning data is limited. (v) Monolingual Models (i, j, t): Monolingual models remain competitive with multilingual models. Notably, mDPR fine-tuned on native language queries (k) performs on par with or better than DPR (t), highlighting the effectiveness of fine-tuning in bridging domain-specific knowledge across low resource languages.

We report Recall@5% performance on the unseen legal query split in Table 3, with detailed results available in Appendix Tables 8-10. (i) Performance on the unseen split shows significant improvement over zero-shot results indicating that fine-tuning helps models learn transferable features that generalize to new topics. However, it is lower than on the seen split, suggesting the need for better handling of query-side distribution shifts and domain adaptation strategies that require minimal labeled data while avoiding overfitting to seen queries.(ii) mBERT (a, b, e, f) performs better in the Siamese setting, while mLegal (u, v, y, z) excels in the two-tower configuration. mDPR (k, m, p, q) achieves comparable performance across both architectures, suggesting that future work should explore when Siamese or two-tower setups are most effective for generalization. (iii) Training on translated queries improves performance for mBERT and mDPR, but not for mLegal (App. 9), possibly due to overfitting to domain-specific linguistic patterns in native language legal texts. (iv) Among monolingual models, BERT (i, j) slightly outperforms DPR (t) and all the multilingual models. The weaker performance of multilingual models highlight the need for more effective methods to improve the generalization of semantic embeddings, for unseen query topics across languages.

6 Conclusion

We introduced LexCLiPR, a cross-lingual dataset designed for the task of paragraph-level retrieval from case law judgements of European Court of Human Rights (ECtHR) based on a legal query. We curate this dataset using the multilingual case-law guides produced by the court's registry with paragraph-level citations to ECtHR judgements. Our experiments in a zero-shot setting revealed significant limitations in pre-trained multilingual models, especially for low-resource languages and underscored the importance of retrieval-specific fine-tuning. We further demonstrated in fine-tuning, that two-tower models excel in cross-lingual retrieval, while siamese architectures are more suited for monolingual tasks. Fine-tuning multilingual models on native language queries improved performance but struggled to generalize to unseen legal concepts, highlighting the need for more advanced strategies to handle distribution shifts. We hope that both our dataset and the fine-tuned models will be useful to the research community working in the space of legal information retrieval.

Limitations

Our experiments are conducted on the LexCLiPR dataset, which spans seven languages and focuses on European Court of Human Rights (ECtHR) judgments. While this dataset offers valuable insights for cross-lingual retrieval tasks, its focus on a single jurisdiction restricts the applicability of our findings to other legal systems that may differ significantly in terminology, structure, interpretative frameworks, and linguistic nuances. To develop more universally applicable retrieval models, future work should expand to broader datasets that capture the diversity of global legal systems and multilingual complexities.

Furthermore, our study focuses exclusively on the pre-fetcher stage of retrieval systems, which is responsible for retrieving potentially relevant paragraphs. Consequently, our evaluation prioritizes recall-based metrics, while we leave an exploration of the re-ranking stage—which emphasizes precision-based metrics—for future work. A notable limitation is our treatment of paragraphs as independent units during training, which disregards the inter-paragraph and cross-document context that is often critical in legal texts. While segmenting documents into shorter chunks for retrieval is a common practice in information retrieval, this approach can strip paragraphs of essential contextual information, such as that provided by citations, sequential structures, and cross-references. This challenge is especially pronounced in the legal domain, where the interplay between paragraphs and documents is fundamental to accurate interpretation and relevance estimation.

Ethics Statement

LexCLiPR dataset, was curated based on the publicly available sources such as case law guides and HuDOC, the official database of the court and it complies with the ECtHR data policy. These decisions, although not anonymized, include the real names of individuals involved. However, our work does not engage with the data in a way that we consider harmful beyond this availability. We acknowledge the potential for biases inherent in legal data, which may arise from systemic factors or representational imbalances in the dataset. It is crucial to scrutinize these biases thoroughly to ensure that the systems developed promote fairness and do not inadvertently reinforce existing inequalities.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv e-prints*, pages arXiv–1611.

Jason R Baron, David D Lewis, and Douglas W Oard. 2006. Trec 2006 legal track overview. In *TREC*. Citeseer.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241.

Odysseas S Chlapanis, Dimitrios Galanis, and Ion Androutsopoulos. 2024. Lar-echr: A new legal argument reasoning task and dataset for cases of the european court of human rights. *arXiv preprint arXiv:2410.13352*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. 2010. Overview of the trec 2010 legal track. In *TREC*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Mark W Davis and Ted E Dunning. 1995. A trec evaluation of query translation methods for multi-lingual text retrieval. In *Fourth Text Retrieval Conference*, volume 483.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Erwin Filtz, María Navas-Loro, Cristiana Santos, Axel Polleres, and Sabrina Kirrane. 2020. Events matter: Extraction of events from court decisions. *Legal Knowledge and Information Systems*, pages 33–42.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the competition on legal information, extraction/entailment (coliee) 2023. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 472–480.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, pages 1–38.

Lena Held and Ivan Habernal. 2023. Lacour!: Enabling research on argumentation in hearings of the european court of human rights. *arXiv preprint arXiv:2312.05061*.

David A Hull and Gregory Grefenstette. 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–57.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella, and Ashutosh Modi. 2023. U-creat: Unsupervised case retrieval using events extraction. *arXiv preprint arXiv:2307.05260*.

Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka.

1999. Overview of ir tasks at the first ntcir workshop. In *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition*, pages 11–44.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Soha Khazaeli, Janardhana Punuru, Chad Morris, Sanjay Sharma, Bert Staub, Michael Cole, Sunny Chiu-Webster, and Dhruv Sakalley. 2021. A free format legal question answering system. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 107–113.

Mi-Young Kim, Ying Xu, Randy Goebel, and Ken Satoh. 2014. Answering yes/no questions in legal bar exams. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2013 Workshops, LENLS, JURISIN, MiMI, AAA, and DDS, Kanagawa, Japan, October 27–28, 2013, Revised Selected Papers 5*, pages 199–213. Springer.

Mi-Young Kim, Ying Xu, Yao Lu, and Randy Goebel. 2016. Legal question answering using paraphrasing and entailment analysis. In *Tenth International Workshop on Juris-informatics (JURISIN)*.

Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. 2016. Multi-dimension diversification in legal information retrieval. In *Web Information Systems Engineering–WISE 2016: 17th International Conference, Shanghai, China, November 8-10, 2016, Proceedings, Part I 17*, pages 174–189. Springer.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Steven A Lastres. 2015. Rebooting legal research in a digital age.

Dawn Lawrie, James Mayfield, Douglas W Oard, and Eugene Yang. 2022. Hc4: A new suite of test collections for ad hoc clir. In *European Conference on Information Retrieval*, pages 351–366. Springer.

Dawn Lawrie, James Mayfield, Douglas W Oard, Eugene Yang, Suraj Nair, and Petra Galuščáková. 2023. Hc3: A suite of test collections for clir evaluation over informal text. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2880–2889.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1253–1256.

Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1261–1264.

Daniel Locke, Guido Zuccon, and Harrisen Scells. 2017. Automatic query generation from legal texts for case law retrieval. In *Information Retrieval Technology: 13th Asia Information Retrieval Societies Conference, AIRS 2017, Jeju Island, South Korea, November 22-24, 2017, Proceedings 13*, pages 181–193. Springer.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Antoine Louis and Gerasimos Spanakis. 2021. A statutory article retrieval dataset in french. *arXiv preprint arXiv:2108.11792*.

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*, pages 2342–2348.

Prasenjit Majumder, Mandar Mitra, Dipasree Pal, Ayan Bandyopadhyay, Samaresh Maiti, Sukomal Pal, Deboshree Modak, and Sucharita Sanyal. 2010. The fire 2008 evaluation exercise. *ACM Transactions on Asian Language Information Processing (TALIP)*, 9(3):1–24.

Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, and Saptarshi Ghosh. 2017. Overview of the fire 2017 irled track: Information retrieval from legal documents. In *FIRE (Working Notes)*, pages 63–68.

J Scott McCarley. 1999. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 208–214.

Raquel Mochales and Aagje Ieven. 2009. Creating an argumentation corpus: do theories apply to real arguments? a case study on the legal argumentation of the echr. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 21–30.

Raquel Mochales and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In *Proceedings of the 2008 conference on legal knowledge and information systems*, pages 11–20.

Marie-Francine Moens. 2001. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law*, 9:29–57.

Suraj Nair, Petra Galuscakova, and Douglas W Oard. 2020. Combining contextualized and non-contextualized query translations to improve clir. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1581–1584.

Marıa Navas-Loro and Vıctor Rodriguez-Doncel. 2022. Whenthefact: Extracting events from european legal decisions. In *Legal Knowledge and Information Systems*, pages 219–224. IOS Press.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

Douglas W Oard. 1998. A comparative study of query and document translation for cross-language information retrieval. In *Conference of the Association for Machine Translation in the Americas*, pages 472–483. Springer.

Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. 2022. Africlirmatrix: Enabling cross-lingual information retrieval for african languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8721–8728.

Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. 2022. Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11139–11146.

Carol Peters. 2019. Information retrieval evaluation in a changing world lessons learned from 20 years of clef.

Florina Piroi, Mihai Lupu, and Allan Hanbury. 2013. Overview of clef-ip 2013 lab: Information retrieval in the patent domain. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization: 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings 4*, pages 232–249. Springer.

Joshua Poje. 2014. Legal research. *American Bar Association Techreport*, 2014.

Prakash Poudyal, Teresa Gonçalves, and Paulo Quaresma. 2019. Using clustering techniques to identify arguments in legal documents. *ASAIL@ ICAIL*, 2385.

Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. Echr: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Gerard Salton. 1970. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

TYS Santosh, Mahmoud Aly, and Matthias Grabmair. 2024a. Lexabsumm: Aspect-based summarization of legal decisions. *arXiv preprint arXiv:2404.00594*.

TYS Santosh, Mahmoud Aly, Oana Ichim, and Matthias Grabmair. 2025a. Lexgenie: Automated generation of structured reports for european court of human rights case law. *arXiv preprint arXiv:2503.03266*.

TYS Santosh, Rashid Gustav Haddad, and Matthias Grabmair. 2024b. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. *arXiv preprint arXiv:2404.00596*.

TYS Santosh, Elvin Quero Hernandez, and Matthias Grabmair. 2024c. Query-driven relevant paragraph extraction from legal judgments. *arXiv preprint arXiv:2404.00595*.

TYS Santosh, Kristina Kaiser, and Matthias Grabmair. 2024d. Cusines: Curriculum-driven structure induced negative sampling for statutory article retrieval. *arXiv preprint arXiv:2404.00590*.

TYSS Santosh, Irtiza Chowdhury, Shanshan Xu, and Matthias Grabmair. 2024e. The craft of selective prediction: Towards reliable case outcome classification–an empirical study on european court of human rights cases. *arXiv preprint arXiv:2409.18645*.

TYSS Santosh, Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024f. Incorporating precedents for legal judgement prediction on european court of human rights cases. *arXiv preprint arXiv:2409.18644*.

Tyss Santosh, Oana Ichim, and Matthias Grabmair. 2023a. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 593–605.

TYSS Santosh, Isaac Misael OlguÃn Nolasco, and Matthias Grabmair. 2025b. Lecopcr: Legal concept-guided prior case retrieval for european court of human rights cases. *arXiv preprint arXiv:2501.14114*.

Tyss Santosh, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023b. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1103–1111.

Tyss Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–494.

Peter Schäuble and Páraic Sheridan. 1998. Cross-language information retrieval (clir) track overview. *NIST SPECIAL PUBLICATION SP*, pages 31–44.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Peng Shi and Jimmy Lin. 2019. Cross-lingual relevance transfer for document retrieval. *arXiv preprint arXiv:1911.02989*.

Francesco Sovrano, Monica Palmirani, Biagio Distefano, Salvatore Sapienza, and Fabio Vitali. 2021. A dataset for evaluating legal question answering on private international law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 230–234.

Shuo Sun and Kevin Duh. 2020. Clirmatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170.

Minh Van Nguyen, Tuan Ngo Nguyen, Bonan Min, and Thien Huu Nguyen. 2021. Crosslingual transfer learning for relation and event extraction via word category and class alignments. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5414–5426.

Aayushi Verma, Jorge Morato, Arti Jain, and Anuja Arora. 2020. Relevant subsection retrieval for law domain question answer system. *Data Visualization and Knowledge Engineering: Spotting Data Points with Artificial Intelligence*, pages 299–319.

Ellen M Voorhees and Donna Harman. 2000. Overview of the sixth text retrieval conference (trec-6). *Information Processing & Management*, 36(1):3–35.

Ivan Vulic and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, volume 2, pages 719–725. ACL; East Stroudsburg, PA.

Pengfei Wang, Ze Yang, Shuzi Niu, Yongfeng Zhang, Lei Zhang, and ShaoZhang Niu. 2018. Modeling dynamic pairwise attention for crime classification over legal articles. In *the 41st international ACM SIGIR conference on research & development in information retrieval*, pages 485–494.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Shanshan Xu, Leon Staufer, TYS Santosh, Oana Ichim, Corina Heri, and Matthias Grabmair. 2023. Vechr: A dataset for explainable and robust classification of vulnerability type in the european court of human rights. *arXiv preprint arXiv:2310.11368*.

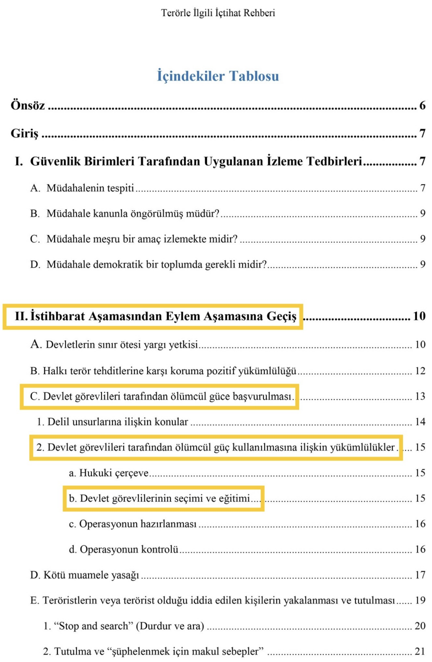
L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Rui Zhang, Caitlin Westerfield, Sungrok Shim, Garrett Bingham, Alexander Fabbri, Neha Verma, William Hu, and Dragomir Radev. 2019. Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. *arXiv preprint arXiv:1906.03492*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. *arXiv preprint arXiv:2108.08787*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *arXiv preprint arXiv:2210.09984*.

Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, 45(1):1–44.



Terörle İlgili İçtihat Rehberi

İçindekiler Tablosu

Figure 1: Illustration of the Table of Contents from the Turkish case law guide on Terrorism, facilitating the derivation of legal concept queries for the LexCLiPR dataset.

A Dataset

A.1 Dataset Curation

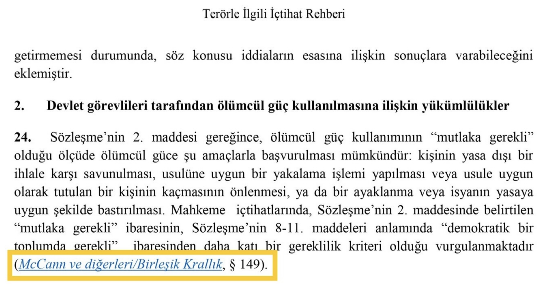
Figure 1 presents a Table of Contents from the Turkish case law guide on Terrorism. This conceptual mapping serves as a foundation for formulating legal concept queries for the LexCLiPR dataset. Figure 2 highlights the structure of case law guides, where each concept is accompanied by explicit references to relevant paragraphs from ECtHR judgments. These references provide relevance signals of paragraphs in judgments that correspond to these legal concepts.

A.2 Dataset Distribution

Figures 3, 4, 5 and 6 present total dataset (all) and language specific distribution of total number of paragraphs per judgement, total number of relevant paragraphs per judgement, number of tokens per query and number of tokens per paragraph respectively.

A.3 Case law Guides and Data split Breakdown

Table 4 presents a detailed breakdown of ECtHR case law guides in various languages, with their usage across different dataset splits. S, U, - indicate that that the case law guide is used in the 'Seen Legal Queries' data split, 'Unseen Legal Queries' data split and unavailable respectively.



Terörle İlgili İçtihat Rehberi

getirmemesi durumunda, söz konusu iddiaların esasına ilişkin sonuçlara varabileceğini eklemiştir.

2. Devlet görevlileri tarafından ölümcül güç kullanılmasına ilişkin yükümlülükler

24. Sözleşme'nin 2. maddesi gereğince, ölümcül güç kullanımının "mutlaka gerekli" olduğu ölçüde ölümcül güce şu amaçlarla başvurulması mümkündür: kişinin yasa dışı bir ihlale karşı savunulması, usulüne uygun bir yakalama işlemi yapılması veya usule uygun olarak tutulan bir kişinin kaçmasının önlenmesi, ya da bir ayaklanma veya isyanın yasaya uygun şekilde bastırılması. Mahkeme içtihatlarında, Sözleşme'nin 2. maddesinde belirtilen "mutlaka gerekli" ibaresinin, Sözleşme'nin 8-11. maddeleri anlamında "demokratik bir toplumda gerekli" ibaresinden daha katı bir gereklilik kriteri olduğu vurgulanmaktadır (*McCann ve diğerleri/Birleşik Krallık*, § 149).

Figure 2: Illustration of Contents of a case law guide, illustrating how legal concepts are discussed with explicit references to relevant paragraphs in ECtHR judgments, enabling the derivation of relevance signals for LexCLiPR.

B Implementation Details

For our zero-shot baseline, we utilize BM25 with hyperparameters $k1 = 1.5$ and $b = 0.75$.. For dense models, we employ max pooling to aggregate the hidden state representations of all tokens from the final layer and use cosine similarity as the similarity function. We use FAISS vector datastore (Johnson et al., 2019) for efficient retrieval. The models are fine-tuned with 7 randomly sampled negatives per query. We sweep through learning rates within the range of $\{1e{-}6, 5e{-}6, 1e{-}5, 5e{-}5\}$ for fine-tuning. The model is trained for 5 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017), and the best model is selected based on the validation results. We use the same setup to train all these dense models: BERT[*], mBERT[*], DPR[*] [*], mDPR[*], mLegalBERT[*] models.

[*] www.huggingface.co/google-bert/bert-base-uncased

[*] www.huggingface.co/google-bert/bert-base-multilingual-cased

[*] www.huggingface.co/facebook/dpr-ctx_encoder-single-nq-base

[*] www.huggingface.co/facebook/dpr-question_encoder-single-nq-base

[*] www.huggingface.co/castorini/mdpr-tied-pft-msmarco

[*] www.huggingface.co/joelniklaus/legal-xlm-roberta-base

ECHR Guide Title	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.
Social Rights	S	S	-	-	-	-	-
Article 6 (Criminal Limb)	S	S	S	S	S	S	S
Rights of LGBTI Persons	S	S	-	-	-	-	-
Article 1 of Protocol No. 1	S	S	-	S	S	S	S
Article 13	S	S	S	S	-	S	-
Prisoners' Rights	S	-	-	S	-	S	-
Article 2 of Protocol No. 4	S	S	-	S	-	-	-
Article 11	S	S	S	S	-	S	-
Article 18	U	U	U	U	-	U	-
Article 4	S	S	S	S	-	S	S
Article 4 of Protocol No. 7	S	S	S	S	S	S	S
Article 5	S	S	S	S	S	S	-
Article 17	S	S	S	S	-	S	-
Article 15	S	S	S	S	S	S	S
Article 14 and Article 1 of Protocol No. 12	S	S	-	S	-	S	-
Article 12	S	S	S	S	-	-	-
Immigration	U	U	-	U	-	U	-
Article 34/35	S	S	S	S	-	-	-
Environment	S	S	-	S	-	-	-
Mass Protests	S	S	-	S	-	S	-
Article 3 of Protocol No. 1	S	S	S	S	-	S	S
Article 8	S	S	-	S	-	S	-
Article 6 (Civil Limb)	S	S	S	S	-	S	S
Article 3	U	U	U	U	U	U	U
Article 1	S	S	S	-	-	S	-
Article 10	S	S	S	S	-	S	-
Article 7	S	S	S	S	S	S	S
Article 9	S	S	S	S	-	-	-
Terrorism	S	S	S	S	-	S	-
Article 2 of Protocol No. 7	S	-	-	S	-	-	-
Article 46	S	S	-	S	-	-	-
Article 3 of Protocol No. 4	S	S	-	S	-	-	-
Article 1 of Protocol No. 7	S	-	-	S	-	-	-
Article 2	S	S	S	S	-	S	S
Data Protection	S	S	S	S	-	-	-
Article 2 of Protocol No. 1	S	S	S	S	-	S	S
Article 4 of Protocol No. 4	S	S	-	S	-	S	S
Article 1 of Protocol No. 7	S	S	-	-	-	-	-
Rights of LGBTI Persons	S	-	-	S	-	-	-

Table 4: List of ECHR Case law guides with their usage across dataset splits and languages. -, S, U represent unavailable, used in the seen split, and used in the unseen split respectively.

Model	**Train Data**	**Model Config**	**Test Data**	**Eng.**	**Fre.**	**Ita.**	**Romn.**	**Rus.**	**Turk.**	**Ukr.**	**Avg.**
						Zero-shot					
mBERT			Ori.	11.53	10.14	10.18	11.06	09.62	08.50	09.80	10.12
			Trans.	11.53	11.44	07.86	11.61	10.61	09.46	07.82	10.05
BERT			Trans.	12.83	12.48	10.74	<u>14.78</u>	13.41	<u>14.64</u>	10.84	<u>12.82</u>
BM25			Trans.	12.24	12.20	07.98	12.19	08.66	06.83	10.32	10.06
mDPR			Ori.	<u>16.07</u>	10.69	09.02	14.62	16.95	06.93	<u>12.46</u>	12.39
			Trans.	<u>16.07</u>	<u>14.03</u>	10.72	14.67	**18.57**	11.29	09.45	13.54
DPR			Trans.	**17.35**	**14.45**	**14.14**	**15.26**	<u>17.53</u>	**15.96**	**13.59**	**15.47**
mLegal			Ori.	11.51	08.82	<u>12.29</u>	10.24	11.23	09.64	11.83	10.79
			Trans.	11.51	09.38	10.49	09.43	12.90	09.94	09.33	10.43
						Fine-tuning					
mBERT	Ori.	Siam.	Ori.	<u>23.73</u>	23.49	22.73	20.09	18.76	17.11	17.75	20.52
			Trans.	<u>23.73</u>	22.60	22.48	24.31	27.04	23.36	20.15	23.38
	Trans.		Ori.	21.70	<u>24.63</u>	24.55	23.26	22.51	<u>25.08</u>	23.43	23.59
			Trans.	21.70	24.11	21.58	**26.12**	28.56	23.80	25.94	24.54
	Ori.	Two-tow	Ori.	16.73	17.79	18.99	22.19	20.45	18.62	20.56	19.33
			Trans.	16.73	17.28	21.04	20.61	20.62	18.48	21.12	19.41
	Trans.		Ori.	19.06	18.64	19.21	21.28	22.38	19.84	20.60	20.14
			Trans.	19.06	20.01	18.96	22.24	23.99	20.76	21.65	20.95
BERT	Trans.	Siam.	Trans.	**26.00**	23.38	23.82	<u>25.69</u>	27.04	**25.60**	25.66	**25.31**
	Trans.	Two-tow	Trans.	22.20	21.97	23.67	25.57	**32.55**	22.59	<u>26.15</u>	<u>24.96</u>
mDPR	Ori.	Siam.	Ori.	23.15	21.69	**26.40**	23.65	25.55	23.14	23.62	23.89
			Trans.	23.15	**24.73**	<u>26.03</u>	22.58	27.94	24.88	21.18	24.36
	Trans.		Ori.	22.05	20.33	22.84	25.13	17.51	23.95	21.63	21.92
			Trans.	22.05	22.71	21.70	24.53	<u>28.89</u>	22.78	24.82	23.93
	Ori.	Two-tow	Ori.	20.12	20.79	19.02	24.32	26.04	22.15	21.75	22.03
			Trans.	20.12	21.03	20.23	23.93	27.77	21.81	23.43	22.62
	Trans.		Ori.	19.96	21.44	20.71	22.28	22.30	19.55	18.82	20.72
			Trans.	19.96	18.83	21.84	19.65	23.78	19.58	20.87	20.64
DPR	Trans.	Two-tow	Trans.	20.51	22.63	23.20	21.67	25.83	20.57	**29.75**	23.45
mLegal	Ori.	Siam.	Ori.	15.26	16.86	18.66	19.78	21.91	15.99	18.88	18.19
			Trans.	15.26	15.91	14.93	13.25	16.87	17.33	15.72	15.61
	Trans.		Ori.	17.66	17.91	19.61	20.65	20.78	2.37	20.52	17.07
			Trans.	17.66	19.62	19.36	18.26	17.89	20.99	17.66	18.78
	Ori.	Two-tow	Ori.	18.76	19.21	19.85	19.58	18.06	21.39	18.23	19.30
			Trans.	18.76	18.93	20.85	20.67	17.94	20.79	18.05	19.43
	Trans.		Ori.	16.13	17.86	21.58	18.52	19.80	18.38	19.99	18.89
			Trans.	16.13	18.99	24.00	19.58	26.53	19.43	24.02	21.24

Table 5: Recall@2% performance on seen legal queries test split.

Model	Train Data	Model Config	Test Data	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.	Avg.
Zero-shot											
mBERT			Ori.	19.91	17.48	21.62	18.79	16.94	16.46	15.17	17.72
			Trans.	19.91	19.16	19.26	21.24	19.58	21.54	14.96	19.72
BERT			Trans.	23.72	21.65	23.06	28.03	25.18	25.16	**26.41**	24.74
BM25			Trans.	25.04	23.17	19.64	25.32	21.12	17.44	17.29	21.29
mDPR			Ori.	**29.83**	22.45	20.68	22.75	**29.79**	14.84	17.57	22.56
			Trans.	**29.83**	**27.55**	**24.70**	27.71	25.22	23.82	18.45	25.33
DPR			Trans.	28.85	25.27	23.78	**28.55**	29.47	**26.57**	22.28	**26.40**
mLegal			Ori.	18.10	17.16	17.76	21.77	18.39	20.35	21.55	19.30
			Trans.	18.10	17.98	18.48	21.28	22.15	19.35	17.01	19.19
Fine-tuning											
mBERT	Ori.	Siam.	Ori.	**44.02**	41.67	40.17	41.17	36.79	34.88	38.44	39.59
			Trans.	**44.02**	42.74	41.27	**48.08**	50.14	42.37	41.14	44.25
	Trans.		Ori.	40.74	41.53	42.83	39.19	41.64	43.68	41.10	41.53
			Trans.	40.74	44.42	41.95	43.41	53.20	44.39	46.33	44.92
	Ori.	Two-tow	Ori.	37.59	39.78	40.03	44.99	51.16	39.23	42.06	42.12
			Trans.	37.59	40.31	39.58	44.75	49.86	39.86	40.77	41.82
	Trans.		Ori.	38.22	39.37	40.18	40.48	50.56	39.67	39.78	41.18
			Trans.	38.22	40.27	37.26	40.75	50.65	41.03	40.47	41.24
BERT	Trans.	Siam.	Trans.	40.83	44.25	41.17	44.85	**55.76**	44.11	44.45	45.06
	Trans.	Two-tow	Trans.	40.46	43.73	**45.30**	45.09	49.26	44.07	44.86	44.68
mDPR	Ori.	Siam.	Ori.	42.15	**45.08**	44.95	45.19	48.93	44.16	**48.91**	**45.62**
			Trans.	42.15	44.74	43.62	41.63	50.93	**47.48**	43.67	44.89
	Trans.		Ori.	40.63	39.20	42.54	42.64	29.58	42.95	36.45	39.14
			Trans.	40.63	41.06	42.19	45.54	48.59	41.62	43.79	43.35
	Ori.	Two-tow	Ori.	40.08	42.89	43.18	46.29	53.37	43.26	41.48	44.36
			Trans.	40.08	42.92	43.74	46.15	55.11	42.89	41.40	44.61
	Trans.		Ori.	40.56	40.15	43.86	41.46	43.15	39.99	43.07	41.75
			Trans.	40.56	40.20	43.56	40.97	44.92	38.64	41.45	41.47
DPR	Trans.	Two-tow	Trans.	41.41	43.53	43.99	42.03	51.96	41.28	47.68	44.55
mLegal	Ori.	Siam.	Ori.	32.65	37.63	40.16	37.94	**45.62**	36.55	38.75	38.47
			Trans.	32.65	34.39	31.83	26.82	31.81	34.77	34.80	32.44
	Trans.		Ori.	37.18	36.10	39.46	40.64	37.76	05.91	34.49	33.08
			Trans.	37.18	36.33	39.38	38.36	38.11	36.52	35.09	37.28
	Ori.	Two-tow	Ori.	39.16	40.49	43.50	41.60	49.86	39.24	41.77	42.23
			Trans.	39.16	40.13	43.08	42.00	47.66	40.07	42.05	42.02
	Trans.		Ori.	37.11	35.95	40.35	38.23	44.87	35.97	39.68	38.88
			Trans.	37.11	37.28	43.46	39.56	44.16	39.88	41.71	40.45

Table 6: Recall@5% performance on seen legal queries test split.

Model	Train Data	Model Config	Test Data	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.	Avg.
Zero-shot											
mBERT			Ori.	31.55	29.20	30.27	32.47	30.45	27.84	34.22	30.86
			Trans.	31.55	31.06	34.70	32.74	32.60	33.52	28.70	32.12
BERT			Trans.	35.88	32.89	35.18	39.33	36.96	38.89	**41.84**	37.28
BM25			Trans.	38.89	36.98	34.74	38.47	38.75	31.64	29.43	35.56
mDPR			Ori.	**41.81**	36.77	32.09	36.21	37.69	25.61	27.76	33.99
			Trans.	**41.81**	37.47	**37.14**	41.88	39.61	38.61	30.87	38.20
DPR			Trans.	40.29	**38.27**	35.90	**44.95**	**43.47**	**39.77**	34.92	**39.65**
mLegal			Ori.	30.62	29.12	31.35	31.79	30.25	35.03	32.29	31.49
			Trans.	30.62	30.27	30.99	32.38	30.97	31.19	31.90	31.19
Fine-tuning											
mBERT	Ori.	Siam.	Ori.	63.45	65.12	65.67	62.91	58.03	59.99	63.07	62.61
			Trans.	63.45	68.22	67.24	69.30	75.10	68.37	**69.29**	68.71
	Trans.		Ori.	**64.80**	67.69	66.49	63.20	63.16	67.34	61.67	64.91
			Trans.	**64.80**	66.76	66.91	66.71	77.55	66.94	69.08	68.39
	Ori.	Two-tow	Ori.	59.86	66.40	66.55	65.95	70.73	63.40	62.74	65.09
			Trans.	59.86	65.98	67.32	65.93	72.61	62.81	61.81	65.19
	Trans.		Ori.	59.08	64.64	64.32	64.73	71.17	63.25	63.12	64.33
			Trans.	59.08	65.76	63.93	64.89	72.08	64.47	63.89	64.87
BERT	Trans.	Siam.	Trans.	60.81	65.41	59.64	62.94	68.76	62.99	65.79	63.76
	Trans.	Two-tow	Trans.	63.72	**68.52**	67.65	**70.03**	77.38	66.62	68.93	**68.98**
mDPR	Ori.	Siam.	Ori.	64.02	68.43	67.60	67.87	75.98	67.77	66.46	68.30
			Trans.	64.02	67.56	65.81	59.60	76.42	**69.69**	63.31	66.63
	Trans.		Ori.	63.47	63.97	66.84	63.29	41.84	69.11	57.75	60.90
			Trans.	63.47	63.53	65.16	65.82	72.69	66.48	64.85	66.00
	Ori.	Two-tow	Ori.	62.70	65.31	67.15	68.18	75.78	65.52	66.18	67.26
			Trans.	62.70	65.73	69.04	67.72	76.07	65.41	65.65	67.47
	Trans.		Ori.	63.47	66.78	**70.79**	66.54	73.20	66.34	66.29	67.63
			Trans.	63.47	66.53	70.41	66.41	73.61	66.19	64.96	67.37
DPR	Trans.	Two-tow	Trans.	62.91	66.16	69.25	63.59	**78.98**	69.17	68.63	68.38
mLegal	Ori.	Siam.	Ori.	56.09	62.74	64.56	63.80	73.35	65.69	63.37	64.23
			Trans.	56.09	60.17	54.93	46.10	60.38	60.02	59.64	56.76
	Trans.		Ori.	59.93	61.37	59.32	63.45	68.95	12.84	54.54	54.34
			Trans.	59.93	61.97	60.47	61.43	68.98	59.85	62.96	62.23
	Ori.	Two-tow	Ori.	64.25	66.81	69.04	65.30	70.38	63.15	62.51	65.92
			Trans.	64.25	66.38	69.15	65.26	70.38	62.97	61.41	65.69
	Trans.		Ori.	62.39	63.24	66.32	64.62	74.59	62.79	66.10	65.72
			Trans.	62.39	62.71	66.55	64.23	72.00	66.41	65.28	65.65

Table 7: Recall@10% performance on seen legal queries test split.

Model	Train Data	Model Config	Test Data	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.	Avg.
Zero-shot											
mBERT			Ori.	08.08	06.72	06.72	07.12	11.66	06.46	09.14	07.99
			Trans.	08.08	08.35	10.44	09.69	14.40	13.15	11.90	10.86
BERT			Trans.	11.84	10.76	12.13	12.92	10.74	14.56	10.52	11.92
BM25			Trans.	**14.41**	10.03	07.21	12.76	16.52	15.88	11.03	12.55
mDPR			Ori.	13.91	09.86	06.63	09.99	15.59	07.33	08.51	10.26
			Trans.	13.91	**14.00**	**12.51**	**13.02**	18.18	**16.85**	08.62	**13.87**
DPR			Trans.	10.69	10.81	11.96	10.67	**20.18**	12.22	14.89	13.06
mLegal			Ori.	09.65	07.86	07.75	11.57	12.55	10.40	**17.47**	11.04
			Trans.	09.65	08.80	08.48	10.53	08.90	11.27	07.93	09.37
Fine-tuning											
mBERT	Ori.	Siam.	Ori.	13.70	13.44	15.81	14.89	18.73	14.99	18.10	15.67
			Trans.	13.70	13.28	15.51	15.97	17.83	16.69	19.20	16.03
	Trans.		Ori.	15.85	13.57	13.63	14.42	12.42	14.33	15.52	14.25
			Trans.	15.85	16.66	15.76	**18.82**	14.91	14.11	21.21	16.76
	Ori.	Two-tow	Ori.	10.28	10.40	13.35	10.74	14.36	11.14	16.44	12.39
			Trans.	10.28	10.50	12.77	10.12	11.08	10.76	13.33	11.26
	Trans.		Ori.	13.69	13.84	14.62	14.55	12.48	14.68	16.09	14.28
			Trans.	13.69	14.06	15.09	13.59	12.51	14.35	15.40	14.10
BERT	Trans.	Siam.	Trans.	15.64	14.50	16.97	17.82	18.55	17.77	16.15	16.77
	Trans.	Two-tow	Trans.	12.44	13.05	14.44	14.01	15.07	13.81	17.93	14.39
mDPR	Ori.	Siam.	Ori.	12.37	13.47	17.27	13.80	19.98	15.37	**22.82**	16.44
			Trans.	12.37	15.07	14.88	14.49	**20.32**	16.72	20.17	16.29
	Trans.		Ori.	**15.92**	16.77	16.50	15.13	15.63	15.20	16.55	15.96
			Trans.	**15.92**	**17.99**	17.36	14.71	19.09	**19.59**	17.64	**17.47**
	Ori.	Two-tow	Ori.	13.88	12.10	11.75	13.02	12.77	15.80	08.51	12.55
			Trans.	13.88	12.01	13.01	14.20	13.66	16.79	11.26	13.54
	Trans.		Ori.	13.22	13.72	14.15	12.86	15.73	14.37	16.55	14.37
			Trans.	13.22	12.07	14.24	14.45	14.78	14.84	19.89	14.78
DPR	Trans.	Two-tow	Trans.	13.74	12.96	14.44	14.80	16.35	12.85	15.69	14.40
mLegal	Ori.	Siam.	Ori.	09.68	14.30	13.47	16.18	13.59	09.71	16.26	13.31
			Trans.	09.68	11.11	08.26	10.48	06.72	09.24	09.37	09.27
	Trans.		Ori.	13.06	13.15	13.13	13.73	14.45	04.58	15.63	12.53
			Trans.	13.06	13.94	14.03	15.70	10.50	13.13	13.45	13.40
	Ori.	Two-tow	Ori.	13.83	15.48	**19.29**	15.07	14.99	17.02	17.53	16.17
			Trans.	13.83	15.60	19.23	16.07	14.46	16.93	17.53	16.24
	Trans.		Ori.	09.22	09.76	12.38	08.68	15.15	10.95	12.30	11.21
			Trans.	09.22	09.87	11.63	10.19	15.93	10.73	11.15	11.25

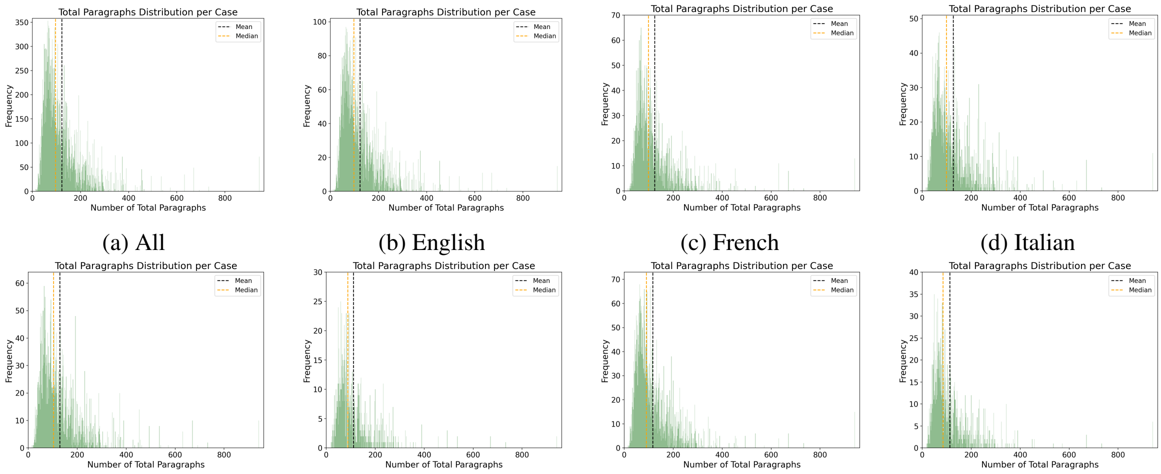
Table 8: Recall@2% performance on unseen legal queries test split.

Model	Train Data	Model Config	Test Data	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.	Avg.
Zero-shot											
mBERT			Ori.	16.76	15.72	14.33	14.49	21.65	12.65	15.57	15.88
			Trans.	16.76	20.02	20.11	17.42	24.49	20.86	<u>23.85</u>	20.50
BERT			Trans.	19.87	18.78	<u>21.21</u>	<u>21.60</u>	22.70	<u>25.98</u>	17.76	21.13
BM25			Trans.	**25.09**	<u>23.12</u>	20.33	**27.18**	**36.08**	**26.44**	**27.13**	**26.48**
mDPR			Ori.	<u>23.59</u>	17.62	15.07	15.68	26.25	12.53	15.06	17.97
			Trans.	<u>23.59</u>	**25.23**	**22.29**	20.72	29.91	25.74	19.37	<u>23.84</u>
DPR			Trans.	20.93	19.18	<u>21.78</u>	20.27	<u>33.11</u>	22.57	20.40	22.61
mLegal			Ori.	16.35	15.19	16.18	17.47	21.86	20.06	22.30	18.49
			Trans.	16.35	16.20	16.85	17.07	19.31	19.70	19.31	17.83
Fine-tuning											
mBERT	Ori.	Siam.	Ori.	27.68	27.89	31.07	29.61	34.78	29.64	35.63	30.90
			Trans.	27.68	28.75	33.16	30.41	36.02	33.39	37.53	32.42
	Trans.		Ori.	29.53	27.40	29.61	32.29	29.31	31.86	39.25	31.32
			Trans.	29.53	28.36	28.25	<u>33.70</u>	29.37	30.64	**41.26**	31.59
	Ori.	Two-tow	Ori.	22.34	22.85	25.86	23.17	26.15	25.12	31.26	25.25
			Trans.	22.34	22.83	26.12	24.22	27.43	24.67	31.95	25.65
	Trans.		Ori.	27.09	28.36	32.12	28.42	33.07	29.38	35.11	30.51
			Trans.	27.09	28.96	32.15	29.20	32.67	29.54	35.29	30.70
BERT	Trans.	Siam.	Trans.	<u>31.01</u>	30.14	<u>34.61</u>	**34.86**	**38.77**	32.07	37.07	<u>34.08</u>
	Trans.	Two-tow	Trans.	30.49	32.44	**34.98**	32.78	33.90	32.30	<u>40.63</u>	33.93
mDPR	Ori.	Siam.	Ori.	25.83	28.04	29.79	27.46	33.81	30.08	38.05	30.44
			Trans.	25.83	31.48	30.17	26.20	<u>38.64</u>	30.29	33.16	30.82
	Trans.		Ori.	**31.28**	**33.05**	33.81	30.68	32.84	<u>34.23</u>	31.55	32.49
			Trans.	**31.28**	<u>32.96</u>	34.53	33.04	38.05	**36.66**	34.31	**34.40**
	Ori.	Two-tow	Ori.	28.34	28.40	30.99	27.97	31.39	33.50	32.99	30.51
			Trans.	28.34	29.22	30.35	29.44	31.69	33.56	34.71	31.04
	Trans.		Ori.	29.49	28.68	29.68	29.81	31.85	31.06	35.17	30.82
			Trans.	29.49	29.03	31.75	31.48	33.33	30.02	36.21	31.62
DPR	Trans.	Two-tow	Trans.	28.15	29.33	31.25	30.38	34.92	30.68	34.54	31.32
mLegal	Ori.	Siam.	Ori.	22.51	27.79	30.27	28.35	30.00	22.54	36.38	28.26
			Trans.	22.51	23.39	20.34	22.42	18.11	22.42	20.98	21.45
	Trans.		Ori.	23.73	22.34	29.84	25.34	29.03	11.90	30.98	24.74
			Trans.	23.73	25.29	23.95	26.31	24.27	28.67	25.00	25.32
	Ori.	Two-tow	Ori.	28.64	28.92	31.72	31.05	29.59	33.43	32.41	30.82
			Trans.	28.64	29.40	31.95	31.50	29.21	31.67	32.41	30.68
	Trans.		Ori.	24.71	26.27	27.59	24.93	32.65	24.77	23.22	26.31
			Trans.	24.71	25.81	28.78	24.94	31.92	26.43	31.09	27.67

Table 9: Recall@5% performance on unseen legal queries test split.

Model	Train Data	Model Config	Test Data	Eng.	Fre.	Ita.	Romn.	Rus.	Turk.	Ukr.	Avg.
Zero-shot											
mBERT			Ori.	27.93	25.88	25.43	26.91	31.41	21.77	26.55	26.55
			Trans.	27.93	30.47	34.25	29.75	33.11	32.2	36.26	32.00
BERT			Trans.	31.98	32.07	35.06	33.5	35.93	**42.09**	32.82	34.78
BM25			Trans.	35.15	34.88	31.31	**38.2**	**48.52**	39.67	**45.29**	**39.00**
mDPR			Ori.	**35.84**	30.59	23.47	28.55	39.42	19.19	29.54	29.51
			Trans.	**35.84**	**37.78**	**36.1**	34.91	45.17	38.22	34.94	37.57
DPR			Trans.	30.76	32.31	35.1	31.15	42.99	34.96	33.85	34.45
mLegal			Ori.	24.37	24.56	26.89	26.82	30.8	30.61	34.6	28.38
			Trans.	24.37	24.98	27.62	26.29	27.41	30.09	29.37	27.16
Fine-tuning											
mBERT	Ori.	Siam.	Ori.	48.03	47.15	47.92	50.21	53.83	46.98	51.32	49.35
			Trans.	48.03	48.05	48.08	48.22	53.7	52.6	55.8	50.64
	Trans.		Ori.	47.72	46.98	49.7	51.71	49.09	52.1	56.72	50.57
			Trans.	47.72	48.5	48.58	**54.24**	50.98	51.27	**59.37**	51.52
	Ori.	Two-tow	Ori.	41.42	44.99	46.83	45.54	48.68	46.92	54.94	47.05
			Trans.	41.42	43.2	44.77	43.9	47.2	45.54	50.86	45.27
	Trans.		Ori.	44.85	47.02	46.84	45.4	48.27	46.88	52.59	47.41
			Trans.	44.85	48.19	47.61	47.41	48.84	45.94	52.24	47.87
BERT	Trans.	Siam.	Trans.	46.88	49.95	49.62	52.16	53.83	48.62	51.55	50.37
	Trans.	Two-tow	Trans.	45.66	46.41	48.69	49.63	48.87	50.32	52.76	48.91
mDPR	Ori.	Siam.	Ori.	43.95	47.73	46.27	48.33	51.42	49.18	56.95	49.12
			Trans.	43.95	45.34	45.58	42.09	52.85	47.95	53.28	47.29
	Trans.		Ori.	**48.94**	51.99	**55.41**	51.88	**56.14**	53.91	52.3	52.94
			Trans.	**48.94**	**53.04**	53.01	52.56	56.03	**55.36**	57.59	**53.79**
	Ori.	Two-tow	Ori.	45.19	49.81	53.08	49.09	51.65	51.27	55.98	50.87
			Trans.	45.19	48.01	50.44	48.08	50.44	52.42	54.25	49.83
	Trans.		Ori.	45.59	47.46	47.74	48.03	49.33	48.71	53.51	48.62
			Trans.	45.59	46.81	48.63	47.93	51.77	48.42	54.2	49.05
DPR	Trans.	Two-tow	Trans.	47.52	50.07	49.19	52.55	50.91	50.56	53.74	50.65
mLegal	Ori.	Siam.	Ori.	43.2	48.23	51.69	49.47	45.59	44.46	53.62	48.04
			Trans.	43.2	42.41	43.1	39.87	39.37	39.83	42.3	41.44
	Trans.		Ori.	41.64	37.42	46.07	41.37	53.79	22.45	53.45	42.31
			Trans.	41.64	44.8	44.1	45.92	43.25	49.92	45.34	45.00
	Ori.	Two-tow	Ori.	46.28	48.31	51.87	49.42	50.96	51.16	54.6	50.37
			Trans.	46.28	48.53	51.87	49.42	50.36	50.38	54.25	50.16
	Trans.		Ori.	45.08	48.18	51.89	46.19	55.01	47.99	49.2	49.08
			Trans.	45.08	47.95	52.47	47.5	55.14	48.43	53.56	50.02

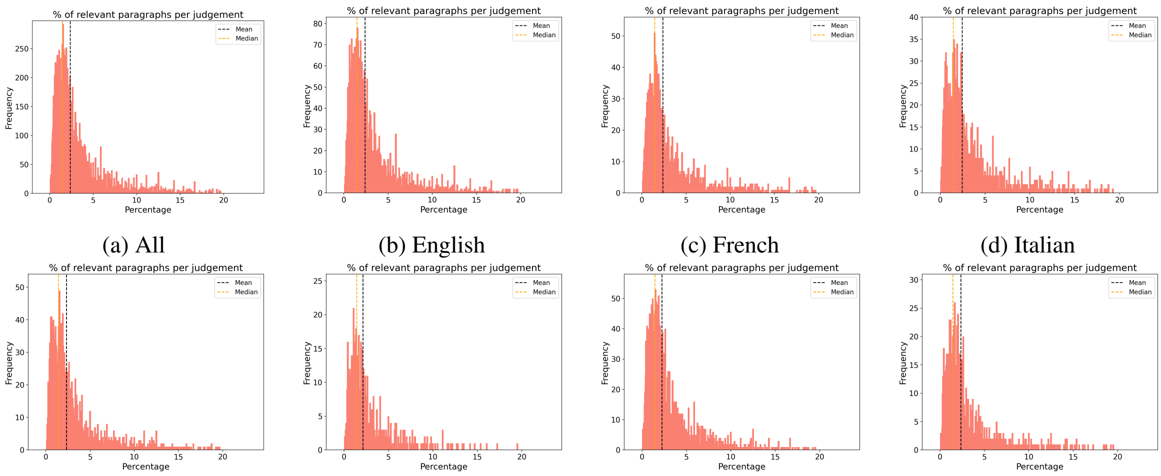
Table 10: Recall@10% performance on unseen legal queries test split.



(a) All (b) English (c) French (d) Italian

(e) Romanian (f) Russian (g) Turkish (h) Ukrainian

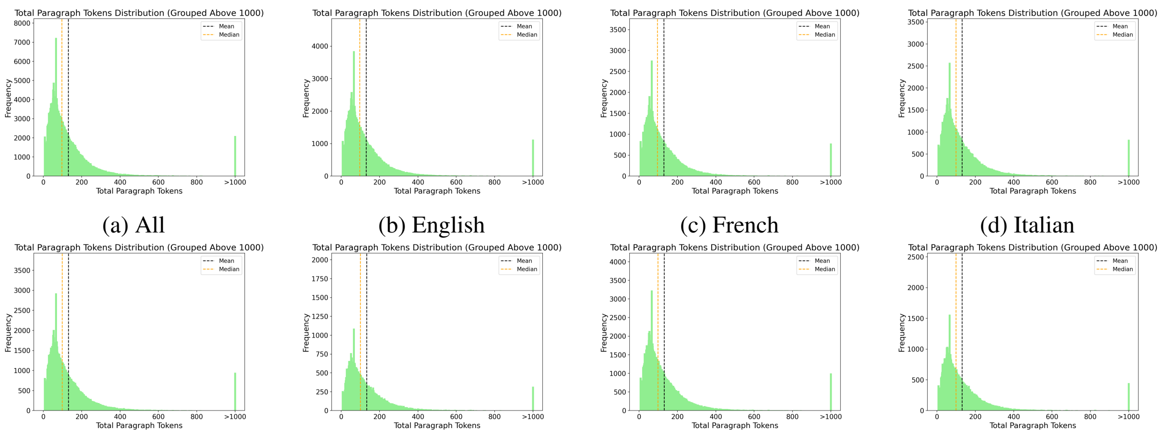
Figure 3: LexCLiPR Data Analysis: Total Number of Paragraphs per Judgement



(a) All (b) English (c) French (d) Italian

(e) Romanian (f) Russian (g) Turkish (h) Ukrainian

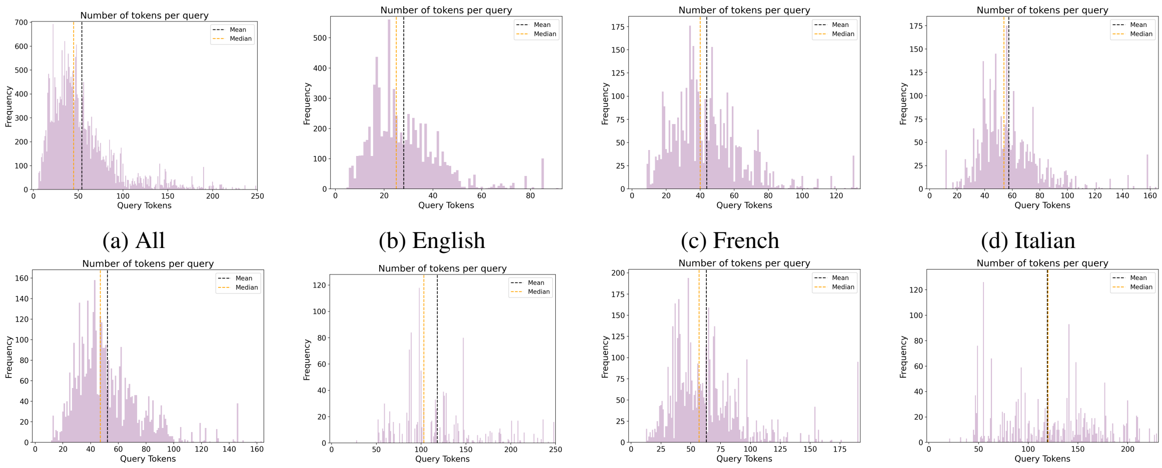
Figure 4: LexCLiPR Data Analysis: Percentage of Relevant Paragraphs per judgement.



(a) All (b) English (c) French (d) Italian

(e) Romanian (f) Russian (g) Turkish (h) Ukrainian

Figure 5: LexCLiPR Data Analysis: Number of Tokens per Paragraph



(a) All (b) English (c) French (d) Italian

(e) Romanian (f) Russian (g) Turkish (h) Ukrainian

Figure 6: LexCLiPR Data Analysis: Number of Tokens per Query