Query-driven Relevant Paragraph Extraction from Legal Judgments

Santosh T.Y.S.S, Elvin Quero Hernandez, Matthias Grabmair
School of Computation, Information, and Technology;
Technical University of Munich, Germany
{santosh.tokala, elvin.quero, matthias.grabmair}@tum.de

Abstract

Legal professionals often grapple with navigating lengthy legal judgements to pinpoint information that directly address their queries. This paper focus on this task of extracting relevant paragraphs from legal judgements based on the query. We construct a specialized dataset for this task from the European Court of Human Rights (ECtHR) using the case law guides. We assess the performance of current retrieval models in a zero-shot way and also establish fine-tuning benchmarks using various models. The results highlight the significant gap between fine-tuned and zero-shot performance, emphasizing the challenge of handling distribution shift in the legal domain. We notice that the legal pre-training handles distribution shift on the corpus side but still struggles on query side distribution shift, with unseen legal queries. We also explore various Parameter Efficient Fine-Tuning (PEFT) methods to evaluate their practicality within the context of information retrieval, shedding light on the effectiveness of different PEFT methods across diverse configurations with pre-training and model architectures influencing the choice of PEFT method.

Keywords: Relevant Paragraph Identification, Parameter Efficient Retrieval, Legal Retrieval

1. Introduction

Legal professionals including lawyers, judges and paralegals, often need to sift through voluminous legal judgments that encompass crucial insights for case law interpretations and judicial reasoning. These judgments, often lengthy, contain nuanced paragraphs holding the key to understanding legal principles, precedents and arguments. Finding relevant case law accounts for roughly 15 hours per week for a lawyer (Lastres, 2015) or nearly 30% of their annual working hours (Poje, 2014). Recent advances in NLP offer new possibilities to bridge this gap by providing summaries of these documents (e.g., Bhattacharya et al. 2019; Shukla et al. 2022 *inter alia*). Nonetheless, practitioners still face challenges in navigating these texts to uncover specific paragraphs that address their queries. The current manual approach is labor-intensive and susceptible to overlooking essential details. Automating this process of identifying paragraphs relevant to the query streamlines legal research, allowing them to access relevant information efficiently.

Finding relevant paragraphs to a query is a challenging task unlike traditional adhoc information retrieval. Firstly, the legal domain is characterized by a vast and intricate vocabulary, interwoven with domain-specific jargon that can vary across different legal jurisdictions. This linguistic complexity demands an in-depth understanding of nuanced legal concepts, posing a substantial challenge for automated systems. The variation in legal writing style further compounds the challenge. Judgments may employ different degrees of formalism and offer varying levels of explicitness. These nuances can lead to difficulties in discerning context and accurately identifying relevant paragraphs that address specific queries. Another key challenge stems from the evolving nature of the legal case law. New legal doctrines, precedents and interpretations continually emerge, leading to an ever evolving array of legal concepts and principles. This dynamism necessitates a flexible and adaptive approach to comprehend new queries and determine relevance.

To investigate the ability of current retrieval models to identify relevant paragraphs, a high-quality labeled dataset is imperative. However, creating such datasets is resource-intensive, often necessitating the involvement of legal experts to produce queries and relevance labels. In this study, we employ distant supervision to construct a dataset tailored for the task of query-driven relevant paragraph extraction from legal judgments by the European Court of Human Rights (ECtHR) which addresses grievances by individuals against states for alleged violations of rights outlined in the European Convention of Human Rights. Our approach capitalizes on the case-law guides available through the ECtHR's Knowledge Sharing platform[1]. We pose the case-law guide's section headers as queries, mirroring the legal concepts professionals utilize when searching within ECtHR judgments. We gather relevance signals by identifying the pinpointed citations to the paragraphs in the judgments within these guides under each section. Further, we meticulously design various splits to assess the generalizability of systems towards new queries (legal concepts), adapting to the evolution of law.[2]

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

[2] Our dataset is made available at https://github.com/TUMLegalTech/ParagraphRetrievalECHR/

LREC-COLING 2024, pages 13442–13454
20-25 May, 2024.

As a second contribution, we assess the performance of current retrieval models in a zero-shot manner using our dataset and further establish fine-tuning benchmarks employing diverse retrieval techniques encompassing dense bi-encoder and cross-encoder architectures. Our experiments reveal the drastic gap between fine-tuned and the zero-shot performance. Furthermore, we investigate into the efficacy of fine-tuning a general pre-trained model that was fine-tuned using other retrieval datasets (such as BERT fine-tuned on MSMARCO), comparing it against a legally pre-trained model (such as LegalBERT) that remains untouched by other retrieval datasets except ours. This investigation revealed that legal pre-training helps to handle distribution shift of the corpus, but still lacks in handling the distribution shift towards unseen queries.

While complete fine-tuning has shown better performance, the trend towards larger models with billions or trillions of trainable parameters makes this fine-tuning process resource-intensive and costly. This spurred the exploration of Parameter Efficient Fine-Tuning (PEFT) strategies which update only a small number of extra parameters while keeping the original pre-trained model parameters frozen. In our study, we delve into this emerging area by evaluating representative methods of PEFT, namely Adapter (Houlsby et al., 2019), prefix-tuning (Li and Liang, 2021) and LoRA (Hu et al., 2021), within the context of our paragraph retrieval dataset. This investigation contributes to the ongoing discourse regarding the practicality of adopting PEFT in the realm of Information Retrieval (Pal et al., 2023; Tam et al., 2022; Ma et al., 2022; Jung et al., 2022). Our experiments demonstrate that PEFT methods achieve comparable performance to full fine-tuning on both seen and unseen queries, with the choice of the best PEFT method contingent on configuration such as general vs. legal pre-training and bi- vs. cross-encoder settings.

2. Related Work

Legal IR Retrieving essential legal information is integral to the workflow of lawyers, encompassing tasks such as searching for legislation (adhoc search or by providing a factual description to identify the relevant statutes (Wang et al., 2018; Paul et al., 2022)), similar prior cases (Rabelo et al., 2022; Mandal et al., 2017), civil codes (Kim et al., 2016, 2014), litigation documents such as technology-assisted-review (Cormack et al., 2010), patents (Piroi et al., 2013) and within law firm's internal support system (Moens, 2001). Our work focuses specifically on legal case retrieval. Most of the existing legal case law retrieval works primarily aim to retrieve entire cases (Sansone and Sperlí, 2022) based on different query granularities, including whole cases (Rabelo et al., 2022; Ma et al., 2021; Mandal et al., 2017) or specific legal queries (Locke et al., 2017; Locke and Zuccon, 2018; Koniaris et al., 2016). In contrast, our approach involves retrieving relevant paragraphs at a finer granularity, providing practitioners with a more targeted means of identifying essential information. At the paragraph granularity level, the legal case entailment task in COLIEE involves identifying a paragraph from existing cases that matches the decision of a new case (Rabelo et al., 2022), but it employs the entire case as the query, in contrast to the short queries used in our work. This paragraph-level retrieval functionality is integral to building legal Question Answering (Khazaeli et al., 2021; Verma et al., 2020) and Query-focused summarization systems.

Tasks on ECtHR Corpora Previous works involving ECtHR corpus has dealt with judgement prediction (Aletras et al., 2016; Chalkidis et al., 2019, 2021; Santosh et al., 2022, 2023; Tyss et al., 2023; Xu et al., 2023b), argument mining (Mochales and Moens, 2008; Habernal et al., 2023; Poudyal et al., 2019, 2020), vulnerability detection (Xu et al., 2023a), event extraction (Filtz et al., 2020; Navas-Loro and Rodriguez-Doncel, 2022). In this work, we capitalize on the case law guides maintained by registry of ECtHR to derive a query-driven relevant paragraph extraction dataset. We offer this dataset to the research community to facilitate advancements in area of AI-enabled tools for legal practitioners.

Parameter Efficient Retrieval With sizes of pre-trained language models soaring up (Brown et al., 2020), full-parameter fine-tuning has become more challenging, this has created an interest in PEFT methods such as prompt tuning (Li and Liang, 2021; Lester et al., 2021; Liu et al., 2022), adapters (Houlsby et al., 2019; Pfeiffer et al., 2021; Mahabadi et al., 2021), additive methods (Hu et al., 2021; Guo et al., 2021; Zhang et al., 2020) and hybrid methods (Mao et al., 2022; Chen et al., 2022). Specifically in IR, Ma et al. 2022 conducted a comprehensive study of several PEFT methods for both the retrieval and re-ranking stages. Jung et al. 2022 has explored prefix-tuning and LoRA on bi-encoder models. Tam et al. 2022 examined the effect of these methods on in-domain, cross-domain and cross-topic retrieval. Pal et al. 2023 studied the effect of adapters on sparse retrieval models contrary to dense models. We contribute to this ongoing discourse using both bi- and cross-encoders using our paragraph retrieval dataset on legal judgements.

3. Task & Dataset

Our task of query-driven 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 Creation

Judgements Collection We acquire ECtHR judgements collection as an HTML data dump from HUDOC [3], the publicly available database of the ECtHR, along with their associated metadata. We retain only the English documents based on their metadata (Document Type: 'HEJUD'). The parsing of judgment into paragraphs posed challenges due to inconsistent HTML structure, the presence of sub-paragraph numbers within each paragraph and the occurrence of spurious paragraph numbers resulting from verbatim text copied from other documents to cross-reference those paragraphs. To address these issues, we devised a range of hand-crafted heuristics to segment the judgment documents into paragraphs. Each paragraph is uniquely identified by its paragraph number at the beginning, facilitating cross-referencing.

Queries and Paragraph Relevance Collection We curate our query-paragraph relevance dataset using case-law guides accessible on ECtHR Knowledge Sharing Platform[4]. This platform, maintained by the court's registry, analyzes case law development for each convention article (e.g., Article 4 - Prohibition of slavery and forced labor[5]) and transversal themes (e.g., Data Protection[6], Rights of LGBTI persons[7]). It comprises 28 article and 8 theme-related case law guides, updated weekly, making them up-to-date with evolving case law with every new judgement and our proposed task can in turn assist registry in achieving this goal of updating these guides regularly.

Obtaining queries The case law guides provide the details of the key concepts involved under each article/theme and discuss them in detail by providing references to the relevant judgements. The legal concepts involved under each article/theme are structured in a hierarchical fashion, with sub-concepts enumerated. A representative index structure of a case law guide is illustrated in Figure 1. For instance, this is a hierarchical path of sections within the theme guide of Rights of LGBTI persons → Freedom of expression and association → Imposed silence and legal bans concerning homosexuality. We can extract this hierarchical structure by parsing the PDF case law guides' structural information. To construct each query, we combine these multiple concepts along the path (from the article or theme title to the leaf node in the PDF structure) by using a delimiter. This approach generates queries that mirror lists of legal concepts, akin to those sought after by legal practitioners when searching in ECtHR judgments. These queries/legal concepts could be used to index legal analytics databases that inform litigation strategies.

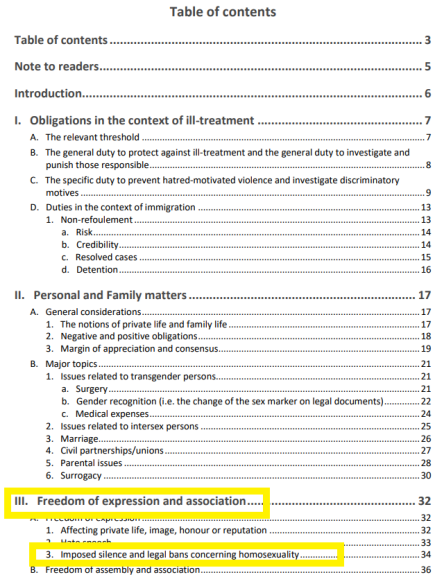
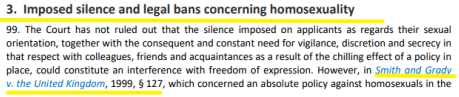


Table of contents

Table of contents 3
Note to readers 5
Introduction 6
I. Obligations in the context of ill-treatment 7
A. The relevant threshold 7
B. The general duty to protect against ill-treatment and the general duty to investigate and punish those responsible 8
C. The specific duty to prevent hatred-motivated violence and investigate discriminatory motives 9
D. Duties in the context of immigration 13
1. Non-refoulement 13
a. Risk 14
b. Credibility 14
c. Resolved cases 15
d. Detention 16
II. Personal and Family matters 17
A. General considerations 17
1. The notions of private life and family life 17
2. Negative and positive obligations 18
3. Margin of appreciation and consensus 19
B. Major topics 21
1. Issues related to transgender persons 21
a. Surgery 21
b. Gender recognition (i.e. the change of the sex marker on legal documents) 22
c. Medical expenses 24
2. Issues related to intersex persons 25
3. Marriage 26
4. Civil partnerships/unions 27
5. Parental issues 28
6. Surrogacy 30
III. Freedom of expression and association 32
1. Affecting private life, image, honour or reputation 32
3. Imposed silence and legal bans concerning homosexuality 34
B. Freedom of assembly and association 36

Figure 1: Query construction process from case law guide. The above table of contents is obtained from 'Rights of LGBTI persons' guide.


3. Imposed silence and legal bans concerning homosexuality

99. The Court has not ruled out that the silence imposed on applicants as regards their sexual orientation, together with the consequent and constant need for vigilance, discretion and secrecy in that respect with colleagues, friends and acquaintances as a result of the chilling effect of a policy in place, could constitute an interference with freedom of expression. However, in *Smith and Grady v. the United Kingdom*, 1999, § 127, which concerned an absolute policy against homosexuals in the

Figure 2: Illustration of pin-pointed paragraph relevance in case law guides.

Obtaining relevant paragraphs in Judgements These case-law guides provide in-depth discussions of each legal concept, offering pin-pointed paragraph references to the judgements from the ECtHR. An example of a legal concept description from a case-law guide is depicted in Fig. 2, demonstrating how relevant paragraphs are referenced under each query. We gather all paragraph references in a specific judgement under each legal con-

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

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

[5] https://ks.echr.coe.int/web/echr-ks/article-4

[6] https://ks.echr.coe.int/web/echr-ks/data-protection

[7] https://ks.echr.coe.int/web/echr-ks/rights-of-lgbti-persons

cept and mark all of them as relevant corresponding to the given query in that judgement. However, it's worth mentioning that all judgements are not exhaustively covered in the case-law guide unless they contribute to the expansion or contraction of existing case law. Taking this into account, we pair queries with specific judgements referenced within them, subsequently extracting relevant paragraphs from these judgements. This contrasts with using all the paragraphs from all the judgements as the candidate set for identifying relevance. While our proposed methodology could theoretically be applied to all judgements across the corpus, we opt to restrict each query to the judgements specifically referenced under it. This deliberate limitation aims to ensure a high-quality evaluation setup, controlling false negatives.

We filter out those query-judgement pairs in which reference to judgement is missing paragraph-level reference. Finally, we map back judgements in query-judgement pairs to our judgements collection, removing the ones which we could not map back as some may refer to non-English documents which have not considered in our collection.

3.2. Data Splits & Analysis

We eventually end up with 4109 query-judgement pairs with 708 unique queries. The number of paragraphs in Judgement range from 21 to 942 with a mean of 102.78 (Fig. 3a) . The percentage of relevant paragraphs in each query-judgement pair range from 0.10% to 15% to the total number of paragraphs in that judgement with a mean around 1.95%, depicted in Fig. 3b. The queries and paragraph have a mean length of 36 and 135 tokens, illustrated in Figures 3c and 3d respectively.

We partition the article/theme case law guides into two distinct splits: one exclusively designated for testing with 403 query-judgment pairs (111 unique queries) derived from these case law guides, referred to as 'Unseen article/themes'. This creates a rigorous unseen evaluation scenario, assessing the model's performance on unfamiliar legal concepts from themes and articles that were not encountered during training. Queries originating from the other split are further divided into two subsets, resulting in 'Seen article/theme, Unseen Query' with 694 pairs (120 unique queries) and 'Seen article/theme, Seen Query' with 3012 pairs (477 unique queries). The former, reserved for testing, exposes the model to previously encountered themes/articles, but with new queries. The latter group is further divided into training (2230 pairs), validation (302 pairs), and test (480 pairs) sets. The test set within the 'Seen article/theme, Seen Query' category assesses the model's comprehension of familiar legal concepts on new judgments in the test set.

4. Retrieval Models

We benchmark our task of identifying relevant paragraphs from a legal judgement given a query using the following models. We compute relevance score for each paragraph in given judgement with respect to the query and obtain the top-k most relevant paragraphs with the highest scores.

BM25 (Robertson et al., 1995) is a bag-of- words approach that estimates paragraph relevance to a query by considering the presence of query terms in the paragraph.

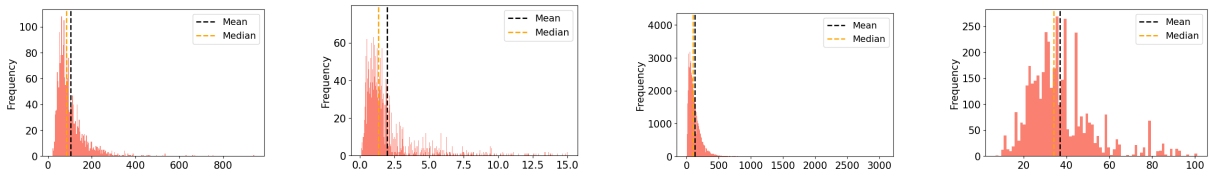
Bi-encoders employ separate encoders to encode queries and paragraphs into low-dimensional representations independently, leveraging neural architectures to capture semantic relationship and the final relevance score is computed using dot-product between the representations of query and paragraph obtained from encoder 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 is to learn representations such that relevant pairs of query and paragraphs will have higher similarity than the irrelevant ones. To reduce the training cost given there are lot of irrelevant paragraphs, negative sampling has been employed. Let $\{<q_i, p_i^+, p_{i,1}^-, \ldots p_{i,n}^->\}_{i=1}^m$ be the training data that consists 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. We optimize negative log likelihood loss function as:

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

Following Karpukhin et al. 2020, we consider negatives chosen from the irrelevant paragraphs randomly and the top paragraphs returned by BM25 which are not relevant to the query. We refer this approach as Dense Passage Retrieval **(DPR)**.

Recently, Xiong et al. 2020 proposed Approximate nearest neighbor Negative Contrastive Learning (**ANCE**) mechanism for dense retrieval. Instead of random or static BM25 negatives, ANCE constructs negatives using the being-optimized dense retrieval model. This helps to align the distribution of negative samples based on the models' training dynamics. While the model undergoes updates with each iteration, it would be expensive to update the negatives for every batch based on the updated model. Hence we asynchronously refresh the negatives at every checkpoint to reduce the computational cost to construct them.

These above methods follow a single-vector paradigm where each query and each paragraph



(a) Number of paragraphs per judgement.

(b) % of relevant paragraphs per judgement.

(c) Number of tokens per paragraphs.

(d) Number of tokens per query.

Figure 3: Data Analysis

is encoded into a single high-dimensional vector which is used to calculate relevance using a dot product. Khattab and Zaharia 2020 proposed a late interaction method named contextualized late interaction over BERT (**ColBERT**) where queries and documents are encoded at a finer granularity into multi-vector representations and relevance is estimated using interactions between these two sets of vectors. ColBERT produces an embedding for every token in the query and the paragraph and computes relevance as the sum of maximum similarities between each query vector and all vectors in the document as $rel(q,p) = \sum_{i=1}^{N} \max_{j=1}^{M} Q_i.D_j^\intercal$ where Q is a query encoding matrix corresponding to N token vectors and D denotes the paragraph encoding matrix corresponding to M token vectors.

Cross-encoders concatenate both of them before being provided to the model instead of encoding query and paragraph separately. The relevance score is directly computed by feed-forward network using the combined representation of the both (Yates et al., 2021) as $rel(q,p) = f(E_\phi(q,p))$ where E_ϕ represents a pre-trained model such as BERT and f denotes a feed-forward network which takes [CLS] representation as input to compute relevance score and is trained end-to-end with binary cross entropy loss. This allows for deeper interaction between the query and paragraph but this effectiveness comes with a cost on efficiency as it now involves whole pass through the model for each query paragraph pair, instead of being able to precompute all the paragraph representations and use the model once to obtain query representation to calculate the relevance score as in bi-encoders.

5. Zero-shot & Fine-tune Experiments

Initially, we investigate the performance of retrieval models in a zero-shot evaluation scenario, where models trained on the MS MARCO paragraph ranking dataset (Bajaj et al., 2016) - a large-scale adhoc retrieval dataset derived from the Bing search log containing 8.8 million passages and around 800K queries for training, are directly evaluated on our legal judgement paragraph ranking dataset. We examine the following models: (i) DPR[8] (ii) ANCE[9] (iii) ColBERT [10] (iv) Cross encoder[11]. We also evaluate a legal-domain-specific encoder model, LegalBERT (Chalkidis et al., 2020) which is pre-trained on diverse English legal texts encompassing legislative content, court cases, and contracts using cosine similarity between obtained [CLS] embeddings as relevance score. Notably, LegalBERT has been exposed to case law from ECtHR.

Subsequently, we fine-tune these models on the training split of our legal judgment paragraph extraction dataset. We create two variants of each model, with distinct initializations: (i) model already fine-tuned on MSMARCO (models used in the zero-shot evaluation) and (ii) LegalBERT.

Implementation Details For DPR, we use mix of negatives from BM25 and random in ratio of 4:1 and train with total of 5 negatives per query-positive pair. For ANCE, we use same number of negatives derived from model. While for COLBERT and cross encoders, we use seven negatives samples for every positive query, where 4 are sampled randomly and 3 are from BM25 negatives. We sweep over learning rates $\{1e-5, 3e-5, 5e-5, 1e-4, 3e-4\}$ and the model is trained end-to-end for 5 epochs with Adam optimizer (Kingma and Ba, 2014) and we select the best model based on the performance on the validation set.

Metrics We evaluate the performance using Recall@k% (R@K%). Recall@k% measures the proportion of relevant paragraphs in the top-k% of the total paragraphs in the judgement and we report mean across all instances. We report for $k = \{2, 5, 10\}$. We use the k as percentage instead of absolute value to account for varying number of

[8] https://huggingface.co/facebook/dpr-question_encoder-multiset-base

[9] https://huggingface.co/sentence-transformers/msmarco-roberta-base-ance-firstp

[10] https://github.com/stanford-futuredata/ColBERT

[11] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

paragraphs across different judgements. Higher recall scores indicate better performance.

5.1. Results

We report the results of both the zero-shot and the fine-tuning experiments in Table 1.

Zero-shot: We observe neural models demonstrate better performance across all the splits compared to BM25, bridging the lexical gap issue. ANCE displays slightly better performance than DPR demonstrating effectiveness of its dynamic negative sampling. COLBERT demonstrates superior performance across all variants, with a larger margin. This can be owed to its multi-vector representations at the granularity of each token and its training with distillation loss from re-ranker models. We notice cross encoder are comparable to other dense models except to COLBERT, due to its ability to act better in re-ranking stage rather than retrieval stage. The performance order of these models is consistent with the out-of-domain zero-shot results on BEIR leaderboard[12] (Thakur et al., 2021). Surprisingly, LegalBERT performs comparably similar or less than BM25 and significantly below retrieval fine-tuned general models, contrary to what one might expect that legal pre-training would mitigate distribution shift on the corpus side to capture relevance. This points out general masked language model objective can not effectively translate to capture relevance in retrieval settings and calls for investigation of pre-training objectives suitable for retrieval such as inverse cloze task (Lee et al., 2019), masking salient spans (Singh et al., 2021) to handle phrase level query matching and contrastive based pre-training (Izacard et al., 2022).

Zero-shot vs Fine-tune: All the fine-tuning models (both MSMARCO and LegalBERT initialized ones) substantially improve over zero-shot variants in all the three splits. This difference highlights the need for future research to improve the generalization ability of current IR models to domains without any relevance label by handling distribution shifts from both the query and corpus side.

Fine-tune: Despite COLBERT demonstrating a better zero-shot performance, cross encoders performed better with fine-tuning due to their deep interactions through concatenations, but that comes at a cost of efficiency to compute joint representation. Among bi-encoders, COLBERT perform well compared to ANCE followed by DPR due to its late interaction using multiple vector representations. The difference between them gets closer with fine-tuning on the 'Seen Article, Seen Query' split, adapting the model to those specific queries. Across the other splits, we notice fine-tuning in general brings improvement over the zero-shot. However, the difference of improvement decreases with 'Seen Article, Unseen Query' setting which further decreases with 'Unseen article' setting. This highlights the need of effective strategies for domain adaptation with minimal labeled domain data without getting overfitted to those specific seen queries and handle distribution shift on query side.

MSMARCO vs Legal Across all the four models, we observe LegalBERT initialization outperforms MSMARCO variant, despite the opposite trend in zero-shot performance. This is more noticeable in unseen splits, where the legal pre-training helps the model in grasping context from the under specified queries compared to general pre-trained model with exposure to general factual-based QA instances. To unveil this capability of LegalBERT in zero-shot setup, it is crucial to design a pre-training objectives closely related to the retrieval task, as discussed before, to address the task shift.

This meticulous design of three different splits, coupled with these results highlight that this dataset can serve as a testbed to study how to adapt these IR models to the distribution shifts between the source training task (such as MS MARCO) and the target tasks (such as ours) in zero-shot setup and also with minimal labeled data with some specific queries.

6. Parameter-Efficient Retrieval

PEFT aims to tune only a small portion of parameters rather than the full parameters as in traditional fine-tuning. PEFT approaches fall into three primary categories: Parameter Composition, Input Composition, and Function Composition (Ruder et al., 2022). Given a neural network $f_\theta : X \rightarrow Y$, it is decomposed into a sequence of functions $f_\theta = f_{\theta_1} \odot f_{\theta_2} \odot \dots f_{\theta_l}$, where $\theta_1, \theta_2, \dots, \theta_l$ represent parameters which are held constant in PEFT and a module with parameters ϕ is introduced, which are updated during training to modify the i^{th} sub-function as follows: Parameter composition involves interpolating models' parameter with new parameters as $f'_i(x) = f_{\theta_i \odot \phi}(x)$. Input Composition augments a model's input with a learnable parameter vector as $f'_i(x) = f_{\theta_i}([x, \phi])$ Function composition augments a model's functions with new task-specific functions as $f'_i(x) = f_{\theta_i} \odot f_\phi(x)$. We pick one representative method from each category and study their performance on our retrieval task.

[12] https://github.com/beir-cellar/beir/wiki/Leaderboard

			Seen Article Seen Query			Seen Article Unseen Query			Unseen Article		
			2%	5%	10%	2%	5%	10%	2%	5%	10%
Zero shot	BM25		0.07	0.17	0.29	0.09	0.23	0.37	0.10	0.25	0.40
	DPR		0.11	0.22	0.33	0.14	0.26	0.42	0.14	0.30	0.47
	ANCE		0.12	0.23	0.34	0.16	0.28	0.44	0.17	0.34	0.48
	COLBERT		0.16	0.32	0.47	0.17	0.34	0.51	0.24	0.41	0.56
	CrossEncoder		0.08	0.20	0.35	0.15	0.28	0.42	0.20	0.36	0.50
	LegalBERT		0.06	0.16	0.32	0.09	0.23	0.37	0.08	0.21	0.36
Fine tune	DPR	MSMARCO	0.21	0.41	0.60	0.22	0.40	0.60	0.25	0.45	0.64
		Legal	0.28	0.47	0.65	0.24	0.46	0.67	0.29	0.50	0.68
	ANCE	MSMARCO	0.22	0.43	0.62	0.24	0.41	0.61	0.26	0.46	0.66
		Legal	0.28	0.48	0.67	0.24	0.47	0.68	0.26	0.51	0.69
	COLBERT	MSMARCO	0.25	0.45	0.64	0.27	0.46	0.66	0.25	0.49	0.69
		Legal	0.29	0.49	0.69	0.29	0.49	0.69	0.27	0.51	0.70
	Cross Encoder	MSMARCO	0.26	0.48	0.69	0.30	0.50	0.71	0.31	0.51	0.70
		Legal	0.30	0.50	0.70	0.31	0.54	0.72	0.32	0.57	0.74

Table 1: Results of various systems on our Query-driven Paragraph retrieval task. For zero-shot settings, all these splits are unseen, as they are not fine-tuned on any task related data.

Adapters (Houlsby et al., 2019) fall under the category of function composition where we inject two small modules between the self-attention sub-layer and the feed forward sub-layer inside each layer of transformer sequentially. The adapter module consists of a down-projection, an up-projection and a nonlinear function between them with a residual connection across each module.

$$Adapter(h) = h + W_{up}^{\intercal}\psi(W_{down}^{\intercal}h) \quad (2)$$

where $W_{down} \in \mathbb{R}^{D_{hidden} \times D_{mid}}$ and $W_{up} \in \mathbb{R}^{D_{mid} \times D_{hidden}}$, D_{mid} denote the bottleneck dimension and ψ is a nonlinear RELU activation function.

Prefix-Tuning (Li and Liang, 2021) falls under the category of input-composition where we prepend a fixed number of trainable vectors to the input of multi-head attention in each Transformer layer, which the original tokens can attend to as if they were virtual tokens. Specifically two prefix matrices P_K and $P_V \in \mathbb{R}^{L \times D_{hidden}}$ are prepended to K and V where L denotes prefix length.

$$h = Attention(Q, [P_k, k], [P_v, v]) \quad (3)$$

LoRA (Hu et al., 2021) Low-Rank Adaptation falls under the category of parameter-composition, introduces trainable low-rank matrices and combines them with the original matrices in the multi-head attention. Specifically, it learns two low-rank matrices $W_{down} \in \mathbb{R}^{D_{hidden} \times D_{mid}}$ and $W_{up} \in \mathbb{R}^{D_{mid} \times D_{hidden}}$ for each of the query and value projections along with their original matrix W_Q and $W_V \in \mathbb{R}^{D_{hidden} \times D_{hidden}}$. Taking W_Q as example:

$$Q = (W_Q^{\intercal} + \alpha W_{up}^{\intercal} W_{down}^{\intercal})h_{in} \quad (4)$$

where α is a tunable hyper-parameter. Once after the training is complete, we can sum up these additional LoRA weights to the original weights, thus making the inference overhead to zero.

7. PEFT Experiments

We investigate the effect of PEFT by applying each method separately on bi-encoder and cross-encoder, using MSMARCO and LegalBERT initializations. Among bi-encoders, we choose COLBERT due to its better performance in full fine-training. We report Recall@k% for $k = \{2, 5, 10\}$ in Table 2.

Implementation Details We use the AdapterHub library[13] for implementing PEFT methods. For Prefix-tuning, we use prefix lengths of 10, 15 and 30. For Bottleneck adapters, we used reduction factors of 8, 16, and 32. In case of LoRA, we use configuration of rank and alpha in $\{8, 16\}$. We sweep over learning rates $\{1e-5, 3e-5, 5e-5, 1e-4, 3e-4\}$ select the best model based on the performance on the validation set. We train the model for 15 epochs with Adam optimizer (Kingma and Ba, 2014).

7.1. Results

CrossEncoder (MSMARCO): We observe all the PEFT methods under perform than full fine-tuning across all the splits. Among them, LORA underperforms consistently across all the splits, while prefix tuning is better among them. However, Adapter takes the lead in 'unseen article' split and this can be attributed to better generalization capability derived through adding new functional composition rather than additional input tokens in case of pre-

[13] https://docs.adapterhub.ml

		% train	Seen Article Seen Query			Seen Article Unseen Query			Unseen Article		
			2%	5%	10%	2%	5%	10%	2%	5%	10%
Cross Encoder MSMARCO	Full	100	0.26	0.48	0.69	0.30	0.50	0.71	0.31	0.51	0.70
	Adapter	1.6	0.25	0.45	0.63	0.28	0.47	0.67	0.30	0.50	0.68
	Pre. Tun.	0.5	0.27	0.48	0.65	0.31	0.51	0.69	0.28	0.47	0.66
	LORA	0.5	0.24	0.42	0.60	0.26	0.45	0.64	0.26	0.46	0.63
Cross Encoder Legal	Full	100	0.30	0.50	0.70	0.31	0.54	0.72	0.32	0.57	0.74
	Adapter	1.3	0.30	0.52	0.71	0.28	0.49	0.68	0.26	0.48	0.70
	Pre. Tun.	0.8	0.30	0.52	0.71	0.29	0.48	0.68	0.27	0.49	0.70
	LORA	0.9	0.29	0.51	0.70	0.28	0.48	0.69	0.27	0.49	0.70
COLBERT MSMARCO	Full	100	0.25	0.45	0.64	0.27	0.46	0.66	0.29	0.49	0.69
	Adapter	1.6	0.22	0.41	0.60	0.24	0.43	0.62	0.24	0.43	0.62
	Pre. Tun.	0.5	0.19	0.39	0.58	0.21	0.40	0.59	0.20	0.39	0.60
	LORA	0.5	0.21	0.41	0.60	0.24	0.42	0.62	0.24	0.43	0.62
COLBERT Legal	Full	100	0.28	0.48	0.67	0.24	0.47	0.68	0.26	0.51	0.69
	Adapter	1.6	0.26	0.46	0.64	0.25	0.46	0.67	0.23	0.46	0.64
	Pre. Tun.	0.5	0.20	0.40	0.61	0.21	0.41	0.61	0.19	0.40	0.57
	LORA	0.5	0.26	0.46	0.63	0.24	0.46	0.66	0.24	0.46	0.63

Table 2: Comparison between full fine-tuning and various parameter-efficient tuning methods.

fix tuning, which may to overfit on seen article splits.

CrossEncoder (Legal): We observe all the PEFT methods comparable to each other across all the splits, which can be attributed to domain-specific legal knowledge from base model. On the 'seen query' split, they even surpass the full fine-tuning, demonstrating that with fine-tuning $\sim$1% of the original model parameters, they can achieve comparable performance to the full fine-tuning baseline, makes them to adopt easily in low-compute settings. However, these methods fall back on generalizability, compared to full-tuning, opening up potential directions to tackle in future, how to augment these PEFT methods to handle these distribution shifts to perform effectively on unseen settings.

COLBERT (MSMARCO): PEFT methods underperform compared to full fine-tuning. Among them, Prefix Tuning turns out to be lowest performer and rest of them are comparable to each other, across all the splits consistently. This can be attributed to the short queries in our case. COLBERT (bi-encoder) models encode queries and paragraphs separately, and for shorter queries, they struggle to extract meaningful contextual information using BERT alone. Prefix Tuning, in particular, fails to enhance this contextual information just by adding additional parameters in the input compared to others which can handle the representation embeddings through function or parameter composition.

COLBERT (Legal): We observe similar to the COLBERT(MSMACRO), prefix tuning underperforming compared to rest of the methods.

Cross encoder vs COLBERT: While prefix tuning turned out to be a better PEFT method in cross encoder setting (especially in MSMARCO), it turned out to be lowest in bi-encoder, COLBERT, encouraging further studies to develop model agnostic PEFT methods and analyze the interplay between architecture and the PEFT method.

MSMARCO vs Legal: Overall, legal pre-training helped to account for distribution shift for corpus, demonstrating better results. This coupled with cross-encoder deep interactions, demonstrated parameter efficiency when fine-tuning. Moreover, Legal oriented models witness only a small decline with sparse fine-tuning from full fine-tuning in comparison to MSMARCO variants.

Overall, we empirically demonstrate that PEFT methods can achieve comparable performance to full-parameter fine-tuning not only in seen query setting but also in challenging unseen settings and motivate further work to bridge the existing gap between them, making them more adaptable in low data and compute resource settings.

8. Conclusion

We present an empirical study focused on the task of extracting relevant paragraphs from legal judgments based on the query. We rigorously curate a dataset for this task from ECtHR jurisdiction, leveraging the case-law guides produced by the court's registry. We assess the current retrieval models on this task in a zero-shot way to emphasize the need of retrieval specific pre-training objectives.

We further fine-tune several models encompassing bi- and cross-encoders for this task. We evaluate the generalizability of different fine-tuning models when faced with unseen concepts or queries to illustrate how legal pre-training can effectively address distribution shifts on the corpus side but still faces challenges in adapting to shift on the query side. In addition, we demonstrate the efficacy of different PEFT methods on these retrieval methods shedding light on their intricate effects concerning legal pre-training, bi-encoder, and cross-encoder models. Our findings reveal that there is no one-size-fits-all PEFT method that performs well across all settings. 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.

9. Limitations

In this study, we treat each paragraph as an independent unit during the training of neural models. However, it's important to note that paragraphs are not entirely independent; they often constitute small excerpts from longer documents, and their content may not always provide a comprehensive estimate of their relevance. To be more precise, some paragraphs draw context not just from other paragraphs within the same document but also from other documents, as evident through citations and cross-references to other judgments or texts within each paragraph. In the future, harnessing this inter-paragraph and cross-document contextual signal could lead to a more enriched understanding of each paragraph's relevance.

Furthermore, this practice of segmenting documents into smaller chunks is common in retrieval tasks, where documents are broken down into shorter lengths for indexing in retrieval systems. This may lead to losing some of their context in the process. It's worth noting that this chunking effect could be more pronounced in our task compared to more fact-focused general retrieval tasks. As a result, future research should explore methods to effectively capture contextual information from the sequential nature of paragraphs within documents and develop discourse-aware representations.

Most retriever systems follow a two-stage pipeline approach, where a pre-fetcher first aims to return all relevant ones followed by a re-ranker which attempts to make more relevant ones appear before less relevant ones. In this work, we explicitly focused our experiments on the pre-fetcher component leaving the second component for future work.

A specific challenge with respect to PEFT methods are that they converge slower and relatively more sensitive to hyper-parameters such as learning rate than full fine-tuning. We do observe same characteristics during our work and have to bypass the problem by training for more epochs and experimenting with different hyper-parameters. It is thus imperative to analyse them more theoretically and design more robust and stable training strategies for PEFT methods in the future.

Additionally, while our findings of the relevant paragraph retrieval experiments are specific to the ECtHR domain and datasets, comparable experiments in other domains will see variation based on the nature of the legal concepts and the legal documents. Nevertheless, all derivation of insights from legal case data comes with jurisdiction-related limitations.

10. Ethics Statement

With the release of our data as a public resource, we do not foresee any ethical concerns in a repurposed and bundled release of this dataset as both the judgements data and the caselaw guides are already available through the HUDOC and ECHR KS platform respectively 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 believe that this work can foster further research in this data scarce legal NLP field, to build assistive technology for legal professionals. We are conscious by employing pre-trained language models, we inherit the biases they may have acquired from their training corpus and need to be further scrutinized of any biases that may arise and is crucial to ensure that the systems developed are fair.

11. Bibliographical 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.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study

of summarization algorithms applied to legal case judgments. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 413–428. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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, 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.

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.

Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2022. Parameter-efficient fine-tuning design spaces. In *The Eleventh International Conference on Learning Representations*.

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

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.

Demi Guo, Alexander M Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4884–4896.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Euna Jung, Jaekeol Choi, and Wonjong Rhee. 2022. Semi-siamese bi-encoder neural ranking model using lightweight fine-tuning. In *Proceedings of the ACM Web Conference 2022*, pages 502–511.

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.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

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)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

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.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, and Xueqi Cheng. 2022. Scattered or connected? an optimized parameter-efficient tuning approach for information retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1471–1480.

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.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 565–576.

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.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. 2022. Unipelt: A unified framework for parameter-efficient language model tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6253–6264.

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.

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.

Vaishali Pal, Carlos Lassance, Hervé Déjean, and Stéphane Clinchant. 2023. Parameter-efficient sparse retrievers and rerankers using adapters. In *European Conference on Information Retrieval*, pages 16–31. Springer.

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.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

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.

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.

Sebastian Ruder, Jonas Pfeiffer, and Ivan Vulić. 2022. Modular and parameter-efficient fine-tuning for nlp models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 23–29.

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

Tyss Santosh, Oana Ichim, and Matthias Grabmair. 2023. 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, 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.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 1048–1064.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

Weng Lam Tam, Xiao Liu, Kaixuan Ji, Lilong Xue, Xingjian Zhang, Yuxiao Dong, Jiahua Liu, Maodi Hu, and Jie Tang. 2022. Parameter-efficient prompt tuning makes generalized and calibrated neural text retrievers. *arXiv preprint arXiv:2207.07087*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Santosh Tyss, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023. 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 1103–1103.

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.

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 N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Shanshan Xu, Leon Staufer, Santosh T.y.s.s, Oana Ichim, Corina Heri, and Matthias Grabmair. 2023a. VECHR: A dataset for explainable and robust classification of vulnerability type in the European court of human rights. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11738–11752, Singapore. Association for Computational Linguistics.

Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023b. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9576, Singapore. Association for Computational Linguistics.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. 2020. Side-tuning: a baseline for network adaptation via additive side networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 698–714. Springer.