

# LREC-COLING 2024



ICCL International  
Conference on  
Computational  
Linguistics

## Query-driven Paragraph Extraction from Legal Judgements

Santosh T.Y.S.S, Elvin Quero Hernandez, Matthias Grabmair

School of Computation, Information, and Technology; Technical  
University of Munich, Germany

# Introduction

- 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](#)).
- Identifying paragraphs relevant to the query streamlines legal research, allowing them to access information efficiently.
- Challenging task unlike traditional adhoc information retrieval.
  - Characterized by domain-specific jargon, demanding in-depth understanding of nuanced legal concepts
  - Judgments with degrees of formalism and offer varying levels of explicitness, make it difficult in discerning contexts
  - Evolving nature of the legal case law with new legal doctrines, precedents and interpretations leading to an ever evolving array of legal concepts and principles, necessitate ability to comprehend new queries

# Legal IR

- Retrieving essential legal information is integral to the workflow of lawyers,
  - Searching for legislation (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 law firm's internal support system (Moens, 2001).
- In contrast to retrieve whole cases, our task involves retrieving relevant paragraphs at a finer granularity.
- At the paragraph 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.

Our work focuses specifically on legal case retrieval.

# Lack of an Existing Dataset

- We employ distant supervision to construct a dataset for the task of query-driven relevant paragraph extraction from legal judgments
  - In the domain of European Court of Human Rights (ECtHR) which addresses grievances by individuals against states for alleged violations of human rights
  - ECtHR's Knowledge Sharing platform provide case-law guide where we use section headers as queries, mirroring the legal concepts professionals utilize when searching within ECtHR judgments.
  - Relevance signals by identifying the pinpointed citations to the paragraphs in the judgments within these guides under each section.

# Our Contributions

- Assess the performance of current retrieval models in a zero-shot manner
- Establish fine-tuning baselines with both dense bi-encoder and cross-encoder architectures.
- Explore 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 the context of our paragraph retrieval dataset

# Dataset Construction

Given a query  $Q$  and a judgement document  $J$  composed of  $n$  paragraphs  $PJ = \{p_1, p_2, \dots, p_n\}$ , the objective is to identify the subset of paragraphs which are relevant to query.

- **Judgements Collection**
  - Scrap all ECtHR judgements collection as an HTML data dump from HUDOC
  - Retain only the English documents based on their metadata
  - Parse into paragraphs

# Dataset Construction

- Query Collection
  - case-law guides accessible on ECtHR Knowledge Sharing Platform
  - Maintained by the court's registry to analyzes case law development for each convention article
  - 28 article and 8 theme-related case law guides
  - 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 and use them as queries

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
A. Freedom of expression .....	32
1. Affecting private life, image, honour or reputation .....	32
2. Hate speech .....	33
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.

# Dataset Construction

- Paragraph Relevance signal
  - Provide pinpointed paragraph references to the judgements from the ECtHR.
  - Gather all paragraph references in a specific judgement under each legal concept and mark all of them as relevant corresponding to the given query in that judgement.

### 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.

# Dataset Construction

- All judgements may not be exhaustively covered in the case-law guide unless they contribute to the expansion or contraction of existing case law.
- So we pair **queries with specific judgements referenced within them** to extract relevant paragraphs from these judgements.
- Contrasts with using all the paragraphs from all the judgements as the candidate set.
- Deliberately opt to restrict each query to the judgements referenced under to ensure a high-quality evaluation setup, controlling false negatives.

# Data Analysis

- 4109 query-judgement pairs with 708 unique queries.
- Number of total paragraphs in Judgement range from 21 to 942 with a mean of 102.78 .
- 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%.
- Queries and paragraph have a mean length of 36 and 135 tokens

# Data Analysis & Splits

- Partition the article/theme case law guides into two distinct splits:
  - Exclusively for testing with 403 query-judgment pairs (111 unique queries) derived from these - 'Unseen article/themes' - unfamiliar legal concepts from themes and articles not encountered during training.
  - Queries from the other split are further divided into two subsets,
    - 'Seen article/theme, Unseen Query' with 694 pairs (120 unique queries)
      - Exposes the model to previously encountered themes/articles, but with new queries.
    - 'Seen article/theme, Seen Query' with 3012 pairs (477 unique queries).
      - Divided into training (2230 pairs), validation (302 pairs), and test (480 pairs)

# Task Setup & Metrics

- 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.
- Recall@k% measures the proportion of relevant paragraphs in the top-k% of the total paragraphs in the judgement

# Models

- BM25 ([Robertson et al., 1995](#))
- Biencoder - Relevance score using dot product between query and paragraph representations from respective encoders
  - Random negatives as in DPR [Karpukhin et al. 2020](#)
  - Negatives using the being-optimized retrieval model as in ANCE [Xiong et al. 2020](#)
- ColBERT([Khattab and Zaharia 2020](#))
  - Queries and documents encoded at a finer granularity into multiple representations
  - Relevance as the sum of maximum similarities between each query vector and all vectors in the document
- Cross Encoder - relevance score is directly computed by feed-forward network using the combined representation of the both ([Yates et al., 2021](#))

# Zero-shot Performance

		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

- Neural models better than BM25
- COLBERT better than bi-encoders
- Cross encoders only comparable, not better, ability to act better in re-ranking stage rather than retrieval stage.
- LegalBERT falls behind necessitating retrieval specific pre-training objectives.

# Fine-tuning Performance

			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

Fine-tuning models  
(both MSMARCO  
and LegalBERT  
initialized ones)  
improve over  
zero-shot variants

# Fine-tuning Performance

			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

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.

# Fine-tuning Performance

			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

LegalBERT initialization outperforms MSMARCO variant, despite the opposite trend in zero-shot performance.

More noticeable in unseen splits.

# Parameter Efficient Retrieval

Let a neural network  $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$  be decomposed into a composition of functions  $f_{\theta_1} \odot f_{\theta_2} \odot \cdots \odot f_{\theta_l}$ . Each has parameters  $\theta_i, i = 1, \dots, l$ .

Function composition

A module with parameters  $\phi$  can modify the  $i$ -th subfunction as follows:

1. Parameter composition:  $f'_i(\mathbf{x}) = f_{\theta_i \oplus \phi}(\mathbf{x})$

Interpolation, e.g., element-wise addition

2. Input composition:  $f'_i(\mathbf{x}) = f_{\theta_i}([\mathbf{x}, \phi])$

Concatenation

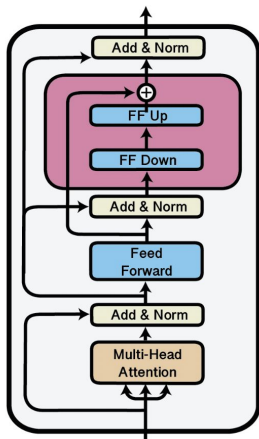
3. Function composition:  $f'_i(\mathbf{x}) = f_{\theta_i} \odot f_\phi(\mathbf{x})$

In practice, typically only the module parameters  $\phi$  are updated while  $\theta$  is fixed.

## Function Composition

- Adapters

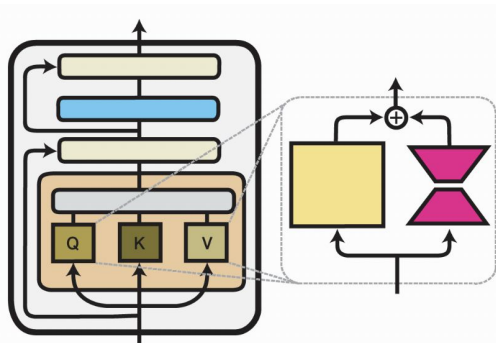
Inject two small modules inside each layer of transformer sequentially.



## Parameter Composition

- LoRA

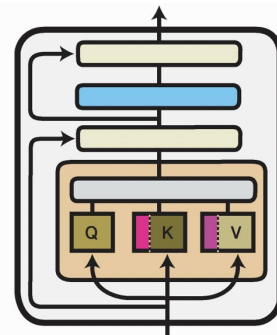
LoRA introduces trainable low-rank matrices and combines them with the original matrices in the multi-head attention.



## Input Composition

- Prefix Tuning

Prepend a fixed number of trainable vectors to the input of multi-head attention in each Transformer layer



# Parameter Efficient Retrieval Findings

		% 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

- LORA underperforms, while prefix tuning is better..
- Adapter takes the lead in 'unseen article' split - better generalization capability derived through adding new functional composition

# Parameter Efficient Retrieval Findings

		% train	Seen Article Seen Query			Seen Article Unseen Query			Unseen Article		
			2%	5%	10%	2%	5%	10%	2%	5%	10%
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

- All the PEFT methods comparable to each other due to domain-specific legal knowledge from base model.
- Still fall back on generalizability, compared to full-tuning
  - how to augment these PEFT methods to handle these distribution shifts in unseen settings.

# Parameter Efficient Retrieval Findings

		% train	Seen Article Seen Query			Seen Article Unseen Query			Unseen Article		
			2%	5%	10%	2%	5%	10%	2%	5%	10%
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

Prefix tuning turned out to be a better PEFT method in cross encoder setting (especially in MSMARCO), but the lowest in bi-encoder settings.

# Conclusions

- Conducted an empirical study on the task of extracting relevant paragraphs from legal judgments based on the query.
- Curate a dataset from ECtHR jurisdiction, leveraging the case-law guides
- Assess the current retrieval models on this task in a zero-shot way to emphasize the need of retrieval specific pre-training objectives.
- Fine-tune several models encompassing bi- and cross-encoders for this task.
  - Legal pre-training can effectively address distribution shifts on the corpus side but still faces challenges in adapting to shift on the query side.
- Efficacy of different PEFT methods on retrieval methods
  - No one- size-fits-all PEFT method that performs well across all settings.