JAIST Repository
https://dspace.jaist.ac.jp/

Title	An Effective Framework for Legal Entailment Retrieval with Large Language Models and Optimal Transport
Author(s)	TRAN, Thanh Cong
Citation	
Issue Date	2024-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/19360
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Master's Thesis

An Effective Framework for Legal Entailment Retrieval with Large Language Models and Optimal Transport

TRAN, Thanh Cong

Supervisor NGUYEN, Minh Le

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Master's Degree)

September, 2024

Abstract

Legal case entailment is a fundamental principle of the legal system in which the verdict of previous cases serves as a guiding precedent for later cases with similar factual circumstances. By following established rulings, this concept ensures judicial consistency and promotes predictability in legal outcomes. Due to the intricate nature of legal documents, identifying entailment between legal cases requires considerable time and effort, necessitating expertise in legal interpretation and analysis. In the field of legal AI development, a prominent initiative is the Competition on Legal Information Extraction & Entailment (COLIEE), held annually to drive advancements in information retrieval and entailment methods for legal texts. To address the legal case entailment task, early approaches from COLIEE utilized Bag-of-Words text representation and employed traditional machine learning methods for entailment prediction. While these approaches are fast and cost-efficient, they lack sufficient semantic and contextual representation for legal texts. Following the emergence of pre-trained language models such as BERT, subsequent methods have leveraged the language modeling capabilities of these architectures for legal case entailment and yielding promising results. Recent task-winners in the COLIEE competition for this task capitalize on Large Language Models (LLMs), particularly leveraging the pre-trained encoder-decoder MonoT5 architecture for entailment ranking and prediction. Despite their works are detailed in competition reports, there exists a significant gap in the literature dedicated specifically to legal case entailment. Furthermore, the performance benchmark of previous methods indicates opportunities for enhancement, underscoring the requirement for high-performance systems that are applicable in real-world scenarios.

To accelerate the process of legal case entailment through high-performance systems, this thesis proposes a two-stage framework centered on entailment information retrieval. We conceptualize this task as a document retrieval problem and develop a cost-efficient system that leverages advanced language models for legal case entailment. In the first stage, we introduce the ColBERT-UOT document retrieval model, which builds upon the ColBERT architecture by incorporating a sparse keyword alignment strategy utilizing the Unbalanced Optimal Transport framework. Our study demonstrates that

by focusing on the interaction of contextually and semantically similar keyword pairs between the query and the document, the proposed alignment method enhances the retrieval capability of ColBERT in the legal domain. In the second stage, we employ a fine-tuned MonoT5 document ranking model to refine the retrieval results and predict entailment instances. As a supplementary study, we benchmark state-of-the-art open-source LLMs in zero-shot legal case entailment to evaluate their performance and potential applications. We formulate the original task as a zero-shot list-wise entailment prediction and evaluate pre-trained LLMs of various sizes with diverse prompt designs to measure the capability of these models in legal reasoning.

We utilize the top-performing systems from COLIEE competitions between 2020 and 2024 as our baseline for comparison, alongside the zero-shot performance of established open-source LLMs. Extensive evaluation demonstrates that our proposed system significantly outperforms previous methods, consistently surpassing the baseline by a notable margin. Additionally, our system surpasses the zero-shot predictions of LLMs by a substantial margin, maintaining an average 3% performance gap in F1 score over the best-performing Llama3 70B LLM. By focusing on entailment retrieval, our system demonstrates a robust capability to identify entailment information within the top five candidates, achieving an average recall of approximately 90%. These results highlight the effectiveness of our approach and the promising potential of the proposed system for real-world applications.

In our analysis section, we examine the cost-effectiveness of ColBERT-UOT, as well as compare the alignment characteristics of the proposed sparse keyword alignment with the baseline approach. The analysis reveals that by focusing on the interaction of semantically and contextually similar keyword pairs between the query and the document, our proposed alignment strategy enhances the retrieval performance of ColBERT for legal texts. In this section, we also evaluate the performance of the two best-performing LLMs in legal case entailment under different prompt designs. Our findings indicate that although LLMs are sensitive to prompt formulation, they exhibit promising zero-shot performance in legal entailment scenarios.

In summary, this thesis investigates the task of legal case entailment and introduces a two-stage framework focused on entailment information retrieval. Based on this framework, our system, which consists of the ColBERT-UOT candidate retrieval model and the MonoT5 entailment prediction model, demonstrates superior performance compared to previous methods on the COLIEE datasets. The benchmarking study also underscores the potential

of LLMs in legal case entailment, despite their sensitivity to prompt design. For future work, we suggest focusing on the development of specialized LLMs tailored to the legal domain, leveraging extensive legal corpora to further support legal professionals and accelerate the analysis process of legal documents.

Acknowledgement

Firstly, I would like to express my gratitude to my supervisor, Prof. Nguyen Le Minh, for his guidance and support throughout my master's study and research at JAIST.

I am also grateful to my second supervisor, Associate Prof. Kiyoaki Shirai, and my minor research supervisor, Associate Prof. Nao Hirokawa, for their ongoing mentorship and support throughout my studies and research. I am immensely thankful to the teachers at JAIST for their invaluable lessons and knowledge transfer. To my colleagues and fellow researchers at Nguyen's Lab, thank you for your invaluable contributions and assistance. The collaborative spirit in the lab, combined with your willingness to share knowledge and offer help, has greatly enriched my understanding of complex concepts and broadened my professional expertise. I also extend a special thanks to Dr. Nguyen Minh Phuong for his guidance and insights into my research.

Lastly, I wish to express my profound gratitude to my family and friends. Their crucial role in my upbringing and their steadfast support have been instrumental in shaping the person I am today. To my grandmother, my parents, and my sister, thank you for always encouraging me and standing by my side during challenging times.

Contents

1 Introduction **1**
- 1.1 Legal Case Entailment Task 1
- 1.2 Objectives 3
- 1.3 Thesis Outline 5

2 Related Works and Background Knowledge **8**
- 2.1 Related Works 8
  - 2.1.1 Legal Case Entailment 8
  - 2.1.2 Optimal Transport for Natural Language Processing . 9
- 2.2 Background Knowledges 11
  - 2.2.1 Optimal Transport for Word Alignment 11
  - 2.2.2 Language Models for Information Retrieval 12

3 Proposed System **15**
- 3.1 Stage 1: Candidate Retrieval 16
  - 3.1.1 Word Alignment as a Similarity Measure 16
  - 3.1.2 Optimal Transport integration with ColBERT 17
- 3.2 Stage 2: Entailment Prediction 19
  - 3.2.1 Fine-tuning MonoT5 for point-wise entailment scoring 19
  - 3.2.2 Zero-shot list-wise entailment prediction with LLMs . . 21

4 Experimentation **26**
- 4.1 Experiment settings 26
  - 4.1.1 Datasets 26
  - 4.1.2 Candidate Retrieval 29
  - 4.1.3 Entailment Prediction 30
- 4.2 Experiment results 32
  - 4.2.1 Candidate Retrieval 32

4.2.2 Entailment Prediction 32

5 Analysis **37**
5.1 The effectiveness of the candidate retrieval stage 37
5.2 Improved alignment led to improved retrieval performance . . 38
5.3 Zero-shot LLMs achieve promising results in legal case entailment 42

6 Conclusion **44**

List of Figures

1.1 Overview of the framework. 4

2.1 ColBERT architecture. 13

3.1 ColBERT-UOT architecture. 17
3.2 An example of a legal entailment prediction prompt and the response from Llama3 70B. 23
3.3 Examples of the prompt templates used for zero-shot legal entailment prediction. 24

4.1 Distribution of the lengths of the entailed paragraphs in COLIEE 2024 dataset. 27
4.2 Distribution of the lengths of the reference paragraphs in COLIEE 2024 dataset. 27

5.1 Number of candidate paragraphs in the COLIEE test sets. . . 38
5.2 Number of input tokens in the COLIEE test sets. 39
5.3 Visualization of keyword matching using the MaxSim^F alignment function. 40
5.4 Visualization of keyword matching using the UOT alignment method. 40

List of Tables

1.1 An example of a decision and the set of reference legal paragraphs 7

4.1 Statistics of the legal case entailment datasets. 28
4.2 Percentages of number of entailment per query across datasets. 28
4.3 Performance of the retrievers in COLIEE datasets, and the average performance. 33
4.4 Performance of the entailment predictors in COLIEE 2020 and 2021 datasets 34
4.5 Performance of the entailment predictors in COLIEE 2022 and 2023 datasets 35
4.6 Performance of the entailment predictors in COLIEE 2024 dataset, and the average performance. 36

5.1 Statistics on the number of alignment links produced by the alignment functions in the COLIEE 2024 dataset. 39
5.2 Performance of LLMs across 4 prompt design categories in the COLIEE 2023 test dataset. 42
5.3 Performance of LLMs across 4 prompt design categories in the COLIEE 2024 test dataset. 42

Chapter 1

Introduction

1.1 Legal Case Entailment Task

The emergence of Large Language Models (LLMs) has marked a paradigm shift in the field of Natural Language Processing (NLP), fundamentally altering the landscape of research and development of NLP systems. State-of-the-art LLMs, such as OpenAI's GPT [43], have achieved unprecedented success in understanding human instruction and performing complex tasks by leveraging vast amounts of data and powerful neural network architectures. The sheer scale of parameters, often numbering in the millions or billions, allows LLMs to capture intricate linguistic patterns and context, enabling them to excel in a wide range of NLP tasks [11] [10] [32].

In the realm of legal studies, the advancement of LLMs presents promising opportunities for developing automated systems capable of handling legal documents and executing intricate legal tasks. Advanced legal AI systems have the potential to make substantial contributions, especially considering that legal documents are inherently complex, featuring intricate language, dense terminology, and nuanced interpretations that demand expert understanding [26]. Furthermore, legal tasks such as case research, document comparison, and precedent identification are highly time-consuming processes. They demand a thorough examination of vast amounts of information, often requiring cross-referencing with multiple sources and historical case law. This complexity and the time-intensive nature of legal tasks pose substantial challenges, highlighting the need for advanced algorithms and models to improve operational efficiency in the legal field.

A key initiative promoting the development of automated systems for legal document processing is the **Competition on Legal Information Extraction & Entailment (COLIEE)** [1], an annual competition that aims to accelerate the advancement of state-of-the-art information retrieval and entailment methods for legal texts. In this work, we focus on the legal case entailment task from COLIEE, which presents a critical aspect of legal reasoning and decision-making. This concept, intrinsic to the common law tradition, emphasizes that judicial rulings are not discrete occurrences but rather interconnected elements shaping a cohesive legal framework, where each decision builds upon and refines the principles established in preceding cases.

At its core, the legal case entailment task involves juxtaposing the decision of a new case (referred to as Q) with relevant case materials (represented by a reference case R) to pinpoint a set of specific paragraphs (referred to as R^*) within case R that entails the decision Q. In this context, "decision" refers not to the final resolution of a case but rather to a specific conclusion articulated by the judge supported by one or more particular paragraphs from the reference case. Unlike conventional information retrieval techniques, which may falter in capturing the nuanced entailment relationships inherent in legal texts, the entailment task demands a deeper semantic understanding and comparison of the textual content between the cases. It is crucial to note that while case R shares relevance with decision Q, mere relevance does not suffice for entailment determination. Within case R, numerous paragraphs may exhibit relevance to decision Q without necessarily entailing it. Therefore, the task necessitates a tailored entailment approach that delves into the semantic congruence and logical coherence between each paragraph in case R and decision Q to identify a concise set of entailment information R^*.

We present a scenario of legal case entailment in Table 1.1. To validate the legal reasoning behind the decision, *"A finding for which no additional evidence is filed should be accorded considerable deference"*, a reference case is provided to identify the supporting information within specific paragraphs. Each paragraph, indexed with a paragraph number, contains information related to specific aspects of the reference case. In this example, paragraph number 001 provides background information on the reference case. Paragraph 033 discusses the correctness standard in the context of findings of fact with additional evidence. It differentiates it from the Registrar's findings on

[1] https://sites.ualberta.ca/ rabelo/COLIEE2024

other facts and thus does not address the situation where no additional evidence is filed. Conversely, the highlighted text fragment in paragraph 034 directly supports the current decision by explicitly stating that a considerable degree of deference should be given to the Registrar's findings when no significant new evidence has been presented. This aligns with the examined decision, which asserts that findings without additional evidence should be accorded considerable deference, thereby establishing an entailment relationship. Lastly, the final two paragraphs do not provide direct support to the decision as they conclude the situation where additional evidence is substantial and potentially alters the court's approach to making findings of fact. Hence, among the 101 paragraphs examined from the reference case, paragraph 034 is the sole entailing paragraph that directly substantiates the decision.

1.2 Objectives

Previous entries in the COLIEE competition have shown promising performance in the legal case entailment task. Early approaches [23] [13] propose to represent legal paragraphs using the Bag-of-Words model and deploy traditional machine learning methods to predict the entailing paragraphs. Recent strategies [49] [48] [38] [41] have leveraged the MonoT5 [42] document ranking model to score reference paragraphs and applied heuristics to identify the entailing paragraphs. However, these methodologies show considerable similarity, with MonoT5 consistently appearing in the leading solutions for the past four years of competition. While these findings are presented in the form of competition reports, there remains a notable gap in the literature specifically focusing on legal case entailment. To encourage the implementation of AI systems in this domain, a thorough evaluation and comprehensive study are essential to establish the baseline approach for future development.

In this thesis, we present a two-stage framework that utilizes language models for legal entailment retrieval and prediction. Our approach is designed to efficiently retrieve entailing paragraphs through the utilization of dense text representations and neural entailment models. Figure 1.1 illustrates the structure of our proposed framework, which consists of two stages: candidate retrieval and entailment prediction.

In the candidate retrieval stage, the objective is to rapidly and efficiently gather a subset of candidates from the initial pool that closely matches a

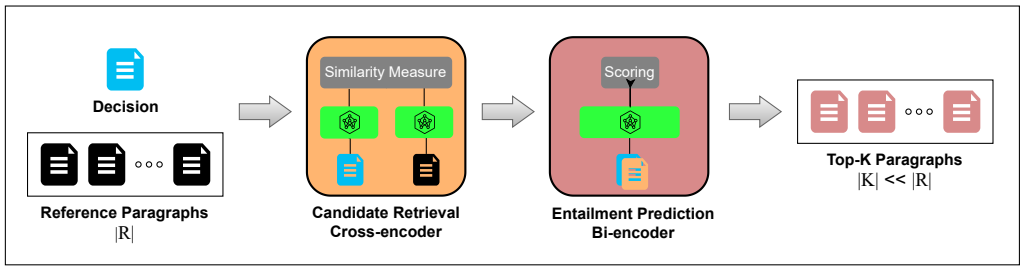

Figure 1.1: Overview of the framework.

given query input. We leverage the state-of-the-art **ColBERT** architecture [51] for this stage to represent legal texts within a Bag-of-Embeddings (BoE) model. This multi-vector text representation approach incorporates the semantic and contextual encoding of neural models while providing interpretability through word similarity and alignment capabilities that neural sentence or document embedding models typically lack. However, CoBERT utilizes the MaxSim approach to conduct one-to-one word alignment, treating the interaction of relevant word pairs and noisy word pairs equally. Building on prior research on word alignment using Optimal Transport (OT) [59] [2], we propose to substitute the MaxSim word alignment method in the original ColBERT architecture with a sparse keyword alignment strategy based on the **Unbalance Optimal Transport (UOT)** framework [14]. This adaptation aims to focus on aligning closely related keyword pairs that directly contribute to the relevance between legal paragraphs.

In the entailment prediction stage, upon receiving the shortlist of candidates from the retrieval stage, the objective is to accurately predict the documents that entail the input query. Extending the methodology from [38], we enhance the fine-tuning process of **MonoT5** to improve the legal entailment scoring capability of the model. The fine-tuned MonoT5 model refines the retrieval results from the initial stages and achieves state-of-the-art performance in the legal case entailment task, significantly outperforming previous approaches across evaluation datasets. In addition to utilizing the specialized MonoT5 model for entailment prediction, we conduct an extensive evaluation of established open-source LLMs in the legal case entailment task. Our motivation stems from the demonstrated reasoning capability of

LLMs across various benchmarks [2] [3]. Through our experiments, we provide valuable insights into the potential applications of current state-of-the-art LLMs in legal case entailment.

In summary, this thesis provides the following contributions:

- We propose integrating the ColBERT document retrieval model with a sparse keyword alignment strategy based on the Unbalanced Optimal Transport (UOT) framework, referred to as ColBERT-UOT. The proposed retrieval method outperforms established baselines across several legal case entailment datasets and demonstrates consistent retrieval performance.
- We formulate the legal case entailment task as a document retrieval problem and introduce a general two-stage framework with a focus on entailment information retrieval. Based on this framework, we employ the proposed ColBERT-UOT model for candidate retrieval and MonoT5 for entailment prediction. The proposed system demonstrates state-of-the-art performance on the COLIEE datasets, surpassing the baselines by a notable margin.
- We conduct an additional study of the performance of open-source LLMs and offer valuable insights into their potential application for the legal case entailment task.

1.3 Thesis Outline

The rest of this thesis is organized as follows:

- Chapter 2 reviews the existing literature pertinent to this study and provides the background knowledge of our methodology.
- Chapter 3 describes in detail our proposed two-stage system for this task and the zero-shot list-wise entailment prediction settings for LLMs.
- Chapter 4 outlines the experimental settings, presents the evaluation results, and provides an analysis of those results.

[2] https://www.vellum.ai/llm-leaderboard
[3] https://rank.opencompass.org.cn/leaderboard-llm

- Chapter 5 provides a detailed examination of the characteristics of the proposed ColBERT-UOT and the variation in zero-shot performance of LLMs.
- Chapter 6 concludes the thesis and discusses potential future directions.

Table 1.1: An example of a decision and the set of reference legal paragraphs

Decision (Q)
A finding for which no additional evidence is filed should be accorded considerable deference.
Reference Paragraphs (R)
001: Since 1984 Garbo Group Inc., or its predecessor, Garbo Creations Inc., has been selling in retail stores across Canada a range of goods that are described compendiously as "women's fashion accessories". Between 1984 and 1991 the goods expanded to include the following: jewellery (precious, semi-precious and costume), hair ornaments, button covers, jacket clips, handbags and belts.
...
033: The provision in s.56 (5) allowing the parties to introduce additional evidence on the appeal may suggest a correctness standard on those findings of fact to which the evidence relates. However, it does not necessarily follow that the same standard should apply to the Registrar's findings on other facts.
034: To conclude, it is my opinion after weighing these factors that, despite the inclusion in the Trade-marks Act of an untrammelled right of appeal and the right to adduce additional evidence, a considerable degree of deference is called for on the part of the appellate court when reviewing the Registrar's finding of confusion, provided at least that no significant new evidence has been adduced on a factual issue and it is not alleged that an error of law has been committed.
...
100: If, on the other hand, the additional evidence goes beyond what was in substance already before the Registrar, then the court should ask whether, in the light of that material, the Registrar reached the wrong decision on the issue to which that evidence relates and, perhaps, on the ultimate decision as well. The more substantial the additional evidence, the closer the appellate court may come to making the finding of fact for itself.
101: For these reasons, the appeal is dismissed. The parties have 14 days from the date of this decision to make submissions in writing on the award of costs. Appeal dismissed.
Entailing Paragraphs (R^*)
034

Chapter 2

Related Works and Background Knowledge

2.1 Related Works

2.1.1 Legal Case Entailment

The task of legal case entailment originates from COLIEE, an annual international event that focuses on advancing research in the field of legal informatics. The competition centers around two primary domains of legal data: case law and statute law [19]. Case law involves collections of judicial decisions and precedents, which are crucial for understanding how legal principles have been applied in previous cases. On the other hand, statute law involves legislative texts such as civil codes or statutes, providing the legal framework within which decisions are made and interpreted. For each domain, COLIEE features both an information retrieval and an information entailment task. In the domain of case law, the first task, named legal case retrieval, involves gathering pertinent supporting cases based on a provided query case. The second task, legal case entailment, requires identifying a paragraph from past cases that logically justifies the decision made in a new case.

In this study, we concentrate on the legal case entailment task. In COLIEE 2019, the best-performing system [23] utilizes histogram feature vectors to represent the similarity between reference paragraphs and decisions, which are then used as input to train a Random Forest classifier for binary entailment classification. Since the introduction of pre-trained language models

like BERT [17], subsequent methods have leveraged these models for legal case entailment. [45] suggested using a threshold-based heuristic on cosine similarity measures and then ensembling the prediction with BERT to form the entailment prediction. [56] proposed to ensemble the scores of BM25 with a fine-tuned BERT architecture for supporting text-pair classification, achieving the highest performance in COLIEE 2020. [28] and [6] conduct experiments on fine-tuning LegalBERT [8], the first language model specifically tailored for legal texts, for legal case entailment. [50] investigated the zero-shot performance of MonoT5, a point-wise document ranking adaptation of the encoder-decoder sequence-to-sequence T5 [47] LLMs, for legal entailment with and achieved first place in COLIEE 2021. Since this initiative, subsequent winning methods for the legal case entailment task have been developed based on performing document ranking with MonoT5. [48] perform zero-shot legal entailment scoring utilizing pre-trained MonoT5 models of various sizes and ensemble their results, securing first place in COLIEE 2022. [38] propose adapting a MonoT5-large model for the legal domain by introducing a fine-tuning procedure that incorporates hard negative sampling, leading to top performance in COLIEE 2023. Following this work, [41] fine-tuned a MonoT5 3B variant using a similar hard negative sampling procedure and prediction heuristic, resulting in the highest performance in COLIEE 2024. In line with the promising results of LLMs for retrieval-based applications [53] [44], [16] attempted to utilize the closed-source GPT-3.5 LLM from OpenAI to perform zero-shot document re-ranking on candidate entailment paragraphs. [39] and [40] employ open-source LLMs to perform zero-shot and few-shot entailment extraction on the top-ranked candidates identified by MonoT5. In this thesis, we propose a general two-stage framework with a focus on legal entailment retrieval performance and efficiency. Based on this framework, we develop a retrieval system composed of language models that achieves state-of-the-art results in legal case entailment.

2.1.2 Optimal Transport for Natural Language Processing

The OT framework, first introduced by [35], aims to determine the optimal plan for minimizing the cost of transporting resources to different target locations. [25] reformulated the transportation problem as one involving the transfer of mass between two probability distributions, which can be

addressed using linear programming techniques. To improve the high computation cost associated with finding the optimal coupling, [15] proposed the entropic regularization version of OT, which serves as an approximation of the original problem and can be solved efficiently with the Sinkhorn-Knop [29] iterative matrix scaling algorithm. To support the transportation of distributions with different masses, the UOT framework [14] relaxes the hard marginal constraints of the original OT problem, encouraging more flexible alignment and avoiding counterintuitive matchings with high transportation costs.

Recently, the OT framework has found many applications in the field of NLP due to its ability to offer a direct solution for aligning text entities. [30] introduced Word Mover's Distance, a metric designed to assess the dissimilarity between two documents based on the OT alignment of pre-trained word embeddings. [61] [12] adopted OT cost as the metric for evaluating machine text generation. [62] employed the OT framework as the matching procedure in bilingual lexicon induction. [1] proposed to use the OT alignment cost as the optimization objective for fine-tuning contextualized embeddings for downstream cross-lingual transfer. [36] utilized OT to learn rationale text matching for measuring scientific document similarity. [60] and [54] define a bi-level optimization problem based on inverse OT for rational alignment in legal case matching. [20] proposed using OT for knowledge distillation in low-resource cross-lingual information retrieval.

In the application of OT for word alignment, [59] proposed an enhanced version of the Word Mover's Distance by decoupling the word embeddings into their norm and direction and employing the OT framework on the direction vectors. [31] proposed an OT-based contrastive learning framework for semantic textual similarity based on word alignment. [2] applied UOT for unbalanced word alignment with null alignment support. In this work, we extend the unbalanced word alignment method to a sparse keyword alignment strategy aimed at measuring similarity between legal texts. We apply the UOT framework to the text embeddings generated by ColBERT and employ custom-designed heuristics for sparse alignment links, resulting in improved performance of ColBERT in retrieving legal texts.

2.2 Background Knowledges

2.2.1 Optimal Transport for Word Alignment

The OT problem addresses the problem of finding the most efficient way to transform the mass of one probability distribution into another while minimizing the cost associated with a distance metric. In the context of word alignment between two paragraphs, the problem of finding the optimal word alignment can be viewed as an OT problem where each paragraph is represented by a discrete probability distribution of its words, and the distance metric is defined by a dissimilarity metric between the word embeddings vectors. Following this perspective, we can define the word alignment problem as an OT problem and utilize optimization algorithms to achieve optimal alignment solutions.

Let $\mathbf{u} \in \sum_n$ and $\mathbf{v} \in \sum_m$ denote the probability distributions associated with the query and document, respectively, where $\sum_n = \{x \in \mathbb{R}_+^n : x^\top \mathbf{1}_n = 1\}$ is the probability simplex with dimension n and $\mathbf{1}_n$ is a n-dimension vector of ones. We define a distance function $f(\cdot, \cdot)$ to measure the dissimilarity between two embedding vectors and form the cost matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$ with $\mathbf{C}_{i,j} = f(\mathbf{E}_{q_i}, \mathbf{E}_{d_j})$. Given the two probability distributions and their associated cost matrix, the OT framework seeks to find the optimal transportation/alignment matrix $\mathbf{P}^* \in \mathbb{R}_+^{n \times m}$ that minimizes the alignment cost between the query Q and the document D:

$$\mathbf{P}^* = \min_{\mathbf{P} \in U(\mathbf{u}, \mathbf{v})} \langle \mathbf{C}, \mathbf{P} \rangle \tag{2.1}$$

with $U(\mathbf{u}, \mathbf{v}) = \{\mathbf{P} \in \mathbb{R}_+^{n \times m} | \mathbf{P}\mathbf{1}_n = \mathbf{u}, \mathbf{P}^\top \mathbf{1}_m = \mathbf{v}\}$ is the space of all possible alignment matrices. The hard constraints on the alignment solutions introduced in this space guarantee that the total mass is preserved after the transformation; the mass of each query token is transferred to tokens on the document side, ensuring that the total transportation mass equals the assigned mass. Consequently, this can lead to counter-intuitive matchings of word pairs with high transportation costs. Following previous works [2] [12], we adopt the same approach by replacing the hard constraints on the alignment matrices with the following regularizations, thereby constituting the UOT problem:

$$\mathbf{P}^\tau = \min_{\mathbf{P} \in \mathbb{R}_+^{n \times m}} \langle \mathbf{C}, \mathbf{P} \rangle + \tau_q KL(\mathbf{P}\mathbf{1}_n, \mathbf{u}) + \tau_d KL(\mathbf{P}\mathbf{1}_m, \mathbf{v}) \tag{2.2}$$

where $KL(\cdot,\cdot)$ is the Kullback-Leibler divergence function and τ_d and τ_q are the regularization terms for penalizing the mass deviation of the alignment matrix from the original mass of the probability distributions. With the regularization introduced in the UOT framework, the search space of OT solvers extends throughout $\mathbb{R}_+^{n\times m}$, enabling a shift in focus towards aligning closely related token pairs while maintaining alignment mass proximity to the original distributions.

The optimization problem in Equation 2.2 can be viewed as a linear programming problem. Linear programming optimization algorithms, such as the network simplex method [5], can be utilized to compute the optimal alignment matrix. However, the optimal solution provided by the LP algorithms requires cubic time complexity [7], which limits its scalability in effectively solving large-scale OT problems. To address this computational issue, [15] proposed to add an entropic regularization term to the original problem to make the problem strictly convex, enabling efficient approximation of the optimal solution using the Sinkhorn iterative matrix scaling algorithm [29]. Following this approach, the UOT problem in Equation 2.2 is regularized as follows:

$$\mathbf{P}^\epsilon = \min_{\mathbf{P}\in\mathbb{R}_+^{n\times m}} \langle \mathbf{C}, \mathbf{P}\rangle + \epsilon H(\mathbf{P}) + \tau_q KL(\mathbf{P}\mathbf{1}_n, \mathbf{u}) + \tau_d KL(\mathbf{P}\mathbf{1}_m, \mathbf{v}) \tag{2.3}$$

where $H(\mathbf{P}) = -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1)$ is the negative entropy of the alignment matrix and ϵ is the regularization factor. Given the query Q, the document D, their corresponding cost matrix $\mathbf{C}$ and the alignment matrix $\mathbf{P}^\epsilon$ from Equation 2.3, the OT distance between Q and D is calculated as follows:

$$W(Q, D) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{A}_{i,j} \tag{2.4}$$

with $\mathbf{A} = \mathbf{P}^\epsilon \odot \mathbf{C}$ is the element-wise product of the alignment and the cost matrix.

2.2.2 Language Models for Information Retrieval

A typical information retrieval system usually consists of a two-stage process: candidate retrieval and re-ranking. In the candidate retrieval stage, the system quickly filters a vast corpus to generate a preliminary set of documents that are potentially relevant to the input query. This is frequently

achieved using embeddings, where both the query and documents are represented as dense vectors in a shared high-dimensional vector space derived from language models. This approach, known as cross-encoding, enables efficient similarity calculations to pinpoint the top candidates. In the re-ranking phase, advanced language models conduct a more sophisticated analysis of the top candidates. This process, referred to as bi-encoding, involves evaluating the relevance of each document alongside the query, taking into account broader contexts and deeper semantic relationships. By refining the initial results, the re-ranking stage ensures that the final set of documents presented to the user not only meets relevance criteria but also prioritizes them according to the user's intent. This two-stage approach leverages the speed and scalability of embeddings for initial retrieval and the nuanced understanding of language models for precise re-ranking, resulting in a more effective and accurate information retrieval system.

In this thesis, we employ the ColBERT architecture for candidate retrieval. ColBERT is a novel neural text embedding method designed for information retrieval and search tasks. Leveraging BERT as its backbone, ColBERT generates rich contextual embeddings and employs a late interaction paradigm to measure similarity scores between query and document embeddings. By incorporating the late interaction mechanism, ColBERT achieves cost-efficiency by supporting the pre-computation and indexing of document embeddings through the separate encoding of queries and documents. Several out-of-domain benchmarks have demonstrated the robust out-of-domain generalization capabilities of ColBERT [51] [55].

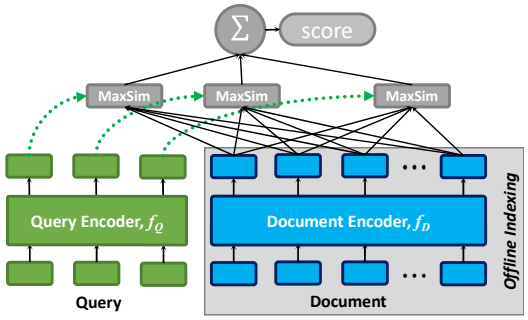

Figure 2.1: ColBERT architecture.

Figure 2.1 illustrates the original architecture of ColBERT. The query and document are first encoded independently following a BoE model into the sets

of contextual embeddings vectors by a shared BERT backbone. During the similarity measurement stage, the MaxSim alignment function is employed to link each query token to the most similar document token based on the dot product of their respective embedding vectors. The relevance score for a query-document pair is calculated by aggregating the maximum similarity score of each query token. For detailed architecture and training procedures of ColBERT, we refer readers to the original thesis [27]. This study concentrates on enhancing the alignment method to achieve improved retrieval performance.

Let $q = [q_0, q_1, \ldots, q_n]$ and $d = [d_0, d_1, \ldots, d_m]$ denote the lists of subword tokens generated from the BERT WordPiece tokenizer for the query Q and document D, respectively. Let $\mathbf{E}_q \in \mathbb{R}^{n \times h}$ and $\mathbf{E}_d \in \mathbb{R}^{m \times h}$ denote the corresponding matrices of query and document embeddings with hidden dimension h obtained from the BERT backbone. Under the MaxSim alignment function, the similarity score of the query Q and document D is calculated as follows:

$$S_M(Q, D) = \sum_{i=0}^{n} MaxSim(q_i, d) = \sum_{i=0}^{n} \max_{j \in [1..m]} \mathbf{E}_{q_i} \cdot \mathbf{E}_{d_j}^{\top} \tag{2.5}$$

In the re-ranking stage, we employ the encoder-decoder MonoT5 architecture to rank the relevancy of query-document pairs. MonoT5 is a novel method for scoring document similarity through fine-tuning the pre-trained encoder-decoder T5 LLM on the MS-MARCO passage ranking dataset [3]. To perform point-wise query-document similarity scoring with the T5 sequence-to-sequence architecture, [42] propose leveraging the following input template:

"Query: {`content of Q`} *Document:* {`content of D`} *Relevant:"*

A pre-trained T5 checkpoint is employed and undergoes fine-tuning to generate either the special `"true"` or `"false"` tokens for a given training query-document pair [42]. The similarity score between query Q and document D is defined as the probability that the first token generated by the model is the special token `"true"`.

Chapter 3

Proposed System

Formally, given a base case B, one fragment of text Q presenting a decision made within the base case B, and a reference case $R = [R_1, R_2, \ldots, R_n]$ (where R_i represents a reference paragraph within R), the objective is to identify the subset of paragraphs $R^* \subset R$ that entail Q. In this thesis, we conceptualize this task as a document retrieval problem, where the decision Q serves as the input query, the set R denotes the candidate pool, and the subset R^* represents the entailing documents to the query Q. For the rest of this thesis, following the information retrieval terminology, we refer to the decision Q as the input query and the reference paragraphs R_i as the candidate documents.

To tackle the entailing document retrieval problem, we propose a two-stage framework. The first stage is candidate retrieval, where we deploy a pre-trained ColBERT model enhanced with the UOT framework to efficiently retrieve a small set of candidates for the given input query. In the second stage, a fine-tuned MonoT5 model is employed to score the candidates identified in the first stage, re-ranking the retrieval results and generating the entailment predictions. As part of our supplementary research, we investigate the performance of open-source pre-trained LLMs in the legal case entailment task. We utilize pre-trained & instruction fine-tuned LLMs to conduct zero-shot list-wise entailment prediction on the top-k candidates retrieved by the fine-tuned MonoT5 model. We describe the details of each stage in the following sections.

3.1 Stage 1: Candidate Retrieval

The candidate retrieval stage aims to retrieve a set of potentially relevant documents that correspond to the input query. An efficient retrieval method can quickly filter irrelevant documents and greatly reduce the computational cost of subsequence stages while maintaining the high recall of the system. In this stage, we utilize ColBERT to encode text into BoE representations. As an enhancement to the word alignment method suggested in the original thesis, we introduce a sparse keyword alignment strategy based on the UOT framework for measuring the similarity of query-document pairs.

3.1.1 Word Alignment as a Similarity Measure

Although the BERT architecture operates at the subword level, the MaxSim function in (2.5) can be interpreted as an effort to perform word alignment between query-document pairs. Each query token, corresponding to a query word, is aligned with the most similar token in the document based on the semantic similarity measured by the dot product of their embeddings. Let $a = [a_0, a_1, \ldots, a_n]$ denote the list of alignment indices for the query tokens, where a_i is the index of the document token that is most similar to the query token q_i. From the MaxSim alignment function, we can form a token alignment matrix $\mathbf{P} \in \mathbb{R}^{n \times m}$ as follows:

$$\mathbf{P}_{i,j} = \begin{cases} 1 & \text{if } a_i = j \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The subword alignment matrix between the query and document demonstrates the interpretable characteristic of the ColBERT architecture. However, the alignment strategy introduced in MaxSim enforces a strict one-to-one alignment, meaning each query token must align with a single document token. We argue that this approach is overly rigid and can lead to noisy alignments in many scenarios. For instance, there are common situations where a single word from the query can be matched with multiple words in the document. In cases where no suitable alignment exists for a query word, the MaxSim alignment strategy will force an alignment to the nearest but irrelevant word in the document. Moreover, in the legal domain, documents frequently contain legal keywords and terminology that are crucial for understanding the content and implications of the text. In legal information

retrieval, we argue that highlighting the presence of semantically and contextually relevant keyword pairs is of high importance, and the outcomes of document retrieval should prioritize the connection between these terms. Following this argument, the alignments between non-relevant or high-frequency words should have a low impact on the retrieval score, as they can introduce noise that negatively affects the retrieval results. To enhance the efficiency of retrieving legal documents, we propose using a sparse keyword alignment strategy that generates weighted alignment links as an alternative to the rigid alignment links introduced in the MaxSim alignment function.

3.1.2 Optimal Transport integration with ColBERT

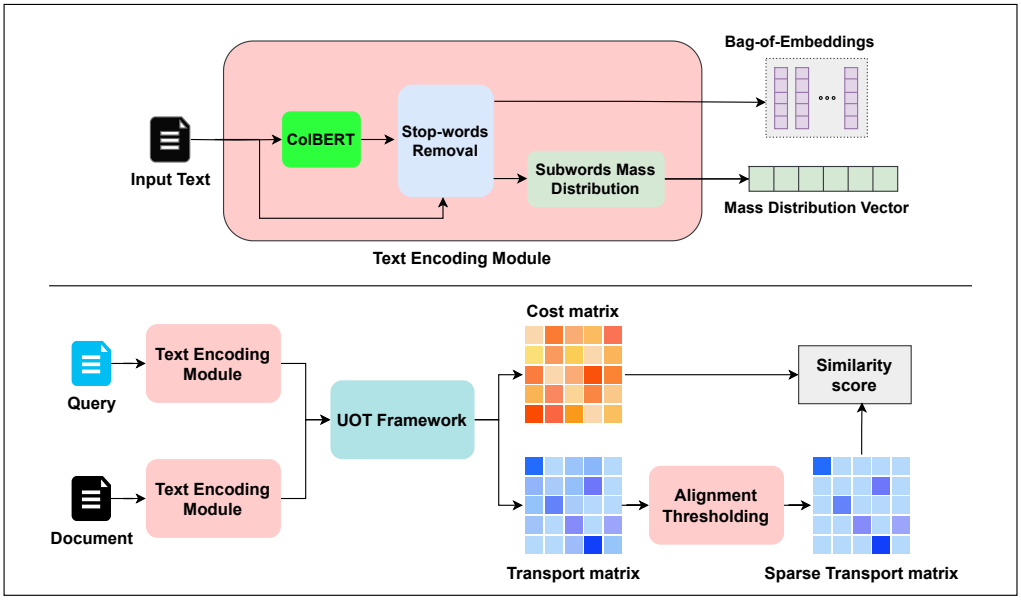

Figure 3.1: ColBERT-UOT architecture.

Departing from the MaxSim approach in the original ColBERT architecture, we propose employing a sparse keyword alignment approach within the UOT framework as an alternative method to measure similarity between query-document pairs. By focusing on the alignment of closely related terms, our method aims to more effectively capture the relevance of legal documents to a given query. Figure 3.1 depicts the architecture of our proposed method.

To adapt the UOT word alignment framework described in Section 2.2.1 into a sparse keyword alignment strategy, we introduce the following processes:

- Stop-words removal: As discussed in section 3.1.1, to discard the alignment links associated with high-frequency but context-irrelevant words, we extract the stop-word tokens from the query Q and document D, along with their corresponding embeddings vector from the embeddings matrices $\mathbf{E}_q$ and $\mathbf{E}_d$ before applying the UOT framework. We note that despite the irrelevancy of the stop-words to the context, they still influence the semantic meaning of other tokens' embeddings. Therefore, we only remove the stop-words from the texts after generating the BoE representation.

- Subwords mass distribution: The alignment process operates at the subword level. For words composed of multiple subwords, a naive approach would treat each subword as an individual token and assign it the same mass as a single subword token. However, this method can result in lengthy words with many subwords carrying excessive mass, leading to over-alignment and noisy outcomes. To address this issue, we evenly distribute the original mass of the word among its subwords. The mass distribution vectors are then normalized to form probability distribution vectors.

- Alignment weight thresholding: One characteristic of the Sinkhorn algorithm is that it produces dense alignment matrices containing numerous entries that are not statistically significant [29]. In the context of document retrieval, sparse alignment between keywords enhances interpretability and diminishes noise in the retrieval score. To promote sparsity in alignment matrices, we employ the following threshold-based heuristics to remove low-weight entries from the alignment matrix $\mathbf{P}^{\epsilon}$. To pinpoint strong alignment links, we use the k-th highest weight as the initial threshold for selecting the first set of alignment links. To capture weaker but potential matching, we select the document token with the highest alignment weight for each query token, thereby forming the second set of alignment links. We merge two sets and apply a final thresholding with value λ to obtain the final set of alignment links. We set $k = 10$ and $\lambda = 0.01$ based on the method's performance on the validation set.

Following previous works [30], we assign uniform mass to the words of the query and document before dividing the mass of multi-token words during the subwords mass distribution process. Since CoLBERT employs the dot product as the similarity measure function, we construct the cost matrix $\mathbf{C}$ by computing the negative dot product of the embeddings matrices, in particular $\mathbf{C} = -\mathbf{E}_q \cdot \mathbf{E}_d^\top$. The UOT alignment problem is constructed as outlined in section 2.2.1, and the similarity score between a query Q and a candidate document D is calculated by taking the negative of the OT distance from (2.4).

3.2 Stage 2: Entailment Prediction

After obtaining the retrieval results from the candidate retrieval stage, the objective of this stage is to identify the documents that entail the input query. In this thesis, we evaluate two approaches for entailment prediction. In the first approach, we fine-tune a MonoT5 model on the COLIEE legal entailment dataset for point-wise legal entailment scoring. In the second approach, we perform zero-shot list-wise entailment prediction by prompting the LLMs to extract entailing documents from a candidate list using a variety of prompt designs.

3.2.1 Fine-tuning MonoT5 for point-wise entailment scoring

In this study, we utilize the MonoT5 architecture and conduct additional fine-tuning for the task of legal document ranking. Algorithm 1 outlines the fine-tuning procedures we implemented for MonoT5. We create a training ranking dataset by leveraging the training segment of the COLIEE legal case entailment dataset. For each training query, we identify the documents that entail the query as the relevant documents. To mine hard negative samples, we sorted the non-entaiment documents based on their scores derived from ColBERT-UOT. In each epoch, we select $n_s = 5$ documents from the top of this sorted list to serve as hard negative samples for each query. These documents are then removed from the list of negative samples, ensuring that new negative samples are rotated in for the following epochs. During the validation phase, we utilize the Mean Reciprocal Rank (MRR) as the primary

Algorithm 1 MonoT5 fine-tuning procedure for legal case entailment

Require: $S_{train} : \text{LIST}[(Q, R^*, R^n)]$ ▷ Training dataset consists of (query, relevant docs, non-relevant docs) tuples
Require: $S_{val} : \text{LIST}[(Q, R^*, R^n)]$ ▷ Validation dataset
Require: $\mathbf{\Theta}_r$ ▷ Candidate retrieval model for hard negative sampling
Require: $\mathbf{\Theta}_p$ ▷ Document ranking model
Require: n_s ▷ Number of negative samples per epoch for each query

$M_p : \text{MAP}[Q, R^*] \leftarrow \{\}$
$M_n : \text{MAP}[Q, R^n] \leftarrow \{\}$
for each $(Q, R^*, R^n) \in S_{train}$ **do**
    $R^n \leftarrow \text{SORT_BY_KEY}(R^n, \mathbf{\Theta}_r(Q, R^n))$ ▷ Sort the non-relevant docs
    $M_p[Q] \leftarrow R^*$
    $M_n[Q] \leftarrow R^n$
end for
$\tilde{M}_n \leftarrow \text{COPY}(M_n)$ ▷ Make an identical copy of M_n for iteration
for each epoch **do**
    $S_e : \text{LIST}[(Q, R^*, R^n)] \leftarrow []$ ▷ Initialize the epoch training dataset
    for each $Q \in M_p$ **do**
        $E_p \leftarrow M_p[Q]$ ▷ Get the list of relevant docs
        for each $D \in E_p$ **do**
            $S_e \leftarrow S_e \cup (Q, D, \texttt{"true"})$
        end for
        $E_n \leftarrow M_n[Q]$ ▷ Get the sorted list of non-relevant docs
        $H_n \leftarrow \text{SLICE}(E_n, 0, n_s)$ ▷ Take the first n_s docs
        $E_n \leftarrow \text{SLICE}(E_n, n_s, |E_n|)$ ▷ Rotate the negative samples
        for each $D \in H_n$ **do**
            $S_e \leftarrow S_e \cup (Q, D, \texttt{"false"})$
        end for
        if $E_n = \emptyset$ **then**
            $M_n[Q] \leftarrow \tilde{M}_n[Q]$ ▷ Repopulate the list of non-relevant docs
        else
            $M_n[Q] \leftarrow E_n$ ▷ Update the list of non-relevant docs
        end if
    end for
    $\mathbf{\Theta}_p \leftarrow \text{TRAIN}(S_e, \mathbf{\Theta}_p)$ ▷ Train document ranking model
    $\mathbf{\Theta}_p^* \leftarrow \text{VALIDATION}(S_{val}, \mathbf{\Theta}_p)$ ▷ Select the best checkpoint
end for
return $\mathbf{\Theta}_p^*$

metric to assess the ranking performance of the checkpoints. The MRR metric is defined as follows:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{3.2}$$

where rank_i denotes the rank position of the first entailing document in the re-ranked document list of the i-th query and $|Q|$ denotes the total number of queries. We fine-tune the `castorini/monot5-3b-msmarco-10k` model with the cross-entropy objective for five epochs with an effective batch size of 64 and a learning rate of 5×10^{-5}.

After undergoing the fine-tuning process, the MonoT5 model operates as a document re-ranking module, refining the retrieval results obtained from the candidate retrieval stage and improving the relative order of the retrieved set of documents. For predicting entailment between documents, we employ a heuristic strategy based on thresholds. We designate the candidate with the highest score obtained from MonoT5 as the mandatory prediction. Additionally, we include other candidates in the prediction under the condition that their similarity score surpasses a threshold t and the disparity between their score and the highest score falls within a specified margin m. We select $t = 0.9$ and $m = 0.05$ to focus on the precision of the entailment predictions.

3.2.2 Zero-shot list-wise entailment prediction with LLMs

In the entailment prediction task, a straightforward approach is to frame it as a point-wise binary classification problem where LLMs are utilized to provide a binary response ("Yes" or "No") for query-document pairs. However, our previous experiments with prompt design following this approach resulted in notably low accuracy with a predominant bias towards "Yes" responses. Upon analyzing the responses and reasoning of the LLMs, we found that this outcome stems from the nature of the legal case entailment task, where the concept of "entailment" is rigorously defined. Since the candidate documents originate from a reference case that shares similar factual scenarios to the query's case, they often contain information related to the query. However, in the context of the legal case entailment task, such related information is insufficient. The entailing documents must provide comprehensive and direct support for the query to meet the strict criteria for entailment. This distinction is crucial, as mere relevance or partial alignment with the query does not

suffice to establish a true entailment relationship [46]. Due to the intrinsic language used in legal texts, LLMs often struggle to grasp this distinction and tend to assign "Yes" to candidate documents that merely exhibit information relevancy to the query. Consequently, an alternative approach is required to effectively utilize LLMs for legal case entailment.

Aligned with recent list-wise re-ranking strategies employing LLMs [53] [44] [34], in this approach, we tackle the legal case entailment problem using list-wise prediction strategy. In contrast to point-wise prediction approach, which assesses the relationship solely between the query and individual documents, the list-wise strategy takes into account the interaction and relevance across multiple candidate documents. By evaluating the entire set of candidate documents simultaneously, this approach can better discern the documents that offer comprehensive and direct support for the query. Moreover, this method helps mitigate the bias towards relevance alone, as the relative importance of each document is assessed within the context of the entire list. By performing list-wise prediction with LLMs, this approach leverages the comparative assessment capabilities of LLMs, enabling a more nuanced understanding of how different documents relate to the query and each other. As a result, the list-wise approach presents a more robust solution for legal case entailment, addressing the limitations observed with point-wise methods.

We transform the initial entailment problem into list-wise entailment prediction as follows: given a query Q and a list of candidate documents $D = [D_1, \ldots, D_k]$ with their respective IDs $I = [I_1, \ldots, I_k]$ from the preceding stage, the goal is to identify the document ID(s) that directly entail or provide support for the query. In the list-wise prediction framework, ColBERT-UOT functions as the initial retrieval method, MonoT5 serves as the high-performance top-k candidate retrieval, and the LLMs are employed as the entailment prediction model. For each query, we select the top-5 documents based on the ranking score derived from the fine-tuned MonoT5 model discussed in Section 3.2.1 to form the input for LLMs. Figure 3.2 demonstrates the input prompt for the entailment query presented in Table 1.1 with the top-5 candidates retrieved from MonoT5, and the Llama3 70B generation.

To compare the legal reasoning performance of LLMs across different sizes, we examine the performance of state-of-the-art, open-source LLMs in two parameter classes: approximately 7 billion parameters and over 70 billion parameters. For the 7 billion parameter class, we select Mistral 7B, Gemma

7B, and Llama3 8B. For the 70 billion parameter class, we choose Llama3 70B, Qwen1.5 72B, and Mistral 8x22B. We evaluate the instruction-finetuned variant of all selected LLMs.

Paragraphs:
ID P0034.txt: To conclude, it is my opinion after weighing these factors that, despite the inclusion in the Trade-marks Act of an untrammelled right of appeal and the right to adduce additional evidence, a considerable degree of deference is called for on the part of the appellate court when reviewing the Registrar's finding of confusion, provided at least that no significant new evidence has been adduced on a factual issue and it is not alleged that an error of law has been committed.
ID P0038.txt: If, on the other hand, the additional evidence goes beyond what was in substance already before the Registrar, then the court should ask whether, in the light of that material, the Registrar reached the wrong decision on the issue to which that evidence relates and , perhaps, on the ultimate decision as well. The more substantial the additional evidence, the closer the appellate court may come to making the finding of fact for itself.
ID P0037.txt: As for the impact on the standard of review of the filing of additional evidence on an issue at the appeal, much will depend on the extent to which the additional evidence has a probative significance that extends beyond the material that was before the Registrar. If it adds nothing of significance, but is merely repetitive of existing evidence without enhancing its cogency, its presence should not affect the standard of review applied by the court on the appeal.
ID P0024.txt: Indeed, even when additional evidence is admitted on appeal, it still may be appropriate to ask whether the Registrar's decision was wrong in the light of that evidence, rather than how the appellate court would have decided the question if it had been before the court de novo. Of course, the more significant the additional evidence the greater the scope for the court to exercise its own independent judgment on the finding of fact in question.
ID P0023.txt: The provision in s.56(5) allowing the parties to introduce additional evidence on the appeal may suggest a correctness standard on those findings of fact to which the evidence relates. However, it does not necessarily follow that the same standard should apply to the Registrar's findings on other facts.
Statement: A finding for which no additional evidence is filed should be accorded considerable deference.
Ascertain the legal paragraph(s) within the given set of paragraphs that best corresponds to, supports, or logically entails the presented legal statement. Only respond with the paragraph ID(s), do not say any word or explain.

Model response: P0034

Figure 3.2: An example of a legal entailment prediction prompt and the response from Llama3 70B.

Prediction-only & Single-answer

Paragraphs:
ID P{I_1}: {D_1}
...
ID P{I_k}: {D_k}

Statement: {Q}

Evaluate the list of legal paragraphs and determine strictly one paragraph that can be considered as entailing or providing valid support for the given legal statement. Only respond with the paragraph ID, do not say any word or explain.

Prediction-reasoning & Multiple-answers

Sift through the historical case law paragraphs and the recent case decision to establish notable legal connections.
Provided paragraphs:
ID P{I_1}: {D_1}
...
ID P{I_k}: {D_k}

Recent case decision: {Q}
Analyze each paragraph from the relevant case and determine which paragraph ID(s) entails the decision of the new case. Support your answer with logical and legal analysis.

Figure 3.3: Examples of the prompt templates used for zero-shot legal entailment prediction.

We conducted experiments on zero-shot list-wise entailment prediction using several different prompt templates. Given that the performance of LLMs is sensitive to prompt formatting [52], we utilize a total of 30 prompt designs, divided into two types of instruction: Prediction-only and Prediction-reasoning, with each type comprising 15 prompt designs. This list of designs consists of both manually crafted and AI-generated by prompting OpenAI's GPT-3.5 to generate additional designs based on a given example for the task of legal case entailment. Based on an observation from the training set indicating that the majority of legal case entailment scenarios involve only one entailing paragraph, we adopt two response modes: Single-answer and Multiple-answers. In the Single-answer mode, the LLMs are directed to provide an answer confined to a single paragraph that most accurately entails the query. This directive is omitted in the Multiple-answers mode. We apply both modes to every design, yielding a total of 60 prompt templates for benchmarking. Figure 3.3 illustrates examples of the prompt tem-

plates used for list-wise entailment prediction. The performance of the LLMs on each template is evaluated using the validation set, and the template demonstrating the highest performance is chosen for evaluation on the test sets. We select the top-5 candidate documents from MonoT5 to compose the input template and prompt the LLMs, ensuring a maximum context length of 5120 tokens. For reproducibility, we set the generation parameters `do_sample=False` and `temperature=0`. We use regular expressions to extract the reference paragraph IDs from the responses of LLMs for entailment prediction.

Chapter 4

Experimentation

4.1 Experiment settings

4.1.1 Datasets

To evaluate the performance of the proposed system, we utilize the legal case entailment datasets from Task 2 of the COLIEE competition from 2020 to 2024 as our benchmark datasets. The legal case documents for this task are selected from an existing collection that primarily comprises Federal Court of Canada case law. Each dataset comprises legal case entailment samples indexed with a three-digit ID starting from "001". Each sample includes a query case document, an entailed paragraph from the query case containing the decision, and a list of candidate paragraphs from a reference legal case. The objective of the task is to identify the paragraph(s) from the provided list that entails the decision of the query legal case. The datasets are divided into two splits: a training set and a test set, with the latter consisting of the last 100 case IDs. To create a validation set, we select the first 100 case IDs from the original training set, leaving the remaining samples as the training data. The performance of the methods is reported on the official test set for each year.

Figure 4.1 and 4.2 depict the distribution of the lengths of the entailed and reference paragraphs in the 2024 dataset, respectively. The entailed paragraphs are relatively short, with lengths ranging from 10 to 140 words and an average length of 38 words. In contrast, the reference paragraphs exhibit a long-tail distribution, with lengths ranging from 20 to more than 1000 words. The majority of reference paragraphs fall within the 20 to 200-

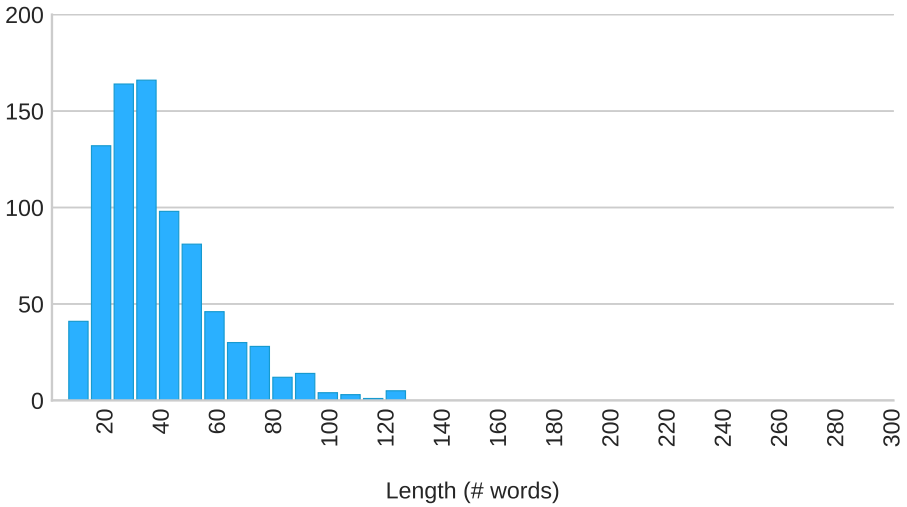

Figure 4.1: Distribution of the lengths of the entailed paragraphs in COLIEE 2024 dataset.

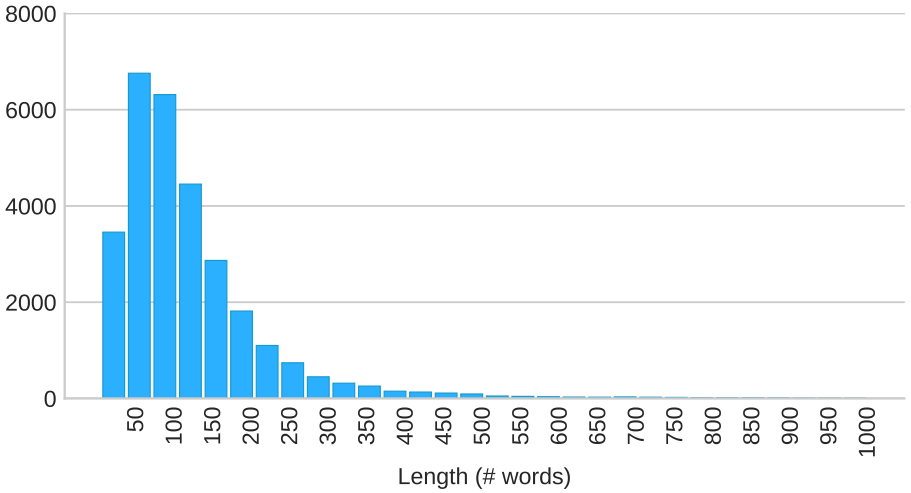

Figure 4.2: Distribution of the lengths of the reference paragraphs in COLIEE 2024 dataset.

word range, with an average length of 120 words. On further analysis, we obtained the insights that long reference paragraphs often include quotations, citations from the courts or defendants, or references to other paragraphs or

articles. To handle these long paragraphs, we established a maximum length of 400 words and we retained only the last 400 words as the input reference text.

Table 4.1: Statistics of the legal case entailment datasets.

Dataset	2020	2021	2022	2023	2024
# samples	425	525	625	725	825
# reference paragraphs	15216	18740	22018	25783	29434
• min	3	7	6	5	7
• max	242	242	138	170	315
• median	38	30	41	32	43
# entailing paragraphs	499	616	734	854	1001
• min	1	1	1	1	1
• max	4	5	5	5	5
• median	1	1	1	1	1

Table 4.2: Percentages of number of entailment per query across datasets.

# entailment	1	2	3	4	5	≤ 2
2020	84.94	13.41	0.94	0.71	-	98.35
2021	85.33	12.95	0.95	0.57	0.19	98.28
2022	84.96	13.44	0.96	0.48	0.16	98.40
2023	85.10	12.83	1.38	0.55	0.14	97.93
2024	82.67	14.30	2.18	0.72	0.12	96.97

Table 4.1 shows the statistics and Table 4.2 shows the entailment percentage of each dataset. The datasets exhibit similar characteristics, with each dataset containing approximately 35 reference paragraphs per sample and an average of 1.2 entailing paragraphs per sample. The number of reference paragraphs fluctuates significantly across samples, ranging from fewer than 10 to over 100, with the median number of reference paragraphs falling between 30 and 40. Similarly, the number of entailing paragraphs differs across samples, with most samples containing 1 or 2 entailing paragraphs and the majority having only 1. These variations underscore the challenge of accurately identifying the entailing paragraphs within the larger set of variable-length reference paragraphs.

4.1.2 Candidate Retrieval

To assess the performance of the candidate retrieval methods, we utilize the Recall@K metric to measure the proportion of entailing documents retrieved among the top-K results. The Recall@K metric is defined as:

$$\text{Recall@}K = \frac{(\#\text{ of retrieved entailing documents in top } K \text{ for all queries})}{(\#\text{ of entailing documents for all queries})} \tag{4.1}$$

We select $K = 20$ to evaluate the entailment retrieval performance of the methods and to highlight the cost-effectiveness of the candidate retrieval stage. We evaluate the performance of the proposed ColBERT-UOT retrieval method against the following established baselines:

- Sparse text representation: This method class represents text documents as sparse vectors where each dimension corresponds to a unique term in the vocabulary. These methods are effective in capturing the basic lexical information but may discard the semantic relationships between words or the context of the documents. For this class, we deploy the following widely-adopted methods as baselines:
  - BM25: the baseline bag-of-words retrieval method. We use the implementation of BM25 provided by the Pyserini toolkit [33].
  - SPALDE++ ED [18]: this state-of-the-art sparse retrieval method uses BERT WordPiece vocabulary vectors to represent documents, with each term's importance determined by logits from a Masked Language Model. We use the `naver/splade-cocondenser-ensembledistil` model for evaluation.
- Dense text embeddings: this method class leverages pre-trained language models to represent text documents as dense, continuous embedding vectors in a high-dimensional space. The embeddings generated by neural models capture contextual information and preserve semantic similarity between words and documents. For this class, we use the following baseline methods:
  - Contriever [21]: this work proposes an unsupervised training framework with text augmentation and sampling strategies for doc-

ument retrieval. We use the `facebook/contriever` checkpoint from the thesis.

- E5 Mistral [57]: this work proposes fine-tuning the Mistral 7B architecture on synthetically generated query-positive-negative triplets to produce document embeddings. We use the `intfloat/e5-mistral-7b-instruct` model for evaluation.
- BGE Large [58]: this work introduces a family of general-purpose embeddings models that are pre-trained on massive text corpora and subsequently fine-tuned through multi-task learning. We use the best-performing `BAAI/bge-large-en-v1.5` model.
- ColBERT-MaxSim [51]: we utilize the original ColBERT architecture with the MaxSim alignment function for direct comparison with the proposed ColBERT-UOT. We benchmark two versions: one employing the original architecture and one incorporating an additional stop-word removal operator before the MaxSim alignment function, denoted as ColBERT-MaxSim^F. We use the `colbert-ir/colbertv2.0` for evaluation.

All the aforementioned neural retrieval models have been extensively trained on large-scale information retrieval datasets, demonstrating notable zero-shot performance across various domains [4] [24]. For this reason, we utilize the original model checkpoint without additional training on legal entailment datasets. For ColBERT-UOT, we use the same BERT model checkpoint with ColBERT-MaxSim for a fair comparison.

4.1.3 Entailment Prediction

Following the COLIEE competition, we use the Micro-F1 as the main evaluation metric and Recall@K with $K = 5$ to access the entailment retrieval performance. The Micro-F1 is defined as:

$$\begin{aligned} \text{Precision} &= \frac{(\#\text{ of retrieved entailing documents for all queries})}{(\#\text{ of retrieved documents for all queries})} \\ \text{Recall} &= \frac{(\#\text{ of retrieved entailing documents for all queries})}{(\#\text{ of entailing documents for all queries})} \\ \text{Micro-F1} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{4.2}$$

We take the metric of the best system for each competition year as the baseline metric for comparison. The following is a brief description of the best system for each year:

- Ensemble of fine-tuned BERT + BM25 (Best of COLIEE 2020 [56]): this work proposes to train a BERT-base architecture on text-pair classification using the COLIEE dataset. An ensemble approach is utilized to integrate the BERT score and BM25 for determining the entailment score.

- Ensemble of MonoT5 + DeBERTa (Best of COLIEE 2021 [49]): the best system constituent of the zero-shot prediction of MonoT5-large and DeBERTa model. Complex heuristic rules were applied to process the concatenated predictions from the two models to form the prediction.

- Ensemble of MonoT5 models (Best of COLIEE 2022 [48]): similar to [49], this work deployed a zero-shot ensemble of MonoT5-3B and MonoT5-base variants. The same heuristic strategy in [49] is used for entailment prediction.

- Fine-tuned MonoT5-large with hard negative sampling (Best of COLIEE 2023 [38]): this work proposes fine-tuning the MonoT5-large variant on the COLIEE legal entailment dataset using a two-phase training approach incorporating hard negative mining strategies. Threshold-based heuristics were subsequently employed to extract entailing documents based on the scores obtained from the fine-tuned model.

- Fine-tuned MonoT5-3B with hard negative sampling (Best of COLIEE 2024 [41]): following the approach of [38], this work fine-tuned a variant of MonoT5-3B using hard negative sampling and applied ratio-based heuristics to form the prediction.

Together with the baselines, we report the performance of zero-shot listwise entailment prediction with LLMs for legal case entailment following the setting described in 3.2.2.

4.2 Experiment results

4.2.1 Candidate Retrieval

Table 4.3 presents the performance of the retrievers for each dataset, as well as their overall performance. The proposed ColBERT-UOT architecture demonstrates competitive performance relative to the baselines, achieving the highest Recall@20 on 3 out of 5 datasets and exhibiting superior entailment coverage on average. In comparison to the ColBERT-MaxSim baseline, our approach integrating the UOT framework achieves an average improvement of approximately 2%, underscoring the efficacy of the sparse keyword alignment strategy. The ColBERT-MaxSim^F variant, which incorporates the stop-word removal operator, consistently demonstrates a recall metric that is equal to or superior to the original version. This highlights the negative impact of noise alignment links on the retrieval results. We conduct further analysis and compare the difference in the word alignments produced by ColBERT-MaxSim variants and ColBERT-UOT in section 5.2. The BM25 baseline, despite its simplicity, still proves to be a strong baseline for text retrieval as it attains good performance compared to the best method on all datasets, even outperforming the SPLADE++ ED sparse representation counterpart and the BGE Large embeddings model. This demonstrates the robustness of traditional lexical retrieval methods in challenging domains such as legal text retrieval. The best dense text representation baseline is E5 Mistral, which leverages the power of LLMs for document embeddings, demonstrating the potential of LLMs for information retrieval. Compared to E5 Mistral, ColBERT-UOT achieves marginal improvements in Recall@20 while benefiting from a significantly smaller model size and inherent interpretability through the word alignment results. Overall, ColBERT-UOT exhibits improved retrieval performance compared to the original architecture, demonstrating robust results in legal document retrieval through the implementation of the sparse keyword alignment strategy.

4.2.2 Entailment Prediction

Table 4.4, 4.5, and 4.6 show the main results of the legal case entailment task. The proposed system with the ColBERT-UOT retrieval model and MonoT5 entailment prediction model achieves the best F1 metric and surpasses the best system at each COLIEE competition by 2 - 7 points on F1

Table 4.3: Performance of the retrievers in COLIEE datasets, and the average performance.

Methods	2020	2021	2022	2023	2024	Average
BM25	98.40	97.44	98.30	97.50	91.84	96.70
SPLADE++ ED	96.00	98.29	96.61	98.33	93.20	96.49
Contriever	96.80	100	99.15	97.50	93.20	97.33
E5 Mistral	97.60	97.44	100	99.17	95.92	<u>98.03</u>
BGE Large	95.20	98.29	99.15	98.33	97.28	97.65
ColBERT-MaxSim	95.20	98.29	97.46	99.17	94.56	96.94
ColBERT-MaxSim^F	96.80	98.29	98.03	99.17	95.91	97.64
ColBERT-UOT	99.20	98.29	100	99.17	96.60	**98.65**

across datasets. Our system showcases good retrieval performance by the ability to locate around 90% of the entailing documents in the top-5 candidates as evidenced in the Recall@5 metric. Compared to the system using only the MonoT5 model (with top-$k = |R|$), the system incorporating the ColBERT-UOT candidate retrieval model achieves comparable and even superior performance on the COLIEE 2024 dataset. Additionally, the system employing ColBERT-UOT surpasses the system utilizing the BM25 retriever, averaging approximately 1 point higher in both F1 and Recall@5 metrics. This demonstrates the effectiveness of an efficient candidate retrieval model in significantly reducing computation during the entailment prediction stage while still maintaining high system performance.

The performance of zero-shot entailment prediction with LLMs varies depending on model size, with the 70B class demonstrating superior performance compared to the 7B counterpart and competing closely with the baseline value. The 7B-class LLMs underperformed by a large margin to the baseline on all datasets. The Llama3 8B achieves superior metrics within its category, surpassing both the Mistral 7B and the Gemma 7B by average margins of 10 and 4 points on F1, respectively. However, the zero-shot performance of Llama3 8B falls behind the baseline value by 3 to 10 points on F1. The 70B-class LLMs, with increased capacity and computation power, perform significantly better than the 7B class as expected and achieve competitive performance with the baseline across datasets. The Llama3 70B achieves the highest metrics of all LLMs on average, lagging by 1-3 points with the baseline on 3 datasets and outperforming the baseline in the COL-

Table 4.4: Performance of the entailment predictors in COLIEE 2020 and 2021 datasets

Methods	Retriever		2020		2021	
	source	top-k	Recall@5	F1	Recall@5	F1
Baseline	-	-	-	67.53	-	69.15
MonoT5 3B	-	$\|R\|$	88.80	68.97	96.58	74.89
	BM25	20	88.00	68.37	94.87	75.77
	ColBERT-UOT	20	88.80	**68.97**	95.72	**75.77**
Mistral 7B			43.20	48.00	47.86	51.61
Gemma 7B			52.00	58.04	60.68	65.44
Llama3 8B			54.40	60.44	61.54	66.36
Llama3 70B			59.20	66.07	70.08	75.57
Qwen1.5 72B			59.20	66.07	66.67	71.89
Mistral 8x22B	MonoT5 3B	5	60.00	66.67	62.39	67.28

IEE 2021 and 2022 datasets. Despite containing an additional 2 billion parameters, Qwen1.5 72B exhibits inferior performance compared to Llama3 70B, with an average gap of approximately 3 points observed in both Recall and F1 metrics. The Mistral 8x22B, the most extensive LLM within our benchmark, demonstrates comparable performance to the Llama3 70B with their average metrics differing by approximately 1 F1 point.

In comparison to the fine-tuned MonoT5 model, the zero-shot entailment prediction of pre-trained LLMs exhibits inferior performance. The 7B class of LLMs demonstrates a significant performance gap relative to MonoT5 3B, while the top-performing LLM, Llama3 70B, trails by around 3.5 points in F1 score on average. This observation highlights the challenge that pre-trained LLMs encounter when addressing legal document processing tasks like legal case entailment, which demand a deep understanding of nuanced language and intricate relationships to make accurate assessments [37]. Our experiment results are consistent with [22], where the LegalBERT language model demonstrates superior performance compared to significantly larger pre-trained LLMs on the LEDGAR subset of the LexGLUE benchmark [9]. Based on these findings, it is recommended that additional training specific to the legal domain be undertaken to improve the performance of LLMs on legal tasks.

Notably, for all LLMs, the prompt utilized for zero-shot prediction that

Table 4.5: Performance of the entailment predictors in COLIEE 2022 and 2023 datasets

Methods	Retriever		2022		2023	
	source	top-k	Recall@5	F1	Recall@5	F1
Baseline	-	-	-	67.83	-	74.56
MonoT5 3B	-	$\|R\|$	95.76	74.67	94.17	77.13
	BM25	20	95.76	73.68	91.67	76.44
	ColBERT-UOT	20	95.76	**74.67**	94.17	**77.13**
Mistral 7B			44.91	48.62	55.00	60.00
Gemma 7B			53.39	57.79	56.67	61.82
Llama3 8B			58.47	63.30	58.33	63.63
Llama3 70B			66.10	71.89	65.83	71.82
Qwen1.5 72B			61.02	66.05	63.33	69.09
Mistral 8x22B	MonoT5 3B	5	66.95	72.48	65.00	70.90

yields the highest F1 metric on the validation set adheres to the Prediction-only and Single-answer mode. In our experiments, limiting the answer to a single paragraph ID leads to a decrease in Recall but a significant improvement in Precision metric, resulting in an enhancement of the F1 metric on the validation set. We further investigate the zero-shot performance of LLMs on legal case entailment in Section 5.3.

Table 4.6: Performance of the entailment predictors in COLIEE 2024 dataset, and the average performance.

Methods	Retriever		2024		**Average**	
	source	top-k	Recall@5	F1	Recall@5	F1
Baseline	-	-	-	65.12	-	-
MonoT5 3B	-	$\|R\|$	87.75	68.84	92.61	72.90
	BM25	20	87.71	67.15	91.60	72.28
	ColBERT-UOT	20	89.80	**69.82**	92.85	**73.27**
Mistral 7B	MonoT5 3B	5	44.22	52.63	47.03	52.17
Gemma 7B			40.82	48.58	52.71	58.33
Llama 8B			48.29	57.49	56.21	62.24
Llama3 70B			53.06	63.16	62.85	69.70
Qwen1.5 72B			50.34	59.92	60.11	66.60
Mistral 8x22B			53.74	63.97	61.62	68.26

Chapter 5

Analysis

5.1 The effectiveness of the candidate retrieval stage

In our framework, the candidate retrieval stages aim to reduce the computational demands of the entailment prediction stage, where we utilize resource-intensive LLMs. In the design of the retrieval module, we aim to minimize the number of reference paragraphs and the number of tokens that need to be processed to predict the entailing paragraphs, without compromising the entailment prediction capability of the system. In our system pipeline, we select the top-k documents retrieved by ColBERT-UOT, with $k = 20$ selected based on the entailment coverage observed on the validation set, for further processing in the subsequent stage. Figure 5.1 and 5.2 illustrate the impact of the ColBERT-UOT retrieval model, demonstrating that the proposed method reduces both the number of paragraphs and the number of tokens to be processed by approximately 50%. This significant reduction in processing requirements highlights the efficiency of our retrieval approach.
The experimental results presented in Section 4.2.2 further show that this reduction does not adversely affect the performance of the MonoT5 model. By pre-filtering the irrelevant reference paragraph, the retrieval module can even enhance the performance of the entailment predictor, particularly in the COLIEE 2024 dataset. These observations underscore the robustness and effectiveness of the candidate retrieval stage, suggesting that this approach not only maintains but also improves the accuracy of legal case entailment prediction while substantially decreasing computational overhead. This efficiency

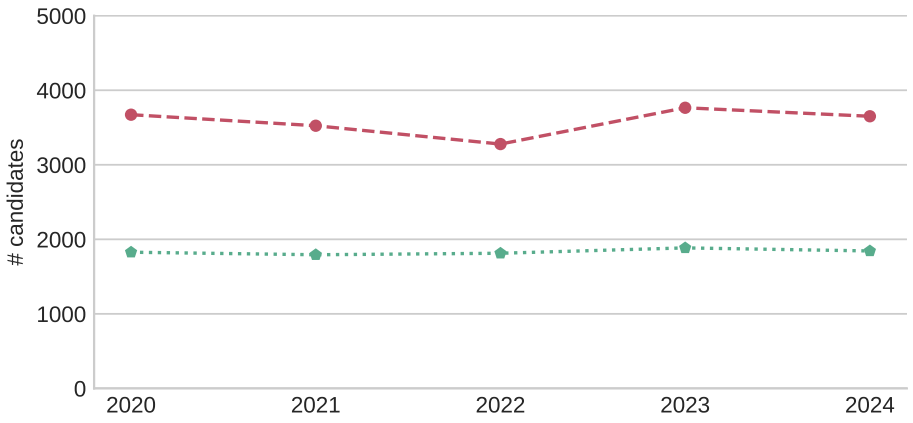

Figure 5.1: Number of candidate paragraphs in the COLIEE test sets. The red line represents the original count, and the green line indicates the number of retrieved candidates for the entailment prediction stage.

gain is crucial for practical applications, where processing large volumes of legal text efficiently and accurately is a key requirement for real-world legal information retrieval tasks.

5.2 Improved alignment led to improved retrieval performance

In this section, we analyze and compare the alignment characteristics of the MaxSim, MaxSim^F, and UOT alignment method to gain a deeper understanding of the performance enhancement achieved through the application of the UOT framework. To analyze the performance of ColBERT with different alignment strategies, we prioritize two critical factors: the sparsity and the quality of the alignment links. The sparsity of alignment links reflects the method's ability to pinpoint relevant connections while disregarding irrelevant links when calculating the similarity between the query and the document. On the other hand, the quality of the alignment links pertains to the relevance of the keywords that are being aligned. High-quality alignment links ensure that semantically and contextually similar keywords between the query and document are aligned, thereby enhancing retrieval results. Together, these factors determine the efficacy of the alignment method in

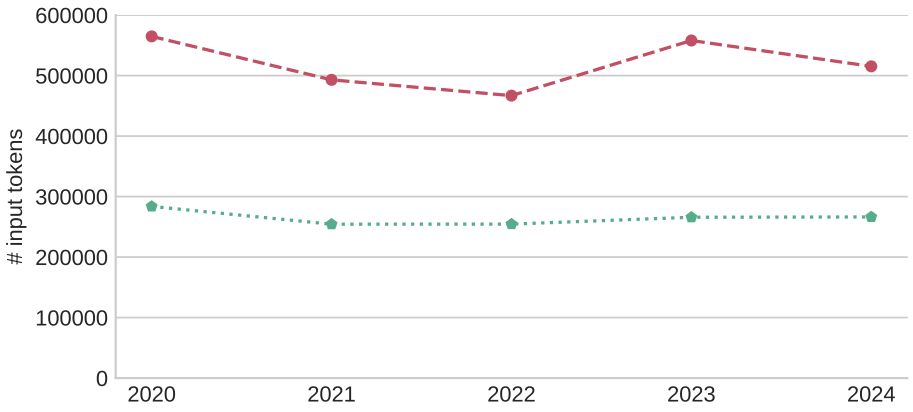

Figure 5.2: Number of input tokens in the COLIEE test sets. The red line denotes the original count, and the green line represents the number of tokens after the candidate retrieval stage.

improving the retrieval performance of the system.

Table 5.1: Statistics on the number of alignment links produced by the alignment functions in the COLIEE 2024 dataset.

Method	Total	link-1	link-2	link-3	link-($\geq$ 4)
MaxSim	748735	86.9%	9.8%	2.5%	0.7%
MaxSim^F	405893	91.7%	6.3%	1.5%	0.5%
UOT	195319	78.5%	20.4%	0.9%	0.2%

In Table 5.1, we present the statistics on the number of alignment links produced by each alignment function. In the context of word alignment, an alignment link-n represents a matching of a word from the query to n words from the document. The MaxSim alignment approach, operating on the original text embeddings, produces the highest number of alignments as expected with with most being link-1 i.e. one-to-one word alignments. We note that despite the one-to-one alignment characteristic of the MaxSim function, it can generate one-to-many alignments in the subword alignment setting for query words composed of multiple subwords, as each subword can link to different words in the document. The MaxSim^F variant, which incorporates stop-word removal, produces significantly fewer alignments by eliminating noisy alignments of stop words. This reduction in noise leads to

improved retrieval performance as demonstrated in Section 4.2.1. The UOT approach generates the most sparse alignment matrices, exhibiting approximately four times fewer alignments compared to the MaxSim function and twice fewer alignments compared to the MaxSim^F function. This illustrates the sparsity feature of the proposed alignment method. Furthermore, the UOT framework promotes one-to-many alignments of semantically and contextually similar words, with the number of link-(≥ 2) alignments comprising more than 20% of the total alignment links. This percentage is twice that observed in the MaxSim alignment function.

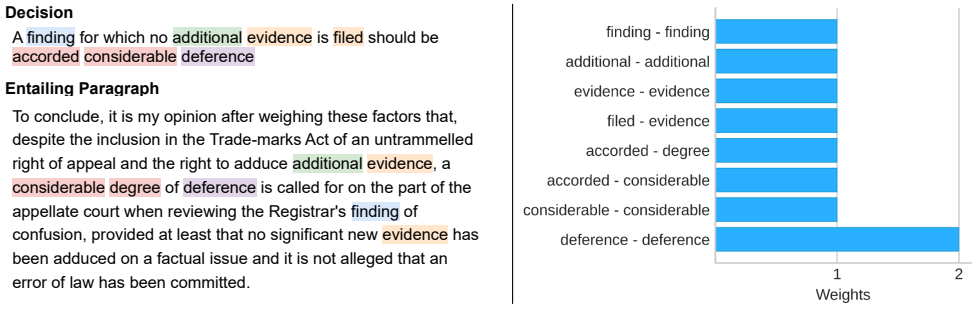

Figure 5.3: Visualization of keyword matching using the MaxSim^F alignment function.

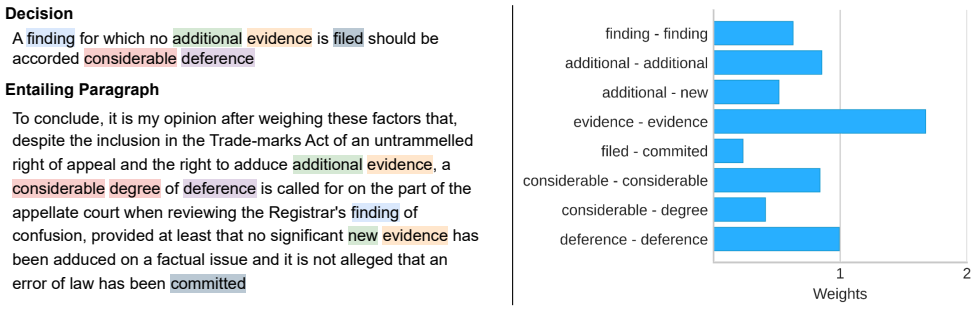

Figure 5.4: Visualization of keyword matching using the UOT alignment method.

We illustrate the word alignment results of the legal case entailment scenario presented in Table 1.1 using the MaxSim^F and the UOT alignment strategy in Figure 5.3 and 5.4, respectively. In each figure, the left

side visualizes the alignment links represented by color, while the right side shows the weight of each alignment link. From the alignment visualization of MaxSim^F, we observe three instances of incoherent alignments: "filed - evidence", "accorded - degree", and "accorded - considerable", with the latter two alignments arising from the tokens "accord" and "##ed" of the word "accorded.". This is attributed to the forced alignment characteristic of the MaxSim function, resulting in noisy alignment links. Moreover, in the MaxSim strategy, all alignment links are uniformly weighted. In instances where words are composed of multiple subwords, such as "deference" being composed of "def" and "##erence" tokens in this example, we aggregate the weights of their subword alignment links. This uniform weighting scheme results in noisy alignment links contributing equally to the similarity score calculation, thereby adversely affecting retrieval performance. Figure 5.4 demonstrates how the UOT framework addresses these issues. In contrast to MaxSim, the UOT alignment strategy bypasses the alignment of the word "accorded" and introduces a contextually appropriate link, "additional - new," resulting in alignments that are more coherent than those generated by the MaxSim function. While the alignment link "filed - committed" lacks semantic suitability, the UOT framework assigns it a relatively low weight, thereby minimizing its impact on the similarity score. Notably, the UOT framework assigns increased weight to the identical alignment link "evidence - evidence" due to the word "evidence" appearing twice in the document. These differences in alignment, compared to MaxSim, result from the unbalanced mass transportation feature of the UOT framework, leading to an uneven distribution of alignment weights. This characteristic proves beneficial when combined with alignment weight thresholding to eliminate irrelevant alignment links or in situations where important keywords occur multiple times in the document. Furthermore, by applying subword mass normalization, we equalize the potential contribution of alignments involving words composed of multiple subwords, thereby positively impacting the stability of the similarity score. In summary, compared to MaxSim, the UOT alignment strategy generates sparser and higher-quality alignments, resulting in an overall improvement in retrieval performance.

5.3 Zero-shot LLMs achieve promising results in legal case entailment

In this section, we seek to gain a deeper understanding of the performance of pre-trained LLMs in the context of legal case entailment. We select the two best-performance LLMs on our benchmark, Llama3 70B and Mistral 8x22B to perform further performance analysis. To analyze how the LLMs performance varies with different prompt designs, we divide the 60 prompt templates into 4 categories: Prediction-only & Single-answer (PO & SA), Prediction-only & Multiple-answers (PO & MA), Prediction-reason & Single-answer (PR & SA), and Prediction-reason & Multiple-answers (PR & MA) (detailed in Section 3.2.2, and Figure 3.3), with each category comprises 15 prompt templates. For each category, we select the prompt template that achieves the highest F1 score on the validation set for evaluation on the test sets.

Table 5.2: Performance of LLMs across 4 prompt design categories in the COLIEE 2023 test dataset.

Model	PO & SA		PO & MA		PR & SA		PR & MA	
	Recall	F1	Recall	F1	Recall	F1	Recall	F1
Llama3 70B	65.83	**71.82**	82.50	68.28	63.33	<u>69.09</u>	81.67	59.94
Mistral 8x22B	65.00	**70.90**	86.67	53.06	55.83	<u>60.91</u>	91.67	45.74

Table 5.3: Performance of LLMs across 4 prompt design categories in the COLIEE 2024 test dataset.

Model	PO & SA		PO & MA		PR & SA		PR & MA	
	Recall	F1	Recall	F1	Recall	F1	Recall	F1
Llama3 70B	53.06	63.16	70.75	**66.45**	46.94	55.87	76.19	<u>63.64</u>
Mistral 8x22B	53.74	**63.97**	81.63	<u>53.81</u>	41.49	49.39	84.35	47.97

We show the Recall and F1 metric for each category in the COLIEE test sets of COLIEE 2023 and 2024 in Table 5.2 and 5.3, respectively. In the COLIEE 2023 test set, the SA mode outperformed the Multi-answers mode, with the PO & SA prompt design achieving the highest F1 score with both LLMs. The Llama3 70B outperforms the Mistral 8x22B in the PO & SA category and maintains consistent metrics across the PO & SA, PO & MA, and

PR & SA categories, while having a significant decrease in performance in the PR & MA category. On the other hand, The Mistral 8x22B demonstrates good performance solely in the PO & SA category and displays subpar metrics in all other categories. This observation holds in the COLIEE 2024 test set, where the Mistral LLM slightly outperforms the Llama3 counterpart in the PO & SA categories but performs poorly in all other prompt categories. Unlike the 2023 test set, the Llama3 LLM achieves the highest F1 metric in the PO & MA category for the 2024 test set due to the latter containing more legal case queries with multiple entailing paragraphs. Notably, in the MA mode, the Mistral LLM achieves higher recall but significantly lower F1 scores compared to the Llama3 70B LLM in both test datasets. This suggests that for legal case entailment, Llama3 70B is generally more effective when high-quality, consistent outputs are required, whereas Mistral 8x22B is better suited for scenarios prioritizing broader entailment coverage and retrieval.

Overall, despite the smaller parameter count, Llama3 70B consistently matches or outperforms Mistral 8x22B in F1 scores across all prompt design categories, indicating a more reliable and consistent capability in entailment prediction. Our observation aligns with the findings from LegalBench[1], a benchmark designed to evaluate the performance of zero-shot LLMs across a wide range of legal tasks. According to LegalBench, the Llama3 70B achieves the highest overall performance among open-source LLMs in legal-related tasks, with the performance on par with close-sourced trillion-parameter LLMs like Claude 3 Opus and GPT4. This phenomenon could be attributed to Llama3 variants being pre-trained on an extensive range of legal domain corpora and potentially fine-tuned on tasks related to legal document processing. It's worth noting that the possibility of Llama3 variants being fine-tuned on the COLIEE 2024 test set is unlikely, as the LLama3 models were released shortly after the release of the COLIEE 2024 datasets. In summary, compared to the methods discussed in Section 4.2.2, the zero-shot Llama3 LLM achieves competitive performance, demonstrating its promising potential for application in legal case entailment scenarios.

[1] www.vals.ai/legalbench

Chapter 6

Conclusion

In this thesis, we investigate the task of legal case entailment and present a general two-stage framework with a focus on legal entailment retrieval. By formulating the original task as a document retrieval problem, our approach leverages state-of-the-art language models in the information retrieval domain to efficiently identify entailment relationships between legal cases. The ColBERT-UOT retrieval architecture, employing a sparse keyword alignment strategy based on the UOT framework, demonstrates enhanced performance in retrieving legal documents compared to the original ColBERT design. Our extensive evaluation using the COLIEE datasets shows that the proposed system comprising ColBERT-UOT and MonoT5 achieves substantial performance enhancements over baseline methods. Furthermore, our benchmarking study reveals the potential of large language models in legal case entailment, despite their sensitivity to prompt formulation. The findings of this research have significant implications for the legal domain, presenting a direct solution to accelerate the time-consuming process of detecting entailment between legal cases. Given the promising zero-shot results of pre-trained LLMs, a promising direction for future work is the development of specialized LLMs tailored to the legal domain. We anticipate that these LLMs, enriched with extensive legal knowledge from vast legal corpora, will have significant potential to assist legal professionals and accelerate the analysis of legal documents.

Bibliography

[1] Sawsan Alqahtani et al. "Using Optimal Transport as Alignment Objective for fine-tuning Multilingual Contextualized Embeddings". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3904–3919. DOI: 10.18653/v1/2021.findings-emnlp.329. URL: https://aclanthology.org/2021.findings-emnlp.329.

[2] Yuki Arase, Han Bao, and Sho Yokoi. "Unbalanced Optimal Transport for Unbalanced Word Alignment". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 3966–3986. DOI: 10.18653/v1/2023.acl-long.219. URL: https://aclanthology.org/2023.acl-long.219.

[3] Payal Bajaj et al. "Ms marco: A human generated machine reading comprehension dataset". In: *arXiv preprint arXiv:1611.09268* (2016).

[4] Payal Bajaj et al. *MS MARCO: A Human Generated MAchine Reading COmprehension Dataset*. 2018. arXiv: 1611.09268 [cs.CL].

[5] Nicolas Bonneel et al. "Displacement interpolation using Lagrangian mass transport". In: *ACM Trans. Graph.* 30.6 (2011), 1–12. ISSN: 0730-0301. DOI: 10.1145/2070781.2024192. URL: https://doi.org/10.1145/2070781.2024192.

[6] M.Q. Bui, D.T. Do, N.K. Le, et al. "Data Augmentation and Large Language Model for Legal Case Retrieval and Entailment". In: *Review of Socionetwork Strategies* 18 (2024), pp. 49–74. DOI: 10.1007/s12626-024-00158-2.

[7] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment Problems*. Vol. 106. Revised reprint. SIAM, 2012.

[8] Ilias Chalkidis et al. "LEGAL-BERT: The Muppets straight out of Law School". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2898–2904. DOI: 10.18653/v1/2020.findings-emnlp.261.

[9] Ilias Chalkidis et al. "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4310–4330. DOI: 10.18653/v1/2022.acl-long.297. URL: https://aclanthology.org/2022.acl-long.297.

[10] Yupeng Chang et al. "A Survey on Evaluation of Large Language Models". In: *arXiv preprint arXiv:2307.03109* (2023).

[11] Yupeng Chang et al. "A Survey on Evaluation of Large Language Models". In: *ACM Trans. Intell. Syst. Technol.* 15.3 (2024). ISSN: 2157-6904. DOI: 10.1145/3641289. URL: https://doi.org/10.1145/3641289.

[12] Yimeng Chen et al. "Evaluating natural language generation via unbalanced optimal transport". In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2020, pp. 3730–3736.

[13] Ying Chen et al. "Legal Information Retrieval by Association Rules". In: *Twelfth International Workshop on Juris-informatics (JURISIN)*. 2018.

[14] Lenaic Chizat et al. "Scaling Algorithms for Unbalanced Transport Problems". In: *Mathematics of Computation* 87 (July 2016). DOI: 10.1090/mcom/3303.

[15] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 2292–2300.

[16] Rohan Debbarma et al. "IITDLI: Legal Case Retrieval Based on Lexical Models". In: *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL)*. 2023.

[17] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[18] Thibault Formal et al. "From Distillation to Hard Negative Sampling: Making Sparse Neural IR Models More Effective". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain: Association for Computing Machinery, 2022, 2353–2359.

[19] R. Goebel et al. "Overview of Benchmark Datasets and Methods for the Legal Information Extraction/Entailment Competition (COLIEE) 2024". In: *New Frontiers in Artificial Intelligence. JSAI-isAI 2024*. Ed. by T. Suzumura and M. Bono. Vol. 14741. Lecture Notes in Computer Science. Singapore: Springer, 2024. DOI: 10.1007/978-981-97-3076-6_8. URL: https://doi.org/10.1007/978-981-97-3076-6_8.

[20] Zhiqi Huang, Puxuan Yu, and James Allan. "Improving Cross-lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation". In: *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. WSDM '23. , Singapore, Singapore, Association for Computing Machinery, 2023, 1048–1056. ISBN: 9781450394079. DOI: 10.1145/3539597.3570468. URL: https://doi.org/10.1145/3539597.3570468.

[21] Gautier Izacard et al. *Unsupervised Dense Information Retrieval with Contrastive Learning*. 2021. DOI: 10.48550/ARXIV.2112.09118. URL: https://arxiv.org/abs/2112.09118.

[22] Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. "Large Language Models are legal but they are not: Making the case for a powerful LegalLLM". In: *Proceedings of the Natural Legal Language Processing Workshop 2023*. Ed. by Daniel Preoțiuc-Pietro et al. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 223–229. DOI: 10.18653/v1/2023.nllp-1.22. URL: https://aclanthology.org/2023.nllp-1.22.

[23] Rabelo Juliano et al. “Legal Information Extraction and Entailment for Statute Law and Case Law”. In: *Twelfth International Workshop on Juris-informatics (JURISIN)*. 2018.

[24] Ehsan Kamalloo et al. *Resources for Brewing BEIR: Reproducible Reference Models and an Official Leaderboard*. 2023. arXiv: 2306.07471 [cs.IR].

[25] Leonid Kantorovich. “On the Transfer of Masses (in Russian)”. In: *Doklady Akademii Nauk* 37 (1942), pp. 227–229.

[26] Daniel Martin Katz et al. “Natural Language Processing in the Legal Domain”. In: *ArXiv* abs/2302.12039 (2023). URL: https://api.semanticscholar.org/CorpusID:256440319.

[27] Omar Khattab and Matei Zaharia. “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*. SIGIR ’20. Virtual Event, China: Association for Computing Machinery, 2020, 39–48. ISBN: 9781450380164. DOI: 10.1145/3397271.3401075. URL: https://doi.org/10.1145/3397271.3401075.

[28] M.Y. Kim, J. Rabelo, K. Okeke, et al. “Legal Information Retrieval and Entailment Based on BM25, Transformer and Semantic Thesaurus Methods”. In: *Review of Socionetwork Strategies* 16 (2022), pp. 157–174.

[29] Philip A. Knight. “The Sinkhorn–Knopp Algorithm: Convergence and Applications”. In: *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pp. 261–275. DOI: 10.1137/060659624. eprint: https://doi.org/10.1137/060659624. URL: https://doi.org/10.1137/060659624.

[30] Matt Kusner et al. “From Word Embeddings to Document Distances”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 957–966.

[31] Seonghyeon Lee et al. “Toward Interpretable Semantic Textual Similarity via Optimal Transport-based Contrastive Sentence Learning”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Mure-

san, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5969–5979. DOI: 10.18653/v1/2022.acl-long.412. URL: https://aclanthology.org/2022.acl-long.412.

[32] Percy Liang et al. *Holistic Evaluation of Language Models*. 2023. arXiv: 2211.09110.

[33] Jimmy Lin et al. "Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations". In: *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2021, pp. 2356–2362.

[34] Xueguang Ma et al. *Zero-Shot Listwise Document Reranking with a Large Language Model*. 2023. arXiv: 2305.02156.

[35] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences* (1781), pp. 666–704.

[36] Sheshera Mysore, Arman Cohan, and Tom Hope. "Multi-Vector Models with Textual Guidance for Fine-Grained Scientific Document Similarity". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 4453–4470. DOI: 10.18653/v1/2022.naacl-main.331. URL: https://aclanthology.org/2022.naacl-main.331.

[37] John J. Nay et al. *Large Language Models as Tax Attorneys: A Case Study in Legal Capabilities Emergence*. 2023. arXiv: 2306.07075.

[38] Chau Nguyen et al. "CAPTAIN at COLIEE 2023: Efficient Methods for Legal Information Retrieval and Entailment Tasks". In: *Workshop of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE'2023) in the 19th International Conference on Artificial Intelligence and Law (ICAIL)*. 2023.

[39] Chau Nguyen et al. "Pushing the Boundaries of Legal Information Processing with Integration of Large Language Models". In: *New Frontiers in Artificial Intelligence*. Ed. by Toyotaro Suzumura and Mayumi Bono.

Singapore: Springer Nature Singapore, 2024, pp. 167–182. ISBN: 978-981-97-3076-6.

[40] Phuong Nguyen et al. "CAPTAIN at COLIEE 2024: Large Language Model for Legal Text Retrieval and Entailment". In: *New Frontiers in Artificial Intelligence*. Ed. by Toyotaro Suzumura and Mayumi Bono. Singapore: Springer Nature Singapore, 2024, pp. 125–139. ISBN: 978-981-97-3076-6.

[41] Animesh Nighojkar et al. "AMHR COLIEE 2024 Entry: Legal Entailment and Retrieval". In: *New Frontiers in Artificial Intelligence*. Ed. by Toyotaro Suzumura and Mayumi Bono. Singapore: Springer Nature Singapore, 2024, pp. 200–211. ISBN: 978-981-97-3076-6.

[42] Rodrigo Nogueira et al. "Document Ranking with a Pretrained Sequence-to-Sequence Model". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 708–718.

[43] OpenAI. *GPT-4 Technical Report*. 2023. URL: https://www.openai.com/research/gpt-4.

[44] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. "RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze!" In: *arXiv:2312.02724* (2023).

[45] Juliano Rabelo, Mi-Young Kim, and Randy Goebel. "Combining Similarity and Transformer Methods for Case Law Entailment". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. ICAIL '19. Montreal, QC, Canada, 2019, 290–296.

[46] Juliano Rabelo et al. "Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021". In: *The Review of Socionetwork Strategies* 16.1 (2022), pp. 111–133. ISSN: 1867-3236. DOI: 10.1007/s12626-022-00105-z. URL: https://doi.org/10.1007/s12626-022-00105-z.

[47] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[48] G. M. Rosa et al. "3B Parameters Are Worth More Than In-domain Training Data: A Case Study in the Legal Case Entailment Task". In: *Sixteenth International Workshop on Juris-informatics (JURISIN)*. 2022.

[49] Guilherme M. Rosa et al. "To Tune or Not to Tune? Zero-Shot Models for Legal Case Entailment". In: *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL)*. 2021.

[50] Guilherme Moraes Rosa et al. "To tune or not to tune?: zero-shot models for legal case entailment". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ICAIL '21. ACM, June 2021. DOI: 10.1145/3462757.3466103. URL: http://dx.doi.org/10.1145/3462757.3466103.

[51] Keshav Santhanam et al. "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3715–3734.

[52] Melanie Sclar et al. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=RIu5lyNXjT.

[53] Weiwei Sun et al. "Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14918–14937. DOI: 10.18653/v1/2023.emnlp-main.923. URL: https://aclanthology.org/2023.emnlp-main.923.

[54] ZhongXiang Sun et al. "Explainable Legal Case Matching via Graph Optimal Transport". In: *IEEE Transactions on Knowledge and Data Engineering* 36 (2024), pp. 2461–2475. URL: https://api.semanticscholar.org/CorpusID:264119987.

[55] Nandan Thakur et al. "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)*. 2021. URL: https://openreview.net/forum?id=wCu6T5xFjeJ.

[56] Nguyen Ha Thanh et al. "JNLP Team: Deep Learning for Legal Processing in COLIEE 2020". In: *COLIEE*. 2020.

[57] Liang Wang et al. "Improving Text Embeddings with Large Language Models". In: *arXiv preprint arXiv:2401.00368* (2023).

[58] Shitao Xiao et al. *C-Pack: Packaged Resources To Advance General Chinese Embedding*. 2023. arXiv: 2309.07597 [cs.CL].

[59] Sho Yokoi et al. "Word Rotator's Distance". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 2944–2960. DOI: 10.18653/v1/2020.emnlp-main.236. URL: https://aclanthology.org/2020.emnlp-main.236.

[60] Weijie Yu et al. "Optimal Partial Transport Based Sentence Selection for Long-form Document Matching". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 2363–2373. URL: https://aclanthology.org/2022.coling-1.208.

[61] Wei Zhao et al. "MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance". In: *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 563–578.

[62] Xu Zhao et al. "A Relaxed Matching Procedure for Unsupervised BLI". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 3036–3041. DOI: 10.18653/v1/2020.acl-main.274. URL: https://aclanthology.org/2020.acl-main.274.