Linking Judgement Text to Court Hearing Videos: UK Supreme Court as a Case Study

Hadeel Saadany[*], **Constantin Orăsan**[*], **Mikolaj Barczentewicz**[*]
Catherine Breslin[†], **Sophie Walker**[**]
[*]University of Surrey, United Kingdom
hadil.saadany@gmail.com, {c.orasan, m.barczentewicz}@surrey.ac.uk
[†]Kingfisher Labs Ltd, United Kingdom
Just Access[**], United Kingdom

Abstract

One the most important archived legal material in the UK is the video recordings of Supreme Court hearings and their corresponding judgements. The impact of Supreme Court published material extends far beyond the parties involved in any given case as it provides landmark rulings on points of law of the greatest public and constitutional importance. Typically, transcripts of legal hearings are lengthy, making it time-consuming for legal professionals to analyse crucial arguments. This study focuses on summarising the second phase of a collaborative research-industrial project aimed at creating an automatic tool designed to connect sections of written judgements with relevant moments in Supreme Court hearing videos, streamlining access to critical information. Acting as a User-Interface (UI) platform, the tool enhances access to justice by pinpointing significant moments in the videos, aiding in comprehension of the final judgement. We make available the initial dataset of judgement-hearing pairs for legal Information Retrieval research, and elucidate our use of AI generative technology to enhance it. Additionally, we demonstrate how fine-tuning GPT text embeddings to our dataset optimises accuracy for an automated linking system tailored to the legal domain.

Keywords: Information Retrieval Dataset, Legal Information Retrieval, Embedding Customisation

1. Introduction

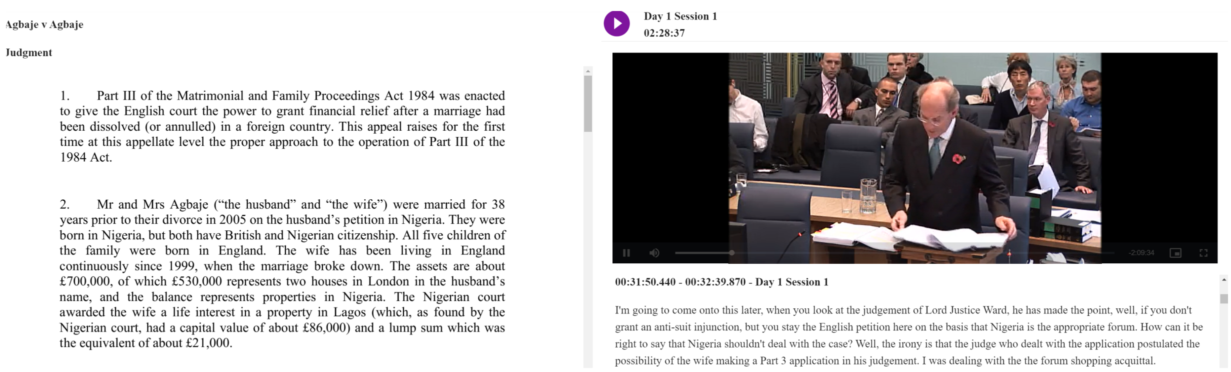
The UK Supreme Court (UKSC) shares live streams and archived recordings of courtroom proceedings, alongside written verdicts for resolved cases. This initiative primarily aims to enhance public accessibility to and comprehension of the UKSC's operations. As the highest court in the UK, the decisions made by the Supreme Court play a pivotal role in shaping the landscape of British law. Additionally, the court's rulings are invaluable for preparing new cases, offering direction for appeals, supporting legal education, and influencing future policy decisions.

Nonetheless, there are two primary challenges hindering the utilisation of this extensive resource to deepen understanding of the legal system and enhance access to justice. Firstly, the video content pertaining to a case often extends across numerous hours and multiple days, posing a significant time and effort investment for legal practitioners aiming to extract pertinent information tailored to their requirements. Secondly, the current demand for legal transcriptions is predominantly fulfilled by human transcribers and is restricted to a select number of UKSC cases (Sturge, 2021). Consequently, this renders the recorded material arduous to employ, whether in textual form or its original audio-visual presentation.

In this study, we explain the methodology employed during the second phase of a collaborative research-industrial endeavour aimed at developing an integrated system for automatically navigating segments within UKSC hearing media data, based on their semantic correlation with specific paragraph(s) in the corresponding judgement text. Leveraging the timing metadata of court hearing transcription segments, we successfully embed bookmarks within video sessions, linking them to semantically relevant paragraphs in the judgement text. The primary goal of this video bookmarking process is to furnish legal professionals and the general public with an automated navigation tool. This tool aims to pinpoint arguments and legal precedents articulated during lengthy hearing sessions, which are particularly pertinent to the judges' decision-making process in the case.

We also release an annotated dataset of UKSC case judgements linked to their relevant court hearing session. This dataset is unique for two reasons. First, it is a document-to-document IR dataset that establishes links between documents from two different linguistic registers: written (the judgement text) and spoken (the video transcription). Linguistically, spoken language is characterised by complex sentence structures with low lexical density (fewer high content words per clause), whereas written language typically contains simple sentence structures with high lexical density (more high content words per clause) (Peters, 2003; Matthiessen and Halliday, 2009). Moreover, the complexity of speech relates to meta-linguistic elements such as intonation, loudness or quietness, pausing, stress, pitch range and gestures communicating semantic con-

LREC-COLING 2024, pages 10598–10609
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



Agbaje v Agbaje

Judgment

1. Part III of the Matrimonial and Family Proceedings Act 1984 was enacted to give the English court the power to grant financial relief after a marriage had been dissolved (or annulled) in a foreign country. This appeal raises for the first time at this appellate level the proper approach to the operation of Part III of the 1984 Act.

2. Mr and Mrs Agbaje ("the husband" and "the wife") were married for 38 years prior to their divorce in 2005 on the husband's petition in Nigeria. They were born in Nigeria, but both have British and Nigerian citizenship. All five children of the family were born in England. The wife has been living in England continuously since 1999, when the marriage broke down. The assets are about £700,000, of which £530,000 represents two houses in London in the husband's name, and the balance represents properties in Nigeria. The Nigerian court awarded the wife a life interest in a property in Lagos (which, as found by the Nigerian court, had a capital value of about £86,000) and a lump sum which was the equivalent of about £21,000.

Day 1 Session 1
02:28:37

00:31:50.440 - 00:32:39.870 - Day 1 Session 1

I'm going to come onto this later, when you look at the judgement of Lord Justice Ward, he has made the point, well, if you don't grant an anti-suit injunction, but you stay the English petition here on the basis that Nigeria is the appropriate forum. How can it be right to say that Nigeria shouldn't deal with the case? Well, the irony is that the judge who dealt with the application postulated the possibility of the wife making a Part 3 application in his judgement. I was dealing with the the forum shopping acquittal.

Figure 1: User-Interface for Linking Judgement to Bookmarks in Video Court Sessions

notations (Halliday, 2007) . Thus, the linking task in our case is nontraditional as it needs careful preprocessing and segmentation of the spoken and the written datasets to establish accurate semantic relevance.

Second, the annotation of this dataset is much more complicated than the typical annotation of semantic text similarity benchmark datasets like the ones developed for the Semantic Text Similarity shared tasks (Cer et al., 2017; Conneau and Kiela, 2018). In the latter, the annotators usually do not require domain-knowledge to determine whether there is semantic linking or not. In contrast, in our use case finding a semantic link between the judgement and the court hearing deliberations entails expert knowledge of UK law.

To give one example, figure 1 shows a snapshot of the UI we created where a paragraph in the judgement (on the left side) is linked to a particular timespan in the court hearing video (on the right side)[1]. The judgement segment is deemed relevant to the court hearing segment as the judgement refers to 'PART III of the Matrimonial and Family Proceedings Act 1984' which provides the English court the power to make amendments to a financial settlement between a married couple after their marriage has been dissolved in a foreign country [2]. The lawyer in the video segment is addressing the jurisdictional requirements of this Act by trying to refute a previous judgement's 'anti-suit injunction'. The 'anti-suit injunction' means that the judgement restrains his client from bringing a claim before the UK courts to amend the inadequate financial provision on her divorce that was issued by a Nigerian court. The semantic relevancy annotation of such segments requires legal-domain knowledge as well as an understanding of the legal terms used. Our research project funded the hiring of post-graduate law researchers to provide relevancy annotations between judgement segments and court hearing transcripts for our compiled datasets.

To summarise, our contributions in this research are:

(a) We introduce an application of Doc2Doc IR where the queries and documents are typically long and come from two distinct linguistic modes, written and spoken, with legal-specific jargon and vocabulary.

(b) We compile and release our first publicly available dataset of UKSC judgement-hearing with gold-standard relevancy annotations, suitable for legal IR as well as Doc2Doc IR in general[3].

(c) We show that the GPT 3 text embeddings produce the best results with respect to the IR document representations and their domain customisation improves the post-fetching results.

To explain the process we followed to build our judgement-hearing linking system and the compilation of our dataset, this paper is divided as follows: section 2 shows how we compiled and preprocessed our dataset using a zero-shot IR approach as a pre-fetching stage. In section 3, we explain our experiments in training an IR system on the compiled dataset and will show that domain-customisation of the GPT latest text embedding model on our dataset produces the best accuracy for our relevancy linking model. In section 4, we conduct an error analysis on a sample of the IR system output and we present the feedback received from stakeholders on the linking tool. In sections 5, we briefly summarise relevant literature in the field of legal NLP in general and legal IR in particular. Finally, in section 6, we write our conclusion on the experiments conducted as well as our future work for improving the linking system.

[1] Due to space limitations, only part of the court hearing transcript is visible in the figure

[2] Matrimonial and Family Proceedings Act 1984

[3] `https://github.com/surrey-nlp/Linking-Judgements`

2. Data Compilation

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 a custom speech-to-text language model we developed in stage one of the project (Saadany et al., 2022) and then segment the judgement into paragraphs. Paragraph(s) 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 $C = \{c_1, c_2, ..., c_n\}$, the task is to find the timespans in the video transcripts $T = \{t_1, t_2, ..., t_n \,|\, t_i \in C \wedge (t_i, q)\}$ 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.

We extracted 7 UKSC case judgements consisting of 1.4M tokens scraped from the official site of the UKSC[4]. The transcription data consists of 53 hours of video material for the selected cases obtained from the UK National Archive[5]. The video sessions were transcribed by our custom speech-to-text model. Next, we ran several preprocessing steps to obtain the best linking accuracy between a judgement segment and the relevant timespans in the transcripts.

2.1. Data Processing and Preparation

The main challenge in preprocessing the dataset was how to segment the judgement text into semantically cohesive sections that would be treated as queries in our IR method. We noticed that typically the 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. 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[6]

2.2. Stage 1: 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. Agirre et al., 2014; Boteva et al., 2016; Thakur et al., 2021). In order to create a human-generated 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 Supreme Court case with, for example, 50 judgement segments and 300 timespans of video transcript, the annotators will need to read 50 x 300 judgement-timespan link, which amounts to 15,000 document-to-document link 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, post-graduate law student, to evaluate the first 20 links produced by the model. 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 done using a specially designed interface which 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 (Robertson et al., 2009): 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,

[4] https://www.supremecourt.uk/decided-cases/

[5] https://www.nationalarchives.gov.uk/

[6] 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.

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

Table 1: Results of Unsupervised IR for linking Judgements to Video Transcripts in One Case

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

Table 2: Results of Unsupervised IR for linking Judgements to Video Transcripts for All Cases

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 (Pennington et al., 2014a) pretrained word embeddings. The GloVe vector embeddings are created by unsupervised model training on general domain data (Pennington et al., 2014b). We create vectors for the judgement segment and the transcripts spans from the mean, minimum and maximum values of the GloVe embeddings.

Entailment Search: We use embeddings from a pretrained model for textual entailment which is trained to detect sentence pair relations, i.e. one sentence entails or contradicts the other. We employed the Microsoft MiniLM model (Wang et al., 2020b) 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 or not 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 (Williams, 2007; Haigh, 2018). For this reason, we employed the Legal BERT (Chalkidis et al., 2020) 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, 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 (Hofstätter et al., 2021) 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 GPT latest text-embedding-ada-002 model which outperforms GPT previous most capable model, Davinci, at most tasks [7]. The context length of the ada-002 model is increased by a factor of four, from 2048 to 8192, making it more convenient to 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.

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:

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

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

[7] https://openai.com/blog/new-and-improved-embedding-model

the model. We calculated MAP for the first 5, 10, and 15 judgement-timespan pairs.

As can be seen from Table 1, the GPT model demonstrated the best performance in comparison to the other models. 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. The human annotators were again asked to judge whether the extracted timespan transcripts are semantically linked or not linked to the judgement paragraph(s). The human annotations were compared to the results of all the embedding models mentioned above.

Again as shown in Table 2, the GPT text embedding model shows superiority over the other models. Thus, the approach of treating the judgement segment as a query and the transcription of the video sessions as the corpus in which we try to find the answer gives the best MAP results for the first 5, 10 and 15 links. It should be pointed out that our use case is different than a typical IR task where the efficacy of the model is evaluated by its ability to get 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-user of our UI can watch or draw the cursor around the bookmarks to get 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. Thus, the zero-shot retrieval by the GPT 3 model was chosen for annotation.

After annotation, we compiled a dataset of 3620 judgement-to-transcript documents annotated with gold-standard similarity labels. In the next section, we describe our method for augmenting this data.

2.4. Stage 2: 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 that was used in the preliminary experiments (Saadany and Orăsan, 2023). The augmented dataset are used in this research along with the gold-standard for training a relevancy model. Recently, several research studies have managed to successfully use ChatGPT prompt engineering as an aiding tool for several NLP tasks (e.g. Qin et al. (2023); Wang et al. (2023); Törnberg (2023)). One successful use of prompt engineering has been the use of ChatGPT as a substitute for crowd-sourced paraphrasing. Research has shown that ChatGPT-generated paraphrases are lexically and syntactically more diverse than human-generated ones (Cegin et al., 2023).

Accordingly, we used the InstructGPT API *set role* prompt strategy to extract paraphrases for the transcript side of the positive instances in our dataset (Ouyang et al., 2022). 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:*

The paraphrases were created by the text-davinci-002 model 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 sample of the AI-generated paraphrases by a legal expert in our research team to make sure the paraphrased transcript reflect the same meaning as the original.

In order to generate negative samples, we adopted two techniques. 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$42M tokens. We used both the gold-standard and the augmented datasets to build a judgement-hearing relevancy model. Our experiments are explained in the following section.

3. Experiments

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 relevancy 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 show our experiments with training several models on both our gold-standard seed and our augmented dataset. The evaluation of each model is based on the provided gold-standard labels for each judgement-hearing segment.

3.1. Baseline Model

For our baseline, we train a logistic regression model with the GPT 3 embedding representations with and without data augmentation. We use the concatenated vectors of each judgement-segment pair in one setting and in another we add the cosine-similarity as a scalar feature.

3.2. Cross-encoder

Recently, one of the most accurate methods of sentence comparison in IR tasks is the cross-encoding. In a cross-encoder, two sequences are concatenated and sent in one pass to the sentence pair model, which is built atop 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 (Liu et al., 2022). We trained a cross-encoder built on top of the distilled version of the RoBERTa-base model (Sanh et al., 2019) from the Huggingface library[8]. 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.

3.3. 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 Carlsson et al. (2020) contrasting learning (CT), 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 (Sanh et al., 2019) from the Huggingface library[9]. 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 e^{-5}.

3.4. GPT3 Embedding Customisation

To optimise the performance of the best IR model, we customised the GPT embeddings to be more domain specific. The GPT embedding model used for our retrieval is trained on different datasets used for text search, text similarity, and code search. In order to customise the GPT embeddings to our legal dataset, we follow the OpenAI method for embedding customisation (Sanders, 2023). We train a classification model on our human-annotated data with the following objective:

$$SE_{\mathsf{min}} = \min \mathsf{SE}(x) \mid x \in \{-1, -0.99, \ldots, 1\} \quad (2)$$

where x is the cosine similarity threshold between the positive and negative class which we obtain by sweeping between cosine similarity scores from -1 to 1 in steps of 0.01 to get the lowest standard error of mean SE_{min} for the cosine similarity distribution. The output of this training is a matrix M that we multiply by the embedding vector v of each judgement and transcript segment. This multiplication produces customised embeddings which are more adapted to our legal dataset relevancy distribution. After training, the customised GPT embeddings reduced the overlap between the relevant and irrelevant judgement-hearing links from 70.5% ± 2.7% to 73.0% ± 2.6%. We used the customised embeddings in training a regression model on both our augmented and non-augmented dataset. We also experimented with incorporating the customised cosine similarity scores as a scalar feature.

3.5. Results

Table 3 shows the results of the different models on a test set with gold-standard labels. As can be seen from the table, the concatenation of the GPT 3 customised embeddings for both the judgement and the hearing segments with the 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. 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. Moreover, generally speaking, the model's performance improves by augmenting the seed dataset with AI-generated samples. Since our aim is to extract as many relevant judgement-hearings links as possible from our UKSC cases, the GPT 3 customised embeddings with their similarity scores renders the best model for our use case.

[8] https://huggingface.co/distilroberta-base

[9] https://huggingface.co/distilroberta-base

Model	Accuracy	Precision	Recall	F1
GPT 3(-)	0.69	0.84	0.64	0.73
GPT 3(+)	0.78	**0.85**	0.75	0.80
GPT 3(+) + cos_sim	0.83	0.91	0.79	**0.85**
GPT 3 Customised(+)	0.83	0.84	0.83	0.83
GPT 3 Customised(+) + cos_sim	**0.85**	**0.85**	0.84	**0.85**
Cross-encoder(-)	0.69	0.61	**0.93**	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

Table 3: Results of Relevancy Models on Augmented (+) and non-Augmented (-) Dataset

4. Error Analysis and User Feedback

We conducted error analysis on a sample of the judgement-hearing segments that had a high relevancy score by our best model, but was deemed irrelevant by our human annotators. We noticed that one of the main causes was phrases that appear with high frequency in both the judgement's paragraph and transcript's segments. For example, in a case concerned with an appeal of a tenant against his evacuation by the London Borough Council [10], the lawyer in the transcript segment repeatedly refers to "arrears 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 "arrears 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.

We have also tested our automatic linking system as a real-life tool by presenting the UI we created to a number of stakeholders. 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 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.

5. Related Literature

Recently, there has been a great interest in utilising Natural Language Processing (NLP) methodologies to enhance the analysis of legal texts (Elwany et al., 2019; Nay, 2021; Mumcuoğlu et al., 2021; Frankenreiter and Nyarko, 2022). This interest primarily revolves around tasks such as summarising legal documents (Shukla et al., 2022; Hellesoe, 2022), predicting judicial outcomes (Aletras et al., 2016; Trautmann et al., 2022), and preprocessing and generating contracts (Hendrycks et al., 2021; Dixit et al., 2022). Furthermore, NLP techniques for extracting information and determining entailment in legal contexts have been widely applied (Zheng et al., 2021). These methods aim to either locate answers to legal inquiries within legal documents (Zheng et al., 2021) or establish connections between textual data (Rabelo et al., 2020).

For instance, Chalkidis et al. (2021) conducted experiments employing various Information Retrieval (IR) models to extract relevant legislative acts from the European Union (EU) and the United Kingdom (UK), crucial for organisations' regulatory compliance. They found that fine-tuning a BERT model for a specific classification task within their domain yielded optimal results. Additionally, there is a growing body of research exploring the utilisation of large language models in legal contexts (Trautmann et al., 2022; Yu et al., 2022; Choi et al., 2023; Blair-Stanek et al., 2023; Trozze et al., 2023; Guha et al., 2023).

In the realm of speech-to-text research, the transcripts of oral hearings at the United States Supreme Court (SCOTUS) have been extensively analysed. It has been demonstrated that the content of oral arguments can be indicative of the final decisions made by the SCOTUS (Shullman, 2004; Epstein et al., 2010; Black et al., 2011; Dietrich et al., 2019; Kaufman et al., 2019). Moreover, researchers have employed machine learning techniques to investigate implicit gender bias in SCOTUS hearings (Rabelo et al., 2020).

While legal Information Retrieval (IR) research has seen a notable advancement in recent times, the utilisation of spoken court hearings for legal IR has not been given comparable focus to comprehending and extracting data from textual legal sources. In our study, we present a commercial solution utilising IR techniques to seamlessly link judgements with video recordings of court proceedings.

6. Conclusion

The most direct benefit of linking of transcribed hearings and Supreme Court judgements is that it

[10] Austin (FC) (Appellant) v Mayor and Burgesses of the London https://www.supremecourt.uk/cases/uksc-2009-0037.html

assists in understanding those judgements. 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, editorialised form. 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. 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 introduced the second phase of our system pipeline, utilising generative AI to automatically connect written judgments from cases in the UK Supreme Court with their corresponding video recordings of hearings. Our information retrieval (IR) system aids users in extracting relevant arguments and data to enhance their comprehension of the specific cases under examination. While our system doesn't explicitly provide answers to legal professionals' inquiries regarding legal precedents, the user interface (UI) 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.

Finally, in future stages, we aim to expand our annotated linking dataset and explore the effectiveness of coupling judgements and video hearings according to common legal entities such as articles, legal provisions and names of similar cases.

7. Bibliographical References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Ryan C. Black, Sarah A. Treul, Timothy R. Johnson, and Jerry Goldman. 2011. Emotions, oral arguments, and supreme court decision making. *The Journal of Politics*, 73(2):572–581.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning?

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Proceedings of the 38th European Conference on Information Retrieval*. `http://www.cl.uni-heidelberg.de/~riezler/publications/papers/ECIR2016.pdf`.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In *International conference on learning representations*.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. Chatgpt to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. *arXiv preprint arXiv:2305.12947*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. 2021. Regulatory compliance through doc2doc information retrieval: A case study in eu/uk legislation where text similarity has limitations. *arXiv preprint arXiv:2101.10726*.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Journal of Legal Education (Forthcoming)*.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Nilaksh Das, Duen Horng Chau, Monica Sunkara, Sravan Bodapati, Dhanush Bekal, and Katrin Kirchhoff. 2022. Listen, Know and Spell: Knowledge-Infused Subword Modeling for Improving ASR Performance of OOV Named Entities. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7887–7891. IEEE.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jetté. 2021. Earnings-21: a practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*.

Bryce J. Dietrich, Ryan D. Enos, and Maya Sen. 2019. Emotional arousal predicts voting on the u.s.supreme court. *Political Analysis*, 27(2):237–243.

Saket Dingliwa, Ashish Shenoy, Sravan Bodapati, Ankur Gandhe, Ravi Teja Gadde, and Katrin Kirchhoff. 2022. Domain prompts: Towards memory and compute efficient domain adaptation of ASR systems. In *Interspeech 2022*.

Abhishek Dixit, Vipin Deval, Vimal Dwivedi, Alex Norta, and Dirk Draheim. 2022. Towards user-centered and legally relevant smart-contract development: A systematic literature review. *Journal of Industrial Information Integration*, 26:100314.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Emad Elwany, Dave Moore, and Gaurav Oberoi. 2019. Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. *arXiv preprint arXiv:1911.00473*.

Lee Epstein, William M. Landes, and Richard A. Posner. 2010. Inferring the winning party in the supreme court from the pattern of questioning at oral argument. *The Journal of Legal Studies*, 39(2):433–467.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Jens Frankenreiter and Julian Nyarko. 2022. Natural language processing in legal tech. *Legal Tech and the Future of Civil Justice (David Engstrom ed.)*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean

Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

Jinxi Guo, Gautam Tiwari, Jasha Droppo, Maarten Van Segbroeck, Che-Wei Huang, Andreas Stolcke, and Roland Maas. 2020. Efficient minimum word error rate training of rnn-transducer for end-to-end speech recognition. *arXiv preprint arXiv:2007.13802*.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Rupert Haigh. 2018. *Legal English*. Routledge.

Michael Alexander Kirkwood Halliday. 2007. *Language and Education: Volume 9*. A&C Black.

Lui Joseph Hellesoe. 2022. *Automatic Domain-Specific Text Summarisation With Deep Learning Approaches*. Ph.D. thesis, Auckland University of Technology.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proc. of SIGIR*.

Oleksii Hrinchuk, Mariya Popova, and Boris Ginsburg. 2020. Correction of automatic speech recognition with transformer sequence-to-sequence model. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 7074–7078. IEEE.

Zhouyuan Huo, Dongseong Hwang, Khe Chai Sim, Shefali Garg, Ananya Misra, Nikhil Siddhartha, Trevor Strohman, and Françoise Beaufays. 2021. Incremental layer-wise self-supervised learning for efficient speech domain adaptation on device. *arXiv preprint arXiv:2110.00155*.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiangyang Li, Edward Lin, and Tie-Yan Liu. 2021. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. *Advances in Neural Information Processing Systems*, 34:21708–21719.

Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2022. Trans-encoder: Unsupervised sentence-pair modelling through self-and mutual-distillations. *arXiv preprint arXiv:2109.13059*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Long Mai and Julie Carson-Berndsen. 2022. Unsupervised domain adaptation for speech recognition with unsupervised error correction. *Proc. Interspeech 2022*, pages 5120–5124.

Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. 2020. Asr error correction and domain adaptation using machine translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348. IEEE.

Christian MIM Matthiessen and Michael Alexander Kirkwood Halliday. 2009. Systemic functional grammar: A first step into the theory.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Emre Mumcuoğlu, Ceyhun E Öztürk, Haldun M Ozaktas, and Aykut Koç. 2021. Natural language processing in law: Prediction of outcomes in the higher courts of turkey. *Information Processing & Management*, 58(5):102684.

Mahdi Namazifar, John Malik, Li Erran Li, Gokhan Tur, and Dilek Hakkani Tür. 2021. Correcting automated and manual speech transcription errors using warped language models. *arXiv preprint arXiv:2103.14580*.

John J. Nay. 2021. *Natural Language Processing for Legal Texts, DOI=10.1017/9781316529683.011*, page 99–113. Cambridge University Press.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014a. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014b. Glove: Global vectors for word representation. Figshare `https://nlp.stanford.edu/projects/glove/`.

J Melanie Peters. 2003. *The Impact of Tele-Advice on the Community Nurses' Management of Leg Ulcers*. University of South Wales (United Kingdom).

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

J Rabelo, MY Kim, R Goebel, M Yoshioka, Y Kano, and K Satoh. 2020. Coliee 2020: Methods for legal document retrieval and entailment, 2020. *URL: https://sites. ualberta. ca/˜ rabelo/COLIEE2021/COLIEE_2020_summary. pdf*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kanishka Rao, Haşim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 193–199. IEEE.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Hadeel Saadany, Catherine Breslin, Constantin Orăsan, and Sophie Walker. 2022. Better transcription of uk supreme court hearings. *arXiv preprint arXiv:2211.17094*.

Hadeel Saadany and Constantin Orăsan. 2023. Automatic linking of judgements to uk supreme court hearings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 492–500.

Ted Sanders. 2023. Customizing embeddings. OpenAI `https://github.com/openai/openai-cookbook/blob/main/examples/Customizing_embeddings.ipynb`.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Hiroaki Sato, Tomoyasu Komori, Takeshi Mishima, Yoshihiko Kawai, Takahiro Mochizuki, Shoei Sato, and Tetsuji Ogawa. 2022. Text-Only Domain Adaptation Based on Intermediate CTC. *Proc. Interspeech 2022*, pages 2208–2212.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv preprint arXiv:2210.07544*.

Sarah Levien Shullman. 2004. The illusion of devil's advocacy: How the justices of the supreme court foreshadow their decisions during oral argument. *J. App. Prac. & Process*, 6:271.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Georgina Sturge. 2021. Court statistics for England and Wales. Technical report, House of Commons Library.

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. `https://openreview.net/forum?id=wCu6T5xFjeJ`.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction.

Arianna Trozze, Toby Davies, and Bennett Kleinberg. 2023. Large language models in cryptocurrency securities cases: Can chatgpt replace lawyers?

Ramya Vunikili, Hitesh Ochani, Divisha Jaiswal, Richa Deshmukh, Daniel L Chen, and Elliott Ash. 2018. Analysis of vocal implicit bias in scotus decisions through predictive modelling. *Proceedings of Experimental Linguistics*.

Haoyu Wang, Shuyan Dong, Yue Liu, James Logan, Ashish Kumar Agrawal, and Yang Liu. 2020a. ASR Error Correction with Augmented Transformer for Entity Retrieval. In *Interspeech*, pages 1550–1554.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. 2020. CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *CHiME 2020-6th International Workshop on Speech Processing in Everyday Environments*.

Christopher Williams. 2007. *Tradition and change in legal English: Verbal constructions in prescriptive texts*, volume 20. Peter Lang.

Guangwei Xu, Yangzhao Zhang, Longhui Zhang, Dingkun Long, Pengjun Xie, and Ruijie Guo. 2022. Hybrid retrieval and multi-stage text ranking solution at trec 2022 deep learning track. *TREC 2022 Deep Learning Track*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer.

Li Yujian and Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095.

Yuanyuan Zhang. 2022. Mitigating bias against non-native accents. *Delft University of Technology*.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. 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*, pages 159–168.