

Introduction

- Finding relevant case law accounts for roughly 15 hours per week for a lawyer or nearly 30% of their annual working hours
- Task:** Identifying paragraphs relevant to the query, allowing lawyers to access information effectively.
- In contrast to prior works where they retrieve whole cases, our task involves retrieving relevant paragraphs at a finer granularity.
- Legal case entailment task in COLIEE deals with paragraph level retrieval - Identifying a paragraph from existing cases that matches the decision of a new case, but it employs the "entire case as the query", in contrast to the short queries used in our work.

Dataset Construction

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

Judgements Collection : Scrap all ECtHR judgements collection as an HTML data dump from HUDOC [2] and parse into paragraphs

Query Collection: Case-law guides accessible on ECtHR Knowledge Sharing Platform [1]

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.

Paragraph Relevance Signal : Provide pinpointed paragraph references to the judgements from the ECtHR.

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 issues	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	25
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. Interference with freedom of expression	33
3. Imposed silence and legal bans concerning homosexuality	34
4. Freedom of assembly and association	36

Figure 1. Query Collection

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. Paragraph Relevance Signal.

Data Analysis & Splits

- 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
- Partition the article/theme case law guides into two distinct splits:
- (A) 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.
- (B) Queries from the other split are further divided into two subsets,
 - (I) 'Seen article/theme, Unseen Query' with 694 pairs (120 unique queries): Evaluates the model to previously encountered themes/articles, but with new queries.
 - (II) 'Seen article/theme, Seen Query' with 3012 pairs (477 unique queries). Divided into training (2230 pairs), validation (302 pairs), and test (480 pairs)

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% - Proportion of relevant paragraphs in the top-k% of total paragraphs in judgement

Retrieval Models

- BM25
- Biencoder: Relevance score using dot product between query and paragraph representations from respective encoders; Random negatives as in DPR; Negatives using the being-optimized retrieval model as in ANCE
- ColBERT : Queries and documents encoded at a finer granularity into multiple representations; Relevance as sum of maximum similarities between each query vector and all document vectors.
- Cross Encoder: relevance score is directly computed by feed-forward network using the combined representation of the both

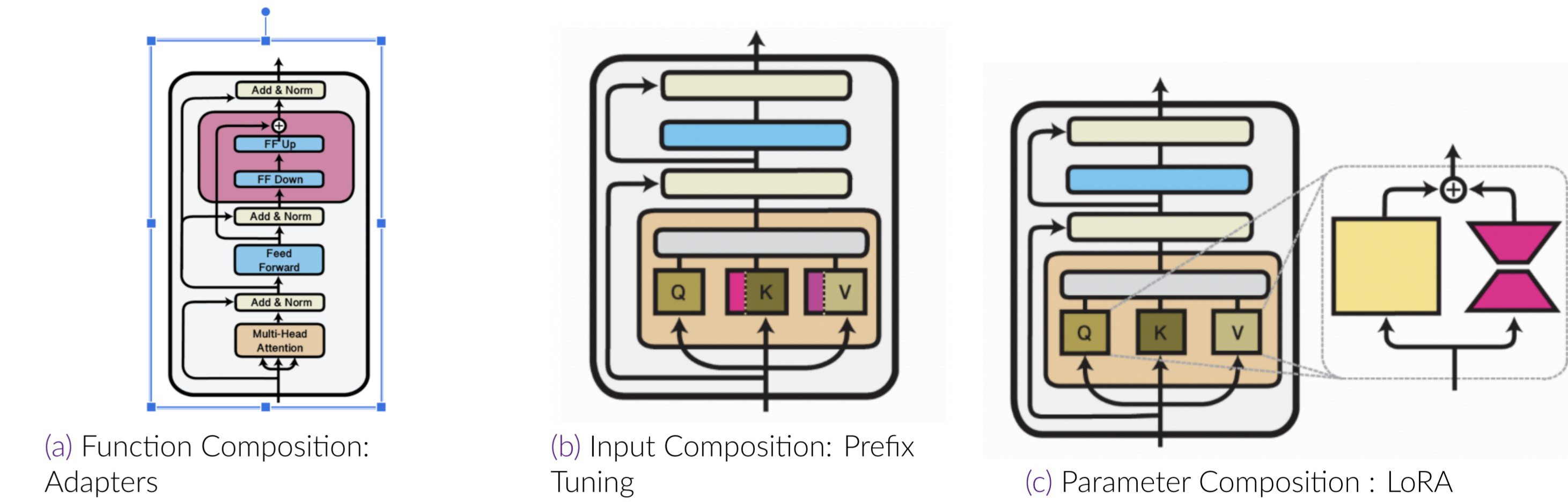
Zero-shot/Fine-tune Results

		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
	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	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.

- 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 models (both MSMARCO and LegalBERT initialized ones) improve over zero-shot variants
- 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.
- LegalBERT initialization outperforms MSMARCO variant, despite the opposite trend in zero-shot performance. More noticeable in unseen splits.

Parameter Efficient Methods



Parameter Efficient Retrieval Results

		% 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. LORA	0.5	0.27	0.48	0.65	0.31	0.51	0.69	0.28	0.47	0.66
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. LORA	0.8	0.30	0.52	0.71	0.29	0.48	0.68	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. LORA	0.5	0.19	0.39	0.58	0.21	0.40	0.59	0.20	0.39	0.60
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. LORA	0.5	0.20	0.40	0.61	0.21	0.41	0.61	0.19	0.40	0.57

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

- CrossEncoder (MSMARCO):** LORA underperforms, while prefix tuning is better..
- Adapter takes the lead in 'unseen article' split - better generalization capability derived through adding new functional composition
- CrossEncoder (Legal):** 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.
- COLBERT (MSMARCO/Legal):** Prefix tuning turned out to be a better PEFT method in cross encoder setting (especially in MSMARCO), but the lowest in bi-encoder settings.

References

- ECtHR Knowledge Sharing Platform, <https://ks.echr.coe.int>.
- HUDOC - European Court of Human Rights, <https://hudoc.echr.coe.int>.