

## Article

# Employing AI for Better Access to Justice: An Automatic Text-to-Video Linking Tool for UK Supreme Court Hearings

Hadeel Saadany <sup>1,\*</sup> , Constantin Orăsan <sup>2</sup> , Catherine Breslin <sup>3</sup>, Mikolaj Barczentewicz <sup>2</sup>  and Sophie Walker <sup>4</sup><sup>1</sup> School of Computing, Birmingham City University, Belmont Row, Birmingham B4 7RQ, UK<sup>2</sup> Centre for Translation Studies, University of Surrey, Guildford, Surrey GU2 7XH, UK; c.orasan@surrey.ac.uk (C.O.)<sup>3</sup> Kingfisher Labs Ltd., Cambridge, UK<sup>4</sup> Just Access, Leeds LS1 2BH, UK

\* Correspondence: hadeel.saadany@bcu.ac.uk

## Abstract

The increasing adoption of artificial intelligence across domains presents new opportunities to enhance access to justice. In this paper, we introduce a human-centric AI tool that utilises advances in Automatic Speech Recognition (ASR) and Large Language Models (LLMs) to facilitate semantic linking between written UK Supreme Court (SC) judgements and their corresponding hearing videos. The motivation stems from the critical role UK SC hearings play in shaping landmark legal decisions, which often span several hours and remain difficult to navigate manually. Our approach involves two key components: (1) a customised ASR system fine-tuned on 139 h of manually edited SC hearing transcripts and legal documents and (2) a semantic linking module powered by GPT-based text embeddings adapted to the legal domain. The ASR system addresses domain-specific transcription challenges by incorporating a custom language model and legal phrase extraction techniques. The semantic linking module uses fine-tuned embeddings to match judgement paragraphs with relevant spans in the hearing transcripts. Quantitative evaluation shows that our customised ASR system improves transcription accuracy by 9% compared to generic ASR baselines. Furthermore, our adapted GPT embeddings achieve an F1 score of 0.85 in classifying relevant links between judgement text and hearing transcript segments. These results demonstrate the effectiveness of our system in streamlining access to critical legal information and supporting legal professionals in interpreting complex judicial decisions.

**Keywords:** automatic speech recognition; legal transcription; UK Supreme Court; legal information retrieval; embedding customisation



Academic Editor: Douglas O'Shaughnessy

Received: 25 June 2025

Revised: 6 August 2025

Accepted: 8 August 2025

Published: 21 August 2025

**Citation:** Saadany, H.; Orăsan, C.; Breslin, C.; Barczentewicz, M.; Walker, S. Employing AI for Better Access to Justice: An Automatic Text-to-Video Linking Tool for UK Supreme Court Hearings. *Appl. Sci.* **2025**, *15*, 9205. <https://doi.org/10.3390/app15169205>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

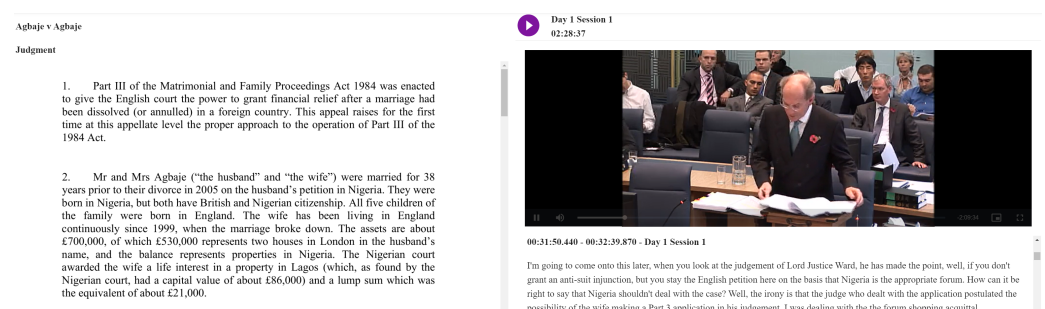
## 1. Introduction

Despite the growing interest in applying Natural Language Processing (NLP) techniques to legal texts, spoken court proceedings, particularly those of the UK Supreme Court (SC), remain under-explored. These hearings are rich in legal reasoning and play a pivotal role in shaping landmark judgements, yet they often span several hours of audio content, making manual navigation and comprehension time-consuming and inefficient for legal professionals and the public alike. Currently, the transcription of such hearings is largely performed by human transcribers, a process that is costly, labour-intensive, and unable to scale to meet the demand of over 449,000 cases annually across UK tribunals [1]. While generic Automatic Speech Recognition (ASR) systems offer a potential solution, they are typically trained on general-domain data and fail to capture the specialised vocabulary and

discourse patterns of legal proceedings, which results in suboptimal transcription quality. Moreover, there are no existing automated systems that semantically link segments of hearing transcripts to the corresponding paragraphs in written judgements, which are essential for understanding the rationale behind judicial decisions. This lack of semantic alignment between spoken and written legal content creates a barrier to efficient legal analysis, training, and public accessibility. To address these challenges, we propose a domain-adapted ASR and semantic linking system that automatically transcribes UK SC hearings and connects them to relevant sections of the final judgement text. This system aims to enhance legal transparency, reduce manual effort, and support informed legal interpretation.

In this paper, we summarise our combined research–industry project for building an automated tool designed for linking segments in UK SC text judgements to semantically relevant timespans in the videos of their relevant hearings [2–4]. The project involved two stages: (1) building a customised Automatic Speech Recognition (ASR) system that is specifically tailored for SC court hearings to ensure accurate transcription of the court hearings and (2) deploying the customised ASR system to transcribe a large dataset of UK SC hearing videos and using it to build an Information Retrieval system that links paragraphs in the text of a case judgement to their relevant spans in the court hearing video. The main objective of this tool is to provide legal professionals, as well as the general public, with an automatic navigation tool that pins down the arguments and legal precedents presented in the long hearing sessions and identifies those that are of particular importance in how the judges made their decision on the case. Although the system presented in this paper was developed in the context of linking court hearings with the case judgement, the methodology described can be adapted to other scenarios which require linking of spoken data with textual information.

Figure 1 shows a screenshot of the user interface created in the project. On the left side of the screen, the paragraphs of the written judgement are displayed. The user can then scroll down to choose a specific paragraph in the judgement. On the right side, the timespan in the court hearing video that is semantically relevant to the legal point mentioned in the selected judgement paragraph is displayed along with temporal metadata (session number, day and time). The user can play the particular timespan and go back and forth around it, as well as read our tool’s transcription of the speech.



**Figure 1.** User interface for linking judgement to bookmarks in video court sessions [3].

When we started the project, we realised that there is an increasing interest in employing NLP techniques to aid the textual processing of the legal domain [5–8]; in contrast, the processing of spoken court hearings has not received the same level of attention. In the UK legal system, the court hearing sessions have a unique tradition of verbal argument that is used to clarify arguments, challenge assumptions, and explore the broader implications of each case. Moreover, UK Supreme Court hearings crucially aid in new case preparation, provide guidance for court appeals, help in legal training, and even guide future policies. However, as mentioned above, the length of the audio materials and the fact that transcription is usually manual limit the access to the information

Although there are numerous speech-to-text (STT) technology providers that could be used to transcribe this data automatically, most of these systems are trained on general domain data, which may result in domain-specific transcription errors if applied to a specialised domain. One way to address this problem is for end-users to train their own ASR engines using their in-domain data. However, in most cases the amount of data available is too low to enable them to train a system from scratch that can compete with well-known cloud-based ASR systems that are trained on much larger datasets. At the same time, in commercial scenarios, using generic cloud-based ASR systems to transcribe a specialised domain may result in suboptimal quality transcriptions for clients who require this service.

This also holds true when transcribing the UK SC proceedings. When applying a generic cloud-based ASR system (in our case Amazon Transcribe) on SC hearings, the Word Error Rate (WER) remains relatively high due to the challenges presented by the presence of several speakers, complex speech patterns, and more crucially, unique pronunciations and domain-specific vocabulary. The examples in Table 1 show some common problems that we faced when transcribing UK court hearings using off-the-shelf ASR systems such as Amazon Web Services (AWS) Transcribe (<https://aws.amazon.com/transcribe/>, accessed on 12 August 2025). The references in the table are taken from human-generated ground-truth transcripts of real UK SC hearings (<https://www.supremecourt.uk/decided-cases/index.html>, 19 June 2024) created by the legal editors in our project's team. The first error is due to the special pronunciation of the phrase "my lady" in British courtrooms, as it is pronounced "mee-lady" when barristers address a female judge. Similarly, in the second example, the error relates to the linguistic etiquette of UK court hearings which the ASR system consistently fails to recognise. The error in the third example, on the other hand, is related to legal terminology critical to the specific transcribed case. Errors similar to the third example are numerous in our dataset and also affect named entities such as numbers and names that are vital in understanding the legal argument in the transcribed cases. These errors can lead to serious information loss and cause confusion.

**Table 1.** Examples of errors produced by Amazon Transcribe in legal hearings. The references were produced by editors with legal background. The errors and their correct forms are in bold.

Model	Transcript
Reference	So <b>my lady</b> um it is difficult to...
AWS ASR	So <b>melody</b> um it is difficult to...
Reference	<b>All rise</b> ...
AWS ASR	<b>All right</b> ...
Reference	it makes further <b>financial order</b>
AWS ASR	it makes further <b>five natural</b>

The first stage of the project described in this paper and presented in Section 3.1 focused on domain adaptation of a generic ASR system to mitigate the errors in the automated UK court transcription services. We address this by fine-tuning off-the-shelf ASR systems with a custom language model (CLM) trained on legal documents as well as 139 h of human-edited transcriptions of UK SC hearings. We also employ NLP techniques to automatically build a custom vocabulary of common multi-word expressions and word n-gram collocations that are critical in court hearings. We infuse our custom vocabulary in the CLM at transcription time. In this study, we evaluate the benefits of our proposed domain adaptation methods by comparing the word error rate of the CLM output with two off-the-shelf ASR systems: AWS Transcribe (commercial) and the OpenAI Whisper model (open source) [9]. We also compare the general improvement in the ASR system's ability to correctly transcribe

legal entities with and without adopting our proposed methods. In addition, we discuss the transcription time with different ASR settings since transcription time is critical for the commercial pipeline implemented by the industrial partner of the project.

In Section 3.2, we present the second stage of the project where we use the SC audio material transcribed by our customised ASR system to construct an integrated system for the automatic navigation of segments in the media data of UK SC hearings based on their semantic relevance to particular paragraph(s) in the text of the judgement issued following the hearing. Based on the timing metadata of the court hearing transcription segments, we assign bookmarks on the video sessions and link them to their semantically relevant paragraphs in the judgement text.

Thus, the major contributions of our research can be summarised as follows:

1. **Customised ASR System for Legal Domain**

We developed and fine-tuned an ASR system specifically for UK SC hearings. This involved training a custom language model using legal documents and human-edited transcripts and integrating domain-specific vocabulary to enhance transcription accuracy and efficiency.

2. **Semantic Linking Between Judgement Text and Hearing Video**

We designed an automated information retrieval system that links paragraphs in written judgements to semantically relevant segments in hearing videos, enabling precise navigation and contextual understanding of legal arguments.

3. **Integrated User Interface for Legal Professionals and Public Use**

We created a user-friendly interface that synchronises textual and audiovisual data and allows users to select judgement paragraphs and view corresponding video segments with playback and transcript functionality.

4. **General Framework for Audio-Text Alignment in Specialised Domains**

We proposed a scalable and adaptable methodology for linking audiovisual content with textual information, applicable to other specialised domains such as education, healthcare, and policy analysis.

The rest of the paper is structured as follows: in Section 2, we briefly summarise the more relevant research in ASR as well as recent information retrieval systems used within the legal domain. In Section 3, we illustrate our pipeline used to build the text-to-video linking tool. Section 4 presents the results of the two stages of our project: customising the ASR model and building the IR system which act as the back-end of our tool. In Section 5, we illustrate both an error analysis of the experiment as well as the feedback we attained from our stakeholders. Finally, in Section 6, we discuss our conclusions of the overall experiment as well as the benefits of our AI tool in providing a better access to justice.

## 2. Related Work

In this section, we provide a review of the relevant literature. We start with a brief overview of research carried out in the field of Automatic Speech Recognition in Section 2.1, followed by a discussion about Information Retrieval methods used in the legal domain in Section 2.2.

### 2.1. Automatic Speech Recognition

Automatic Speech Recognition (ASR) models convert audio input to text, and they have optimal performance when used to transcribe data similar to the data on which they were trained. However, performance degrades when there is a mismatch between the data used for training and that which is being transcribed. Additionally, some types of audio material are intrinsically harder for speech recognition systems to transcribe. In practice, this means that Automatic Speech Recognition system performance degrades

when, for example, there is background noise [10], non-native accents [11,12], young or elderly speakers [11], or a shift in domain [13].

Performance degradation is typically mitigated by adapting or fine-tuning ASR models towards the domain of the targeted data by using a domain-specific dataset [14–16]. Some methods for domain adaptation adopt NLP techniques such as using machine translation models to learn a mapping from out-of-domain ASR errors to in-domain terms [17]. An alternative approach is to build a large ASR model with a substantially varied training set so that the model is more robust to data shifts. An example of this latter approach is the recently released OpenAI Whisper model, which is trained on 680k hours of diverse domain data to effectively generalise a range of unseen datasets without the need for explicit adaptation [9].

ASR models are usually evaluated using Word Error Rate (WER), which treats each incorrect word equally. However, ASR models do not perform equally on different categories of words. Performance is worse for categories like names of people and organisations as compared to categories like numbers or dates [18]. For this reason, there is ASR research targeted specifically for improving specific errors such as errors related to entities using NLP techniques [19,20]. Correct recognition of entities is particularly important for our purpose as the entities can serve as important anchor points between the paragraphs in the court judgement and video timestamps.

To solve the challenges in transcribing the special domain of UK SC hearings, we adopt simple techniques to improve the effect of the domain mismatch between a generic ASR model and the specialised domain of British courtroom hearings. Our proposed method improves both the system's WER rate as well as its ability to capture case-specific terms and entities. In Section 3.1, we present the setup of our experiments and the evaluation results.

## 2.2. Information Retrieval for Legal Domain

Our literature survey has shown that there has recently been an increased interest in employing NLP techniques to aid the textual processing in the legal domain [5–8]. The main focus has been on legal document summarisation [21,22], predicting judgements [23,24], and contract preprocessing and generation [25,26]. Moreover, NLP methods for Information Extraction and Textual Entailment have been extensively used in NLP legal to either find an answer to a legal query in legal documents [27] or to connect textual data [28].

Chalkidis et al. [29] experimented with different IR models to extract relevant EU and UK legislative acts that are important for organisations' *regulatory compliance*, which they need to ensure compliance with the relevant laws. Their experiments show that fine-tuning a BERT model on an in-domain classification task is the best pre-fetcher for their dataset. Similarly, Kiyavitskaya et al. [30] use textual semantic annotation to extract government regulations in different countries which companies and software developers are required to comply with. They show that AI-based IR tools are effective in reducing the human effort to derive requirements from regulations.

Although there has recently been a significant increase in legal IR research, the processing and deployment of spoken court hearings for legal IR has not received the same attention as understanding and extracting information from textual legal data. The next section shows how we employed ASR technology to build an IR tool which automatically connects judgements and videos of court hearings.

## 3. Models and Methods

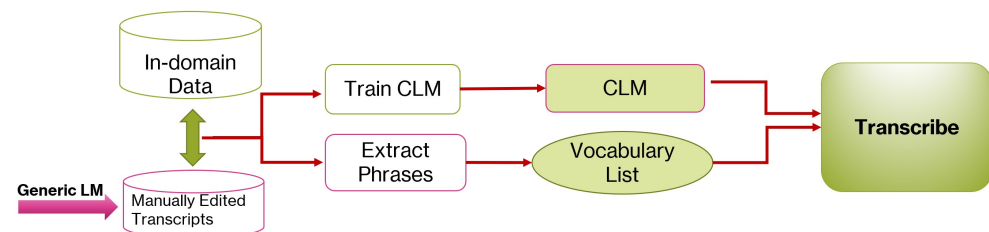
This section describes the two stages of our system. In the first stage, we build a custom language model (CLM) by fine-tuning the base AWS ASR system using two types of training data: (1) textual data from the legal domain and (2) a corpus of human-generated

legal transcriptions. Second, we use NLP techniques to extract domain-specific phrases and legal entities from the in-domain data to create a vocabulary list. We use both the CLM and the vocabulary list for transcribing legal proceedings. The objective of this stage is to obtain a high-quality transcript for our automatic retrieval model in the second stage. This approach is described in Section 3.1.

The second stage consists of building an IR system capable of extracting the best n-links between a paragraph(s) of the judgement text and the timespans of the transcribed video sessions of that particular case. The links are then translated into timestamp bookmarks in the long videos of each case to be used in constructing our UI.

### 3.1. Stage One: A Customised ASR Model

In this section, we explain how we customised the AWS Transcribe base model for the legal domain. Figure 2 illustrates the pipeline used for customising of an ASR system tailored for the UK SC hearings. The pipeline begins by feeding two types of in-domain data into a generic language model: (1) legal texts from Supreme Court judgements and (2) manually edited transcripts of Supreme Court hearings. This data serves a dual purpose. First, it is used to train a customised language model (CLM) that captures the linguistic nuances and domain-specific terminology of UK legal discourse. Second, it is processed by a phrase extraction model to identify and extract key legal phrases prevalent in the textual data. The resulting CLM and the extracted in-domain phrases are then integrated into the ASR system to transcribe Supreme Court hearing audio. This enriched transcription process ensures higher accuracy and contextual relevance, forming the basis for evaluating the performance of the adapted ASR system. We compare the performance of our CLM model with the accuracy of AWS Transcribe base ASR system and OpenAI Whisper open-source ASR system when transcribing  $\approx 12$  h of UK Supreme Court hearings.



**Figure 2.** Pipeline for improving ASR output for legal specific errors [2].

#### 3.1.1.1. Dataset Compilation

For our experiments, we compiled two open-source datasets from the UK legal domain. The first is a dataset of 43 Supreme Court written judgements with over 3.26 million tokens scraped from the official site of the UK Supreme Court (<https://www.supremecourt.uk/cases/>, accessed on 12 August 2025). The second dataset consists of  $\approx 139$  h for 10 SC hearings sessions downloaded from the British National Archive (<https://discovery.nationalarchives.gov.uk/>, accessed on 19 June 2024). To ensure representativeness, we selected cases spanning diverse legal domains including commercial disputes, criminal appeals, and custody challenges. The audio sessions were transcribed using the Amazon Transcribe service. This transcription was then post-edited by a team of legal professionals using a specially designed post-editing tool which enables them to compare the ASR output to the audio files of the SC hearing sessions.

We applied word-level Levenshtein distance [31] between the ASR output and the edited transcripts of a sample of the transcribed cases in our dataset in order to highlight differences between them and identify challenges that cause error in the automatic transcription. The main challenges can be summarised as follows:

- *Overlapping speech and background noise* due to the logistics of the court hearings' settings, as the barristers frequently ask the court to turn to specific pages in the case file.
- *UK legal jargon* was consistently mistranscribed due to special pronunciation of some phrases in the English court as is the case in the first example in Table 1. Also, repeated forms of address that have a special pronunciation lead to transcription errors. For example, a barrister addressing a colleague as "My learned friend" is pronounced as "my learn'id friend" with a stress on the second syllable of "learned".
- *Legal entities* such as case names with non-English names (e.g., Agbaje (Respondent) v Akinnoye-Agbaje (FC) (Appellant)), provisions (e.g., Section 84 1 ), and cardinals crucial to the discussed case were frequently mistranscribed.
- *Legal terms* specific to the deliberated case were often mistaken by the ASR system to phrases with similar pronunciation. For example, the legal term "inherent vice" was consistently mistranscribed as "in your advice". This most likely relates to the fact that the ASR system opts for the most acoustically similar phrase provided by its language model, which is trained on non-domain data.

The first challenge requires processing of the audio signal in order to remove the background noise and can be performed to a certain extent with the latest advances in AI. While this would improve the overall accuracy of ASR, it would not help much with the recognition of domain-specific terminology and jargon. The next sections present how we customised the ASR engine for the legal domain to address the last three challenges identified.

### 3.1.2. Customising the ASR System

The architecture used by AWS Transcribe relies on a recurrent neural network transducer (RNN-T) [32]. This is an end-to-end model for automatic speech recognition (ASR) which has gained popularity in recent years as it folds the separate components of the ASR system (i.e., acoustic, pronunciation and language models) into a single neural network [33]. The RNN-T speech recognition architecture involves determining the most likely word sequence,  $W = w_1, \dots, w_n$ , given an acoustic input sequence,  $x = x_1, \dots, x_T$ , where  $T$  represents the number of frames in the utterance:

$$W^* = \underset{W}{\operatorname{argmax}} P(W|x), \quad (1)$$

which is typically decomposed into three separate models as follows:

$$W^* = \underset{W}{\operatorname{argmax}} \sum_{\phi} P(x, \phi|W)P(W) \quad (2)$$

$$\approx \underset{W, \phi}{\operatorname{argmax}} p(x|\phi)P(\phi|W)P(W) \quad (3)$$

where the acoustic model,  $p(x|\phi)$ , predicts the likelihood of the acoustic input speech utterance given a phoneme sequence,  $\phi$  [34].

The AWS Transcribe platform allows the fine-tuning of this architecture via building custom language models to improve transcription accuracy for domain-specific speech. In contrast to other approaches that require the fine-tuning of audio files and their transcriptions, the custom language model (CLM) for AWS Transcribe is created by infusing a significant amount of text data on top of the base model without the need of audio files. The text data must contain domain-specific vocabulary that would normally appear in the audio files. For training our CLM, we used the two datasets from the legal domain introduced in the previous section.

### 3.1.3. Phrase Extraction Model

The second component we use in customising the base ASR model is a legal-specific vocabulary list that is infused at the fine-tuning time. The list is extracted from the gold-standard transcriptions of SC hearings along with the SC textual judgements dataset used for training the CLM. To create the vocabulary list, we implemented two methods. First, we used the Gensim library (<https://radimrehurek.com/gensim/models/phrases.html>, accessed on 12 December 2024) to train a phrase detection model to automatically detect common phrases—aka multi-word expressions—from the stream of sentences in our dataset. We trained the model to extract bigram collocations based on Pointwise Mutual Information (PMI) scores [35]. PMI is a measure of association between words; it compares the probability of two words occurring together to what this probability would be if the two words were independent. Words that normally occur together in a domain will have higher PMI scores than the rest of words. We train the collocation model using the Gensim 4.3.3 Python library with a minimum score threshold for a bigram to be taken into account set to 1 and PMI as the probability scoring method [36]. The collocation model is trained on the textual data of the SC transcriptions and the SC judgements. The model is then frozen and used to extract a list of most common bigrams in the whole dataset.

The second method we employed to create a list of custom vocabulary is the identification of named entities in our dataset. For this purpose, we used Blackstone (<https://research.iclr.co.uk/blackstone>, accessed on 15 June 2024), an NLP library built on top of spaCy (<https://spacy.io/>, accessed on 15 June 2024) capable of identifying legal entities. The list of legal entities includes Case Name, Court Name, Provision (i.e., a clause in a legal instrument), Instrument (i.e., a legal term of art), and Judge. We concatenated this Blackstone entity list with the list of non-legal entities that spaCy library normally identifies such as Cardinals, Persons and Dates.

Figure 3 shows an example of the type of common phrases extracted by our collocation model along with their frequencies. As can be seen from the figure, the extracted phrases include frequent legal terms (highlighted in blue) as well as named entities such as names of institutions and persons (highlighted in yellow) which are specific to the SC cases included in the training corpus. Thus, our vocabulary list comprised the main error categories which were identified in the output of the base ASR system. We use our vocabulary lists in two ways: (1) as part of the CLM training data and (2) as a custom vocabulary list infused at transcription time. The results of applying our domain-adaptation methods for the transcription of Supreme Court case hearings are explained in the Section 4.

```

17267.38 khan toddler
17267.38 illustrative nonexhaustive
17267.38 alba cora
17267.38 actus reus (legal )
16576.68 sri lanka
15348.78 swimming pool
15348.78 gillette industries (institution)
15348.78 colonel karuna
15348.78 blank cheque
14095.82 et cetera
14095.82 bona fide (legal)
13813.9 pottu amman (names)
13813.9 mens rea (legal)
13813.9 dumbarton oaks
13156.1 al qaeda
12950.53 wayne tank
11937.94 prima facie

```

**Figure 3.** Example of common collocations extracted by the phrase extraction model [2].

### 3.2. Stage Two: Text-to-Video IR System

In this section, we present the method we proposed to link paragraphs from court judgements with timestamps from videos. For the sake of simplicity, we will refer to the former as *paragraphs* and the latter as *timestamps*. In Section 3.2.1, we describe how

we prepared the data for our experiments. Given the lack of training data to train a classifier which determines whether there is a link between a paragraph and a timestamp, we first experiment with existing retrieval methods to determine which one is the most appropriate for our purpose. These experiments are presented in Section 3.2.2. Armed with this information, we annotate a dataset in Section 3.2.3. However, this annotation is still time-consuming. For this reason, we propose an automatic method to augment our dataset in Section 3.2.4. A number of classifiers which are able to determine whether there is a link between a paragraph and a timestamp are proposed in Section 3.2.5.

### 3.2.1. Data Processing and Preparation

We treat the linking of a judgement paragraph(s) to the relevant timespan transcripts of a video session as a semantic search task. We first transcribe the video sessions using our custom speech-to-text language model and then segment the judgement into paragraphs. The paragraphs are treated as a query and the transcript of the case is the corpus in which we search for an answer to that query. More formally, given a judgement segment  $q$  and a set of candidate timespan transcripts  $C = \{c_1, c_2, \dots, c_n\}$ , the task is to find the timespans in the video transcripts  $T = \{t_1, t_2, \dots, t_n \mid t_i \in C \wedge (t_i, q)\}$  where  $(t_i, q)$  denotes a semantic link between the information presented in the timespan transcript and the argument put forward in the judgement segment.

The main challenge in preprocessing the dataset was segmentation of the judgement text into semantically cohesive sections that would be treated as queries in our IR method. We noticed that typically a Supreme Court judgement is structured manually into sections such as: "Introduction", "The context", "Facts of the Case", "The Outcome of the Case", etc. However, after we carefully scrutinised the dataset, we found that the naming of sections is not consistent and therefore building an automatic segmentation method around these labels would be difficult. On the other hand, the judgement texts are consistently divided into enumerated paragraphs (typically a digit(s) followed by a dot). We opted, therefore, for segmenting the judgement text into windows of enumerated paragraphs. After experimenting with different window sizes, the optimum window size consisted of three enumerated paragraphs. The average length of this window was 389 tokens per segment.

The preprocessing of the transcription consisted mainly of excluding very short timespans since they were mostly either interjections (e.g., "Yes, sorry, I'm not following", "I beg your pardon", etc.) or reference to logistics of the hearing (e.g., "This is your paper, isn't it?", "Please turn to the next page", etc.). We chose to exclude transcription spans less than 50 tokens as an empirical threshold for semantically significant conversation units. For both the judgement and transcript data, we cleaned empty lines and extra spaces but kept punctuation intact, especially in judgement segments as it is essential in identifying names of cases and legal provisions (the UK legal system has a unique punctuation style for case names such as "R v Chief Constable of South Wales [2020] EWCA Civ 1058", which are crucial in understanding legal precedents).

### 3.2.2. Zero-Shot Information Retrieval

The ability of an IR system to retrieve the top-N most relevant results is usually assessed by comparing its performance with human-generated similarity labels on a sentence-to-sentence or query-to-document similarity dataset(s) (e.g., [37–39]). In order to create a human-validated evaluation dataset, we needed human annotators to manually check the correct links between judgement segments and the timespans of video hearing transcripts for each of our chosen cases. However, in our use case, this is not feasible since to annotate one SC case with, for example, 50 judgement segments and 300 timespans of video

transcript, as the annotators will need to read and compare  $50 \times 300$  judgement–timespan pairs, which means they need to decide on 15,000 links per case.

To overcome this problem, we adopted a zero-shot IR approach. Accordingly, we embedded all judgement and transcript segments in our corpus into the same vector space and used the cosine similarity as our semantic distance metric to extract the top closest 20 transcript timespans per judgement segment in the vector space. We first experimented with different ways to encode the judgement segments and transcription timespans as numeric vectors for a single case in our dataset. Then, we assigned a human annotator, a post-graduate law student, to evaluate for each segment from the judgement the first 20 links produced by the model. For the case above, this reduced the number of judgements to  $50 \times 20 = 1000$ , which is a more manageable task. The annotator compared each judgement segment against each timespan to choose either “Yes” there is a semantic link or “No” there is not. This was performed using a specially designed interface that also allowed them to play the corresponding video timespan if necessary. The IR models used for our experiments are the following:

#### A. Frequency-based Methods (keyword search)

**Okapi BM25 [40]:** BM25 is a traditional keyword search based on a bag-of-words scoring function estimating the relevance of a document  $d$  to a query  $q$ , based on the query terms appearing in  $d$ . It is a modified version of the tf-idf function where the ranking scores change based on the length of the document  $d$  in words, and the average  $d$  length in the corpus from which documents are drawn.

#### B. Embedding-based Methods

**Document Similarity with Pooling:** We experimented with different pooling methods of the GloVe [41] pretrained word embeddings. The GloVe vector embeddings were trained using an unsupervised method on a general domain dataset [42]. We create vectors for the judgement segments and the transcripts’ timespans from the mean, minimum, and maximum values of the GloVe embeddings.

**Entailment Search:** We used embeddings from a pretrained model for textual entailment that is trained to detect sentence pair relations, i.e., one sentence entails or contradicts the other. We employed the Microsoft MiniLM model [43] which is trained on the Microsoft dataset MiniLM-L6-H384-uncased and fine-tuned on a 1B sentence pairs dataset. The potential link in this case is whether the judgement paragraph(s) entails the particular segment of the video transcript.

**Legal BERT:** Our dataset comes from the legal domain, which has distinct characteristics such as specialised vocabulary, particularly formal syntax, and semantics based on extensive domain-specific knowledge [44,45]. For this reason, we employed the Legal BERT [46] which is a family of BERT models for the legal domain pre-trained on 12 GB of diverse English legal text from several fields (e.g., legislation, court cases, and contracts). The judgement text and the video transcript data were converted into the Legal BERT pretrained word embeddings.

**Asymmetric Semantic Search:** Asymmetric similarity search refers to finding similarity between unequal spans of text, which may be particularly applicable to our case where the judgement text may be shorter than the span of the video transcript. For this purpose, we created the embeddings using the MS MARCO model [47] which is trained on a large-scale IR corpus of 500k Bing query examples.

**GPT Question–answer linking:** In this setting, a question–answer linking approach is adopted where the selected judgement text portion is treated as a question and the segments of the video transcript as potential answers. We use pretrained embeddings obtained from OpenAI’s latest text-embedding-ada-002 GPT model, which outperforms the

previous most capable GPT model, Davinci, at most tasks (<https://openai.com/blog/new-and-improved-embedding-model>, accessed on 19 June 2024). The context length of the ada-002 model is increased by a factor of four, from 2048 to 8192, making it more convenient for the long documents in our dataset. We use the model's contextual representations of our corpus to find answers in video timespans for each segment in the judgement which is treated as a prompt query.

### 3.2.3. Results of Pre-Fetching

To assess the performance of each model in comparison to the human judgement, we calculated the mean average precision (MAP) which is the de facto IR metric:

$$MAP = \frac{1}{Q} \sum_{q=1}^Q AP(q) \quad (4)$$

where  $Q$  is the total number of queries, in our case the judgement segments, and  $AP(q)$  is the average precision of a single query  $q$ .  $AP(q)$  evaluates whether all of the timespans assigned as relevant by the annotator are ranked the highest by the model. We calculated MAP for the first 5, 10, and 15 judgement–timespan pairs. We report the values for the MAP metric in Table 2. For completion, we also report the Recall@ $k$  score calculated with respect to the 20 links annotated in the gold standard:

$$\text{Recall@}k = \frac{|\text{Relevant}_k|}{|\text{Total Relevant}|} \quad (5)$$

where

- $|\text{Relevant}_k|$  is the number of relevant links retrieved among the top  $k$  candidates.
- $|\text{Total Relevant}|$  is the total number of relevant links that exist for a given query.

**Table 2.** Results of unsupervised IR for linking judgements to video transcripts in one case. Bold indicates the best scores.

Model	MAP@5	Recall@5	MAP@10	Recall@10	MAP@15	Recall@15
<b>GPT</b>	<b>0.96</b>	<b>0.33</b>	<b>0.89</b>	<b>0.57</b>	<b>0.85</b>	0.77
<b>Entailment</b>	0.87	0.32	0.85	0.55	0.82	<b>0.79</b>
<b>Glove</b>	0.81	0.27	0.77	0.53	0.61	0.78
<b>BM25</b>	0.87	0.29	0.81	0.53	0.78	0.77
<b>Asymmetric</b>	0.94	0.32	0.88	0.54	0.83	0.77

We report **Recall@5**, **Recall@10**, and **Recall@15** to assess the system's ability to retrieve correct links when various number of links are considered. A high Recall@5 indicates that most relevant links are retrieved within the top 5 results, which is particularly valuable in legal contexts where precision and interpretability are critical. This metric provides insight into the system's effectiveness in surfacing semantically meaningful connections between judgement text and spoken court hearing segments and hence supports legal professionals in navigating lengthy hearing sessions.

As can be seen from Table 2, the GPT model demonstrated the best performance in comparison to the other models when evaluated on one case. Thus, to create a dataset for annotation for the rest of the cases, we extracted the top 15 links for each judgement–transcript segment according to the cosine similarity scores of the GPT embedding model. We also extracted 5 links with the lower ranks (50 to 55) to avoid bias to the GPT model and randomly shuffled the 20 links for each judgement–transcript segments. After this

processing, the dataset constructed for manual annotation consisted of 3620 judgement-to-transcript documents. Human annotators were again asked to judge whether the extracted timespan transcripts were semantically linked or not linked to the judgement paragraph(s).

To assess the relevance of legal information presented in multimedia content, annotators were instructed to follow a structured evaluation protocol. Each annotator was required to (1) listen to the video segment, (2) read the corresponding transcript, and (3) examine the associated judgement paragraph. Based on this multimodal input, annotators determined whether the legal information conveyed in the video span was relevant to the judgement text. The annotators consisted of five undergraduate and postgraduate law students, who were encouraged to apply their domain knowledge and legal reasoning skills to make informed decisions. This protocol was designed to simulate realistic legal assessment scenarios where contextual understanding and interpretive judgment are essential.

The agreement between annotators was evaluated using Cohen's kappa as a statistical measure. The Cohen's kappa score between the annotators was 0.43, indicating a moderate level of agreement. The joint probability of agreement was 0.73, suggesting that annotators concurred on 73% of the cases, although this metric does not adjust for chance agreement. These results suggest that while annotators were reasonably aligned in their assessments, the task involved a degree of subjectivity, likely influenced by individual interpretations and varying levels of legal expertise. The relatively lower kappa score can also be attributed to the inherent complexity of the legal relevance task and the large number of annotators involved, which increases variability in judgment. The human annotations were compared to the results of all the embedding models mentioned above.

Table 3 demonstrates that the GPT text embedding model is again superior to the other models. This is the case when we evaluate the models with the first 5, 10, and 15 hits. It should be pointed out that our use case is different from a typical IR task where the efficacy of the model is normally evaluated by its ability to obtain the best links in the very first few hits (optimally hits 1 to 5). The reason is that the output of the model is used to bookmark the long video sessions at the parts most relevant to the legal argument stated in the judgement segment. The end-users of our tool can watch or draw the cursor around the bookmarks to obtain more information. Accordingly, our system's priority is to extract as many relevant bookmarks as possible from all the true relevant links in the long video sessions. Recall@15 and MAP@15, therefore, are of the highest importance in our retrieval results.

**Table 3.** Results of unsupervised IR for linking judgements to video transcripts for all cases. Bold indicates the best scores.

Model	MAP@5	Recall@5	MAP@10	Recall@10	MAP@15	Recall@15
<b>GPT</b>	<b>0.691</b>	<b>0.391</b>	<b>0.622</b>	0.657	<b>0.711</b>	<b>0.914</b>
<b>Entailment</b>	0.615	0.348	0.568	0.611	0.66	0.885
<b>Glove</b>	0.526	0.316	0.506	0.602	0.607	0.884
<b>BM25</b>	0.655	0.377	0.612	<b>0.659</b>	0.698	0.902
<b>Asymmetric</b>	0.602	0.347	0.553	0.619	0.664	0.908
<b>LegalBert</b>	0.557	0.326	0.531	0.613	0.632	0.896

After annotation, we compiled a dataset of 3620 judgement-to-transcript documents annotated with gold-standard linking labels. The dataset we compiled is relatively small in comparison with the datasets usually used to train classifiers. For this reason, the next section describes our method for augmenting this dataset.

### 3.2.4. Data Augmentation

The task of annotating our dataset is both expensive and time-consuming for two reasons: (1) it requires annotators with legal knowledge and (2) it involves the reading and understanding of the case particulars by the expert annotator in order to understand the latent semantic relevancy that can be used to extract more relevant links. For this reason, we decided to employ AI generative technology to augment our gold-standard dataset by generating more relevant judgement–transcript segment pairs. The augmented dataset was used along with the gold-standard for training the relevancy model.

Recently, several research studies have managed to successfully use GPT-3.5 and GPT-4 prompt engineering as an aiding tool for several NLP tasks (e.g., Qin et al. [48], Wang et al. [49], Törnberg [50]). One successful use of prompt engineering was to substitute crowd-sourced paraphrasing with a GPT chat model. Research has shown that ChatGPT-generated paraphrases are lexically and syntactically more diverse than human-generated ones [51]. Accordingly, we used the InstructGPT API (GPT 3.5) *set role* prompt strategy to extract paraphrases for the transcript side of the positive instances in our dataset [52]. The following prompt was used to create paraphrases of the transcript segments:

*I want you to act like a British lawyer. Paraphrase the following text:  
{original text}*

The paraphrases were created by the text-davinci-002 model, which was the latest model at the time when we run the experiment, and we set the parameters of `max_tokens` to 1400 tokens and the temperature to 0.7 to balance the degree of randomness for the models output. A sanity check was conducted on a randomly selected sample of the AI-generated paraphrases by a senior academic legal expert in our research team to make sure the paraphrased transcript has the same meaning as the original. While synthetic augmentation enhances query diversity, it also introduces a risk of overfitting for the positive links, as the model may learn to overly rely on the stylistic patterns or phrasing conventions introduced by the paraphrasing algorithm rather than generalising to our original transcription data. Thus, in order to avoid overfitting we generated negative samples to be used in training the relevancy model. For creating the negative samples, we adopted two approaches. The first was random shuffling of judgement–hearing segments from different cases. To reduce the effect of randomness, we chose the judgement–hearing segment pairs with the highest cosine similarity scores between their GPT-3 text embeddings. The second technique was the in-batch negative sampling during training which will be explained in the next section.

The augmented dataset amounted to 7248 judgement–hearing links with  $\approx 42$  M tokens. We used both the gold-standard and the augmented datasets to build a judgement–hearing relevancy model. The methods explored to determine whether there is a link between a paragraph and a timestamp are presented in the next section and evaluated in Section 4.

### 3.2.5. Paragraph-Timestamp Link Classifiers

The end-product of our project is a UI that bookmarks important timespans in the UKSC court hearing videos and links them to the judgement segments. Accordingly, we aim to use the compiled dataset to build a model that is capable of extracting as many transcript segments as possible per each judgement segment for the UKSC cases in the dataset. In the following sections, we describe our experiments where we trained several classification models on both the non-augmented and the augmented gold-standard dataset of judgement–hearing segment pairs. We split each of the augmented and the non-augmented datasets into an 80% training and a 20% testing set, ensuring the split was performed using random sampling for unbiased selection.

#### **Baseline Model:**

For our baseline, we train a logistic regression model with the GPT-3 embedding representations of the original data with and without the augmentation. We conduct the two experiments with two settings: (1) we use the concatenated vectors of each judgement–segment pair as the input features, and (2) we add the cosine-similarity score between each judgement–segment pair as an additional scalar feature.

#### **Cross-encoder:**

Recently, one of the most accurate methods of sentence comparison in IR tasks is cross-encoding. In a cross-encoder, two sequences are concatenated and sent in one pass to the sentence pair model, which is built on top of a Transformer-based language model. The attention heads of a Transformer can directly model which elements of one sequence correlate with which elements of the other, enabling the computation of an accurate relevance score [53]. We trained a cross-encoder built on top of the distilled version of the RoBERTa-base model [54] from the Huggingface library (<https://huggingface.co/distilroberta-base>, accessed on 19 June 2024). The hyperparameters we used for training are: batch size 16, num\_epochs 4, warmup\_step 10% of the training data, and a binary classification evaluator every 1000 steps. We trained the cross-encoder on both the augmented and non-augmented dataset.

#### **Cross Tension with In-batch Negative Sampling:**

To minimise the effect of random negative sampling in the augmented dataset, we experiment with an unsupervised learning approach with in-batch negative sampling. Adopting the contrasting learning (CT) from Carlsson et al. [55], we train two independent encoders on judgement–hearing segment pairs initialised with identical weights, where for each randomly selected segment  $s$ ,  $K$  irrelevant segments are sampled along with one relevant segment to create a  $K + 1$  batch as a training sample. The CT objective of the two independent encoders is to maximise the dot product between sentence representations of irrelevant segments and minimise the dot product between relevant ones. We hypothesise that using in-batch negative sampling gives a stronger training signal than the random shuffling of judgement–hearing segments in creating semantic representations. We initialise our two encoder models with distil-bert-base-uncased pretrained embeddings [54] from the Huggingface library (<https://huggingface.co/distilroberta-base>, accessed on 19 June 2024). We train the encoders for four epochs with a batch size of 16 segments with 300 max size tokens and a learning rate of  $5 \times 10^{-5}$ .

#### **GPT-3 Embedding Customisation**

To optimise the performance of our best-performing IR model, we customised GPT embeddings to better reflect the semantic characteristics of our legal dataset. The base GPT embedding model (text-embedding-ada-002) used is trained on diverse corpora including text search, text similarity, and code search tasks. To adapt it to the legal domain, we followed the embedding customisation approach proposed by OpenAI [56] and extended it with a transparent workflow tailored to our annotated legal data.

#### **Workflow Overview:**

1. We start with a set of human-annotated transcript–judgement pairs, labelled as either relevant (positive) or non-relevant (negative).
2. For each pair, we compute the cosine similarity between their original GPT embeddings.
3. We perform a threshold sweep over cosine similarity values  $x \in \{-1, -0.99, \dots, 1\}$  in increments of 0.01.

4. At each threshold  $x$ , we compute the standard error of the mean (SE) for the similarity scores of the positive and negative classes.
5. We identify the threshold  $x^*$  that minimises the standard error:

$$SE_{\min} = \min\{SE(x) \mid x \in [-1, 1]\} \quad (6)$$

6. Using the optimal threshold  $x^*$ , we train a linear transformation matrix  $M$  that maximises the separation between positive and negative pairs in the embedding space.
7. The customised embedding  $v'$  for each segment is computed as

$$v' = Mv \quad (7)$$

where  $v$  is the original GPT embedding and  $M$  is the learned transformation matrix.

To assess generalisability, we applied the customised embeddings to our training and held-out datasets of UK SC cases. The customised embeddings were used to train a regression model on both augmented and non-augmented datasets. Additionally, we experimented with incorporating the transformed cosine similarity scores as scalar features. Results of experiments with customised embeddings are explained in the following section.

## 4. Results

This section presents the evaluation results of our methods.

### 4.1. Results for Stage One

We evaluated our customised ASR model on two court cases with over 12 h of audio recording. We used the Word Error Rate (WER) which is a standard metric used to evaluate the performance of ASR systems. It quantifies the difference between a system-generated transcript and a reference (ground-truth) transcript. WER is defined as

$$WER = \frac{S + D + I}{N} \quad (8)$$

where

- $S$  = Number of substitutions (incorrect words);
- $D$  = Number of deletions (missing words);
- $I$  = Number of insertions (extra words);
- $N$  = Total number of words in the reference transcript.

The numerator represents the total number of errors, and the denominator normalizes this with the length of the reference transcript. A lower WER indicates better transcription accuracy.

Table 4 shows the WER scores and WER average scores for the two transcribed cases with different CLM system settings as well as the two baseline systems: AWS Transcribe (AWS base) and Whisper. The different CLM settings are as follows:

1. **CLM1** is trained on only the texts of the Supreme Court judgements.
2. **CLM2** is trained on both the judgements and the gold-standard transcripts.
3. **CLM2 + Vocab** uses CLM2 for transcription plus the global vocabulary list extracted by our phrase detection model.
4. **CLM2 + Vocab2** uses CLM2 for transcription plus the legal entities vocabulary list extracted by Blackstone and spaCy v3.4 library.

**Table 4.** Average WER and Transcription Time. Bold indicates the best scores.

Model	WER Case1	WER Case2	WER Average	Transcription Time
AWS base	8.7	16.2	12.3	85 min
CLM1	8.5	16.5	12.4	77 min
CLM2	<b>7.9</b>	15.5	<b>11.6</b>	77 min
CLM2 + Vocab	<b>7.9</b>	15.6	<b>11.6</b>	132 min
CLM2 + Vocab2	8.0	15.6	11.7	112 min
Whisper	9.6	<b>15.3</b>	12.4	191 min

As can be seen in Table 4, the ASR performance is consistently better with the CLM models than with the generic ASR systems for the two transcribed cases. CLM2 model, trained on textual data (i.e., the written judgements) and gold-standard court hearing transcriptions, outperforms AWS base and Whisper with 9% and 8% WER improvements, respectively. Moreover, we observe around a 9% improvement in average WER score over the two generic models when concatenating the list of legal phrases that is extracted by our phrase detection model with the CLM2 system. While ASR error correction indicates an improved transcription quality with our proposed domain adaptation methods, we also evaluated the ASR systems performance with specific errors such as legal entities and terms.

Table 5 shows the average ratio of correctly transcribed legal entities in the two studied courtroom hearings. We compare the performance of CLM2 infused with the legal terms list (CLM2 + Vocab) to the two generic ASR systems. The ratios in Table 5 indicate that CLM2 + Vocab is generally more capable of transcribing legal-specific terms than the other two models. It is also better at transcribing critical legal entities such as Provisions. A Provision, a statement within an agreement or a law, typically consists of alphanumeric utterances in British court hearings (e.g., “section 25(2)(a)–(h)” or “rule 3.17”). Such legal terminology needs to be accurately transcribed. Our CLM2 model with legal vocabulary demonstrates better reliability in transcribing these terms.

**Table 5.** Ratio of correctly captured legal entities by the ASR systems.

Entity	AWS BASE	Whisper	CLM2 + Vocab
Judge	0.66	0.77	0.84
CASE NAME	0.69	0.85	0.71
Court	0.98	1	0.93
Provision	0.88	0.95	0.97
Cardinal	1	0.97	1

A similar trend is evident with the legal entity Judge, which refers to the forms of address used in British courtrooms (e.g., “Lord Phillips”, “Lady Hale”). This entity is typically repeated in court hearings whenever a barrister or solicitor addresses the court. We see that both the generic ASR systems perform badly on this category with ratios of 0.66 and 0.69, respectively. On the other hand, we observe a significant improvement in correctly transcribing this type of entities by the CLM2 + Vocab with a ration of 0.84 correct transcriptions. Appendix A shows an example of the output of the AWS base ASR model without our domain-adaptation methods compared to the output of the CLM correcting the mistakes. The transcription errors (highlighted yellow) in the base output includes

legal jargon, legal terms and named entities. The errors are corrected by our CLM model (corrections are highlighted in blue).

In addition to evaluating the output of the ASR engines, we also recorded the time required to produce the transcription. The models based on AWS were run in the cloud using the Amazon infrastructure. Whisper was run on a Linux desktop with an NVIDIA GeForce RTX 2070 GPU with 8G VRAM. For all the experiments, the medium English-only model was used. As expected, the fastest running time is obtained using the AWS base model. Running the best-performing model increases the time by 155%, whilst Whisper more than doubles it. The trade-off between running time and the level of domain-specific accuracy is a variable parameter that can be determined based on the transcription purpose and the end-user needs defined by our project's commercial partner.

#### 4.2. Results for Stage Two

Table 6 shows the results of the different models on the test set extracted from the gold-standard dataset. As can be seen from the table, the concatenation of the GPT-3 customised embeddings (**GPT-3 Customised(+)**) for both the judgement and the hearing segments with their cosine similarity scores produce the best overall scores. Although the performance of a cross-encoder trained with the non-augmented dataset is best in extracting relevant judgement pairs with a recall of 0.93, its precision is much lower than the GPT-3 embeddings with and without data augmentation (**GPT-3 Customised(+)**, **GPT-3(-)**, respectively). Similarly, the recall of the Cross Tension (CT) bi-encoder with in-batch negatives is around 6% higher than the GPT-3 customised model, however, its precision is significantly lower. We have also conducted a bootstrap significance test to compare two binary classification models: GPT-3 with and without customised embeddings. The results yielded a  $p$ -value below 0.05 for both the F1 score and accuracy, indicating that the performance difference between the models is statistically significant. This suggests that customising the embeddings in GPT-3 leads to a meaningful improvement in classification performance, reinforcing the value of domain-specific adaptation in enhancing predictive accuracy and robustness.

**Table 6.** Results of classification models on augmented (+) and non-augmented (-) data. Bold indicates the best scores.

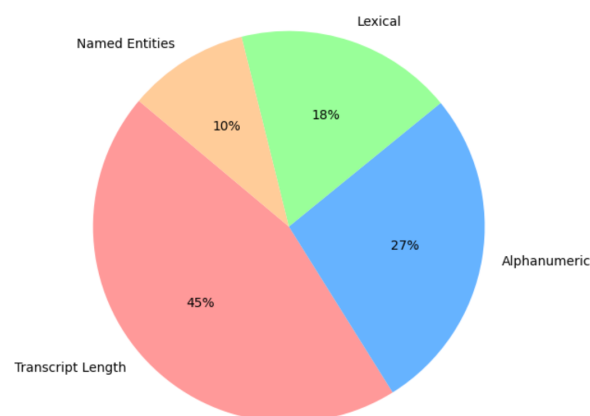
Model	Accuracy	Precision	Recall	F1
GPT-3(-)	0.69	0.84	0.64	0.73
GPT-3(+)	0.78	<b>0.85</b>	0.75	0.80
GPT-3(+) + cos_sim	0.83	0.91	0.79	<b>0.85</b>
GPT-3 Customised(+)	0.83	0.84	0.83	0.83
GPT-3 Customised(+) + cos_sim	<b>0.85</b>	<b>0.85</b>	0.84	<b>0.85</b>
Cross-encoder(-)	0.69	0.61	<b>0.93</b>	0.74
Cross-encoder(+)	0.81	0.79	0.84	0.81
CT with in-batch negatives	0.69	0.63	0.90	0.74

Moreover, generally speaking, the models' performance improves by augmenting the seed dataset with AI-generated samples. Since our aim is to extract as many relevant judgement–hearing links as possible from our UKSC cases, the GPT-3 customised embeddings with their similarity scores renders the best model for our use case.

## 5. Error Analysis and User Feedback

During the evaluation of our best retrieval system's performance on linking judgement paragraphs to corresponding court hearing transcript timestamps, our manual analysis of misclassified instances revealed four predominant error types. First, alphanumeric mismatches emerged when the system incorrectly deemed references as relevant due to superficial similarity. For instance, the transcript may cite "Section 193(1)" while the judgement refers to "Section 193(5)". According to our legal expert annotators, the two alphanumeric terms relate to entirely separate statutory provisions. Second, lexical repetition led to false positives where high-frequency terms in either the transcript or judgement skewed relevance predictions. For example, in a case concerned with an appeal of a tenant against his evacuation by the London Borough Council (Austin (FC) (Appellant) v Mayor and Burgesses of the London <https://www.supremecourt.uk/cases/uksc-2009-0037.html>, accessed on 19 June 2024), the lawyer in the transcript segment repeatedly refers to "arrear of rent". The relevancy model classified this segment as relevant to the introduction paragraph of the judgement whereas our expert annotator decided it was irrelevant as the lawyer is talking about a similar case and not the one brought before the court. We hypothesise that due to the frequency in the judgement segment of the phrase "arrear of rent" and other words from the same field (e.g., "tenancy of the premises", "paying rent", "Housing Act", etc.), the algorithm gave it a high semantic relevancy score. A commonly repeated word, even if contextually insignificant, was often over-weighted by the system.

Third, named entity sparsity posed challenges, particularly when critical entities appeared infrequently in the transcript. For example, "Mr. Ncqv", a senior director in a corporate group and a central figure in the case, was mentioned only once in the timestamp, resulting in a misclassification despite his legal significance. Lastly, transcription span length discrepancies contributed to errors; when the timestamp was considerably shorter than the corresponding judgement paragraph, the system struggled to establish relevance, often misclassifying the pair due to insufficient contextual overlap. Figure 4 shows the distribution of the error types in the analysed sample. Transcript length errors are the most frequent (50%), followed by alphanumeric (30%), lexical (20%), and named entity (10%) errors. These findings underscore the need for more nuanced handling of legal references, term weighting, entity salience, and span alignment in future iterations of the retrieval model.



**Figure 4.** Distribution of misclassification error categories.

In addition to this error analysis, we have also assessed the retrieval tool in collaboration with legal stakeholders to better understand its practical utility and alignment with user expectations. Evaluation of the system with real users showed that their productivity is dramatically increased when using the UI. In a preliminary study where we wanted to better understand the process of linking paragraphs to timestamps, a legal expert needed

15 h to identify 10 links without the help of any tool. When using our UI, the legal expert was able to validate 220 links in 3 h.

We have also tested our automatic linking system as a real-life tool by presenting the UI we created to a number of legal institutions in the UK. We chose entities who would potentially use the tool for a better access to justice. Accordingly, we conducted demos of the UI platform to the UK National Archives, the UK Supreme Court, and a number of industrial stakeholders in the field of legal AI. The tool and the objective behind its construction received positive feedback as well as interest in adopting it in a pipeline of a legal transcription software.

## 6. Discussion

The most direct benefit of linking of transcribed hearings and Supreme Court judgements is that it assists in understanding those judgements [7]. Written versions of the arguments (submissions) made by the advocates before the Supreme Court are not normally publicly available. Moreover, when judgements refer to arguments made by the parties, they do so in a selective, abbreviated, and editorialised form [24]. Thus, hearing recordings are the main source allowing external observers to learn the details of the arguments of the parties. In addition, the recordings of court hearings contain the questions and comments made by the judges, which may shed light on the contents of the judgement [23]. Given the systemic importance of Supreme Court decisions, such additional information about Supreme Court cases is likely to be helpful to academic researchers, practising lawyers, and even other judges aiming to understand the broader consequences of the case in question.

In this study, we illustrated the two stages of our system pipeline, which utilises generative AI to automatically connect written judgments from cases in the UK Supreme Court with their corresponding video recordings of their hearings. We employed NLP techniques to customise an ASR model specific of the Supreme Court acoustic tradition. The high-quality transcriptions of the court hearing sessions were used in building the Information Retrieval system that links important video bookmarks to the paragraphs in the final judgement text. Our IR system aids users in extracting relevant arguments and data to enhance their comprehension of the specific cases under examination.

While our system does not explicitly provide answers to legal professionals' inquiries regarding legal precedents, the UI developed in the project facilitates navigation and filtering of lengthy court hearing videos, allowing users to efficiently search through numerous timestamps. Subsequently, it offers a curated selection of essential bookmarks, crucial for grasping the judgment rendered in each case.

Beyond its utility for legal practitioners and scholars, this tool has broader implications, enhancing public access to court proceedings and fostering a deeper understanding of justice. Moreover, it opens avenues for new research inquiries, such as investigating correlations between courtroom proceedings and judicial decisions. Such analyses could shed light on the relationship between judges' statements during hearings and their ultimate rulings, as well as identifying effective advocacy strategies in influencing judicial outcomes.

## 7. Patents

Our UI is currently in the application process to be published as an innovative patent by the UK Intellectual Property Office (<https://www.gov.uk/government/organisations/intellectual-property-office>, accessed on 19 June 2024).

**Author Contributions:** Conceptualization, H.S. and C.O.; methodology, H.S. and C.O.; software, H.S. and C.O.; validation, H.S., C.B. and M.B.; formal analysis, H.S. and C.O.; investigation, H.S.;

resources, S.W.; data curation, H.S.; writing—original draft preparation, H.S.; writing—review and editing, H.S. and C.O.; visualization, H.S.; supervision, H.S. and C.O.; project administration, C.O.; funding acquisition, C.O., S.W. and C.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Innovate UK project reference 10022430.

**Data Availability Statement:** The datasets utilized for customising the ASR model and fine-tuning the GPT model for building the text-to-video IR system are accessible online. The datasets and model can be accessed at [https://github.com/sadanyh/Linking\\_Judgements](https://github.com/sadanyh/Linking_Judgements) (accessed on 23 October 2024).

**Conflicts of Interest:** Author Catherine Breslin was employed by the company Kingfisher Labs Ltd. Author Sophie Walker was employed by the company Just Access. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
STT	speech-to-text
WER	word error rate
IR	Information Retrieval
PMI	Point Mutual Information
CLM	Customised Language Model
SC	Supreme Court
LLM	Large Language Model
NLP	Natural Language Processing

## Appendix A

Examples of ASR output with and without domain-adaptation. Transcription errors are highlighted in yellow. Blue letters indicate the correct transcription.

	BASE MODEL	Names of people "Mrs Agbaje"	
Legal jargon "my learned"	with my lonely junior Miss Eleanor Harris. I appear on behalf of Mrs Ebadi, the appellate um, and then your friend Mr Timothy Scott, Queen's counsel, and Mr Peter Mitchell appear on behalf of Mr and Party, the respondent to the appeals. If I may, I will have dropped the Family Law convention of calling the party's husband and wife. The principal issue arising on this appeal is, in what circumstances is it appropriate for the English court to make a further financial order when a foreign court has already divorced the spouses and made a financial order? At first blush, it may seem unusual that courts in two different countries can, at different times make financial orders are rising out of the same cause of action, namely, the party's divorced. The power in English court to make a further financial order derives from Part three of the Matrimonial and Family Proceedings Act 1984 and it was intended by its frame as the law commissioners to remit financial hardship arising from two distinct circumstances. I suppose I should make it absolutely plain that this is the work of my predecessor. Yes, the Law Commission. I had no hand in part three at all. Yes, we saw the authors that listed in the report. The two circumstances that the law commissioners were addressing were these firstly, when a foreign court had made no financial order at all and secondly, where the foreign Court had made a financial order. But that order was inadequate and this appeal is concerned only with the second category. It is a condition precedent to the exercise of powers under Part three that the wife has to show in the circumstances of her case that a serious injustice has arisen. The serious injustice that this wife relies upon is that after a marriage of 32 years which produced five Children of the founded and where the assets were around £700,000 the Nigerian court in June of 2005 awarded her a lump sum of £21,000 and a life interest in the house on Tin Can Island in Lagos. This very modest award produced a very significant disparity or discrepancy in the allocation of the assets between the parties		Legal Terms "its framers"
Grammar mistakes "parties divorced"			
Vocabulary "children of the family"			

CLM MODEL Correcting Mistakes		Names
Legal jargon	with <b>my learned</b> Junior Miss Eleanor Harris. I appear on behalf of <b>Mrs Agbaje</b> . The appellate, um, my learned friend, Mr Timothy Scott, Queen's counsel, and Mr Peter Mitchell appear on behalf of Mr M. Party, the respondent to the appeal. If I may, I will have dropped the Family Law Convention of calling the parties husband and wife. The principal issue arising on this appeal is in what circumstances is it appropriate for the English court to make a further financial order when a foreign court has already divorced the spouses and made a financial order? At first blush, it may seem unusual that courts in two different countries can at different times make financial orders arising out of the same cause of action,	
Grammar	namely, <b>the parties divorced</b> . The power in the English court to make a further financial order derives from Part three of the Matrimonial and Family Proceedings Act 1984 and it was intended by <b>its framers</b> , the law commissioners, to remit financial hardship arising from two distinct circumstances. I suppose I should make it <b>absolutely plain</b> that this is the work of my predecessor. Yes, the Law Commission. I had no hand in part three at all. Yes, we saw the authors listed in the report. The two circumstances that the law commissioners were addressing were these firstly, when a foreign court had made no financial order at all and secondly, where the foreign Court had made a financial order. But that order was <b>inadequate</b> and this appeal is concerned only with the second category. It is a condition precedent to the exercise of powers under Part three that the wife <b>has to show</b> in the circumstances of her case that a serious injustice has arisen. The serious injustice that this wife relies upon is that after a marriage of 32 years which produced five <b>Children of the family</b> and where the assets were around £700,000 the Nigerian court in June of 2005 awarded her a lump sum of £21,000 and a life interest in the house on Tin Can Island in Lagos. This very modest award produced a very significant disparity or discrepancy in the allocation of the assets between the parties"	Legal terms
Vocabulary		

## Appendix B

### An Example of the Automatic Linking of Judgement Segment and Transcription Segments by GPT Embeddings

Legal Point	<p><b>Judgement:</b> The complaint, very moderately advanced by Mr. Geekie QC, is that a "starting point" of undesirability, placing the burden upon the person wishing to cross-examine a child to show some "particular justification" for doing so, gives insufficient weight to the Convention rights of all concerned. All the parties in care proceedings are entitled to a fair hearing in the determination of their civil rights and obligations the parents who stand to lose their children if allegations of abuse are made out, the children who stand to lose their parents if allegations of abuse are made out, but also stand to suffer abuse or further abuse if they are left at home because those allegations cannot be proved. And it is not only their article 6 rights which are in play. The civil rights in issue are also Convention rights in themselves the right to respect for the family lives of the parents and their children but also the right to respect for the private lives of the children, which include their rights to be protected from attacks upon their physical and psychological integrity. Even a "stranger" child, whose future is not in issue in the proceedings</p> <p><b>Timestamps:</b> 00:01:31.440,00:03:19.240,spk_2 (The Judge),</p>
Correct link	<p>Looking at what we say should be the starting point. Can I just consider, the issue of the welfare of the child and in the context and this feeds directly into my paragraph 47 where I say the welfare, of the child is irrelevant but not paramount consideration it links into Mr. Geekie's point about whether a non subject child, so a child who is simply a witness within proceedings has Article 8 rights engaged, so not as confident as Mr. Geekie on that. But what I would say is that if the court is considering a child as a witness and in no other capacity, we would say that that child's welfare is relevant and that probably actually answers the Article 8 point if that's engaged. But when we say welfare of a child is relevant, we mean for any child witness, whether a party or no. The way in which we had formulated the factors to be put into the balancing exercise came as is clear from paragraph 45 of the Medway case from the judgement of the Lady Justice Smith, where she went on to look at the various factors, she's put them in a much more generalized form.</p> <p>00:22:40.710,00:24:57.600,spk_1 (The lawyer),</p>
Correct link	<p>An equal starting point for all competing interests and the -- of course, one can't begin to list all competing interests that might arise, but judges can be cautioned or advised to be alert to the Article 6 and 8 rights of witnesses, and in the way that My Lady, we've canvassed already, that would apply to different -- witnesses in the three different positions before the Court. The Article 6 and 8 rights of parents, and the Article 6 and 8 rights of siblings. And if one sets those as separate categories as the appropriate path to begin thinking about, that at least gives a decision-making judge in an individual case the right ambit of decisions to consider. The present authorities rely significantly upon ....</p>

## Appendix C

### Example of Original and Paraphrased Transcript

Original Transcript (Transcript excerpt taken from *Austin v Southwark London Borough Council*, UK Supreme Court. Available at: <https://vlex.co.uk/vid/austin-v-southwark-london-841087421>).

*"My Lords. My Lady, In this matter, I appear with Mr Desmond Rutledge for the Appellant Mr Barry Austin, the Respondent, Southwark London Borough Council is represented by my learned friends Mr Drabble, Queen's Counsel, and Mr Shaw Kelly.*

*My Lords and My Lady will have seen that we have identified five issues on which we would invite the attention of this court and I propose to deal with them in the sequence set out in the agreed statement of facts and issues starting with issue one, which might probably be described as the tolerated trespass question. The question posed for this court by issue one is whether the possession order made against him ended the late Mr Allen Austin's tenancy or whether, in the alternative, he was at the date of his death, still the tenant of the council. That's the issue to which the first of the matters on this appeal goes."*

#### AI-Generated Paraphrasing

*"My Lords and My Lady, I appear before you with Mr Desmond Rutledge to represent the Appellant, Mr Barry Austin. The Respondent, Southwark London Borough Council, is represented by my learned friends Mr Drabble, Queen's Counsel, and Mr Shaw Kelly. We have identified five issues that we would like the court to address, and we will deal with them in the sequence set out in the agreed statement of facts and issues. The first issue is whether the possession order made against the late Mr Allen Austin ended his tenancy or whether, in the alternative, he was still the tenant of the council at the time of his death. This is the first matter on the appeal."*

## References

1. UK Supreme Court. Annual Report and Accounts 2024–25, 2025. Available online: [https://supremecourt.uk/uploads/The\\_Supreme\\_Court\\_Annual\\_Report\\_and\\_Accounts\\_2024\\_25\\_ec4ecbe2e5.pdf](https://supremecourt.uk/uploads/The_Supreme_Court_Annual_Report_and_Accounts_2024_25_ec4ecbe2e5.pdf) (accessed on 30 July 2025).
2. Saadany, H.; Breslin, C.; Orăsan, C.; Walker, S. Better Transcription of UK Supreme Court Hearings. In Proceedings of the Workshop on Artificial Intelligence for Access to Justice (AI4AJ 2023), Braga, Portugal, 19 June 2023; Volume 3435.
3. Saadany, H.; Orăsan, C. Automatic Linking of Judgements to UK Supreme Court Hearings. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, Singapore, 6–10 December; pp. 492–500.
4. Saadany, H.; Orăsan, C.; Walker, S.; Breslin, C. Linking Judgement Text to Court Hearing Videos: UK Supreme Court as a Case Study. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 10598–10609.
5. Elwany, E.; Moore, D.; Oberoi, G. Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. *arXiv* **2019**, arXiv:1911.00473. [CrossRef]
6. Nay, J.J. Natural Language Processing for Legal Texts. In *Legal Informatics*; Cambridge University Press: Cambridge, UK, 2021; pp. 99–113. [CrossRef]
7. Mumcuoğlu, E.; Öztürk, C.E.; Ozaktas, H.M.; Koç, A. Natural language processing in law: Prediction of outcomes in the higher courts of Turkey. *Inf. Process. Manag.* **2021**, *58*, 102684. [CrossRef]
8. Frankenreiter, J.; Nyarko, J. Natural Language Processing in Legal Tech. In *Legal Tech and the Future of Civil Justice*; Engstrom, D., Ed.; Cambridge University Press: Cambridge, UK, 2022.
9. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In Proceedings of the 40th International Conference on Machine Learning **2023** pp. 28492–28518.
10. Watanabe, S.; Mandel, M.; Barker, J.; Vincent, E.; Arora, A.; Chang, X.; Khudanpur, S.; Manohar, V.; Povey, D.; Raj, D.; et al. CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings. In Proceedings of the CHiME 2020-6th International Workshop on Speech Processing in Everyday Environments, Online, 4 May 2020.
11. Feng, S.; Kudina, O.; Halpern, B.M.; Scharenborg, O. Quantifying bias in Automatic Speech Recognition. *arXiv* **2021**, arXiv:2103.15122. [CrossRef]
12. Zhang, Y. Mitigating bias against non-native accents. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Incheon, Republic of Korea, 18–22 September 2022; Delft University of Technology: Delft, The Netherlands, 2022.
13. Mai, L.; Carson-Berndsen, J. Unsupervised domain adaptation for speech recognition with unsupervised error correction. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Incheon, Republic of Korea, 18–22 September 2022; pp. 5120–5124.
14. Huo, Z.; Hwang, D.; Sim, K.C.; Garg, S.; Misra, A.; Siddhartha, N.; Strohmaier, T.; Beaufays, F. Incremental layer-wise self-supervised learning for efficient speech domain adaptation on device. *arXiv* **2021**, arXiv:2110.00155.

15. Sato, H.; Komori, T.; Mishima, T.; Kawai, Y.; Mochizuki, T.; Sato, S.; Ogawa, T. Text-Only Domain Adaptation Based on Intermediate CTC. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Incheon, Republic of Korea, 18–22 September 2022; pp. 2208–2212.
16. Dingliwa, S.; Shenoy, A.; Bodapati, S.; Gandhe, A.; Gadde, R.T.; Kirchoff, K. Domain prompts: Towards memory and compute efficient domain adaptation of ASR systems. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Incheon, Republic of Korea, 18–22 September 2022.
17. Mani, A.; Palaskar, S.; Meripo, N.V.; Konam, S.; Metze, F. Asr error correction and domain adaptation using machine translation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6344–6348.
18. Del Rio, M.; Delworth, N.; Westerman, R.; Huang, M.; Bhandari, N.; Palakapilly, J.; McNamara, Q.; Dong, J.; Zelasko, P.; Jetté, M. Earnings-21: A practical benchmark for ASR in the wild. *arXiv* **2021**, arXiv:2104.11348. [[CrossRef](#)]
19. Wang, H.; Dong, S.; Liu, Y.; Logan, J.; Agrawal, A.K.; Liu, Y. ASR Error Correction with Augmented Transformer for Entity Retrieval. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Incheon, Republic of Korea, 18–22 September 2022; pp. 1550–1554.
20. Das, N.; Chau, D.H.; Sunkara, M.; Bodapati, S.; Bekal, D.; Kirchoff, K. Listen, Know and Spell: Knowledge-Infused Subword Modeling for Improving ASR Performance of OOV Named Entities. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 7887–7891.
21. Shukla, A.; Bhattacharya, P.; Poddar, S.; Mukherjee, R.; Ghosh, K.; Goyal, P.; Ghosh, S. Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv* **2022**, arXiv:2210.07544. [[CrossRef](#)]
22. Hellesoe, L.J. Automatic Domain-Specific Text Summarisation with Deep Learning Approaches. Ph.D. Thesis, Auckland University of Technology, Auckland, New Zealand, 2022.
23. Aletras, N.; Tsarapatsanis, D.; Preoțiuc-Pietro, D.; Lampos, V. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Comput. Sci.* **2016**, *2*, e93. [[CrossRef](#)]
24. Trautmann, D.; Petrova, A.; Schilder, F. Legal Prompt Engineering for Multilingual Legal Judgement Prediction. *arXiv* **2022**, arXiv:2212.02199. [[CrossRef](#)]
25. Hendrycks, D.; Burns, C.; Chen, A.; Ball, S. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv* **2021**, arXiv:2103.06268. [[CrossRef](#)]
26. Dixit, A.; Deval, V.; Dwivedi, V.; Norta, A.; Draheim, D. Towards user-centered and legally relevant smart-contract development: A systematic literature review. *J. Ind. Inf. Integr.* **2022**, *26*, 100314. [[CrossRef](#)]
27. Zheng, L.; Guha, N.; Anderson, B.R.; Henderson, P.; Ho, D.E. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, São Paulo, Brazil, 21–25 June 2021; pp. 159–168.
28. Rabelo, J.; Kim, M.; Goebel, R.; Yoshioka, M.; Kano, Y.; Satoh, K. COLIEE 2020: Methods for Legal Document Retrieval and Entailment. 2020. Available online: [https://sites.ualberta.ca/~rabelo/COLIEE2021/COLIEE\\_2020\\_summary.pdf](https://sites.ualberta.ca/~rabelo/COLIEE2021/COLIEE_2020_summary.pdf) (accessed on 19 June 2024).
29. Chalkidis, I.; Fergadiotis, M.; Manginas, N.; Katakalous, E.; Malakasiotis, P. Regulatory compliance through Doc2Doc information retrieval: A case study in EU/UK legislation where text similarity has limitations. *arXiv* **2021**, arXiv:2101.10726.
30. Kiyavitskaya, N.; Zeni, N.; Breaux, T.D.; Antón, A.I.; Cordy, J.R.; Mich, L.; Mylopoulos, J. Automating the extraction of rights and obligations for regulatory compliance. In Proceedings of the Conceptual Modeling-ER 2008: 27th International Conference on Conceptual Modeling, Barcelona, Spain, 20–24 October 2008; Proceedings 27; Springer: Berlin/Heidelberg, Germany, 2008; pp. 154–168.
31. Yujian, L.; Bo, L. A normalized Levenshtein distance metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1091–1095. [[CrossRef](#)]
32. Graves, A. Sequence transduction with recurrent neural networks. *arXiv* **2012**, arXiv:1211.3711. [[CrossRef](#)]
33. Guo, J.; Tiwari, G.; Droppo, J.; Van Segbroeck, M.; Huang, C.W.; Stolcke, A.; Maas, R. Efficient minimum word error rate training of RNN-Transducer for end-to-end speech recognition. *arXiv* **2020**, arXiv:2007.13802.
34. Rao, K.; Sak, H.; Prabhavalkar, R. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 193–199.
35. Bouma, G. Normalized (pointwise) mutual information in collocation extraction. *Proc. GSCS* **2009**, *30*, 31–40.
36. Řehůřek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, 22 May 2010; pp. 45–50. Available online: <http://is.muni.cz/publication/884893/en>.

37. Agirre, E.; Banea, C.; Cardie, C.; Cer, D.; Diab, M.; Gonzalez-Agirre, A.; Guo, W.; Mihalcea, R.; Rigau, G.; Wiebe, J. Semeval-2014 task 10: Multilingual semantic textual similarity. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 81–91.
38. Boteva, V.; Gholipour, D.; Sokolov, A.; Riezler, S. A Full-Text Learning to Rank Dataset for Medical Information Retrieval. In Proceedings of the 38th European Conference on Information Retrieval, Padua, Italy, 20–23 March 2016. Available online: <http://www.cl.uni-heidelberg.de/~riezler/publications/papers/ECIR2016.pdf>.
39. Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; Gurevych, I. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), Online, 6–14 December 2021. Available online: <https://openreview.net/forum?id=wCu6T5xFjeJ>.
40. Robertson, S.; Zaragoza, H.; et al. The probabilistic relevance framework: BM25 and beyond. *Found. Trends<sup>®</sup> Inf. Retr.* **2009**, *3*, 333–389. [[CrossRef](#)]
41. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
42. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation, 2014. Available online: <https://nlp.stanford.edu/projects/glove/>.
43. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *arXiv* **2020**, arXiv:2002.10957.
44. Williams, C. *Tradition and Change in Legal English: Verbal Constructions in Prescriptive Texts*; Peter Lang: Lausanne, Switzerland, 2007; Volume 20.
45. Haigh, R. *Legal English*; Routledge: London, UK, 2018.
46. Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; Androutsopoulos, I. LEGAL-BERT: The Muppets straight out of Law School. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 2898–2904. [[CrossRef](#)]
47. Hofstätter, S.; Lin, S.C.; Yang, J.H.; Lin, J.; Hanbury, A. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 11–15 July 2021.
48. Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; Yang, D. Is ChatGPT a general-purpose natural language processing task solver? *arXiv* **2023**, arXiv:2302.06476.
49. Wang, J.; Liang, Y.; Meng, F.; Shi, H.; Li, Z.; Xu, J.; Qu, J.; Zhou, J. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv* **2023**, arXiv:2303.04048. [[CrossRef](#)]
50. Törnberg, P. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv* **2023**, arXiv:2304.06588.
51. Cegin, J.; Simko, J.; Brusilovsky, P. ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness. *arXiv* **2023**, arXiv:2305.12947. [[CrossRef](#)]
52. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
53. Liu, F.; Jiao, Y.; Massiah, J.; Yilmaz, E.; Havrylov, S. Trans-encoder: Unsupervised sentence-pair modelling through self-and mutual-distillations. *arXiv* **2022**, arXiv:2109.13059.
54. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
55. Carlsson, F.; Gyllensten, A.C.; Gogoulou, E.; Hellqvist, E.Y.; Sahlgren, M. Semantic re-tuning with contrastive tension. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
56. Sanders, T. Customizing Embeddings, 2023. OpenAI. Available online: [https://github.com/openai/openai-cookbook/blob/main/examples/Customizing\\_embeddings.ipynb](https://github.com/openai/openai-cookbook/blob/main/examples/Customizing_embeddings.ipynb).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.