CLosER: Conversational Legal Longformer with Expertise-Aware Passage Response Ranker for Long Contexts

Arian Askari
a.askari@liacs.leidenuniv.nl
Leiden University

Mohammad Aliannejadi
m.aliannejadi@uva.nl
University of Amsterdam

Amin Abolghasemi
m.a.abolghasemi@liacs.leidenuniv.nl
Leiden University

Evangelos Kanoulas
e.kanoulas@uva.nl
University of Amsterdam

Suzan Verberne
s.verberne@liacs.leidenuniv.nl
Leiden University

ABSTRACT

In this paper, we investigate the task of response ranking in conversational legal search. We propose a novel method for conversational passage response retrieval (ConvPR) for long conversations in domains with mixed levels of expertise. Conversational legal search is challenging because the domain includes long, multi-participant dialogues with domain-specific language. Furthermore, as opposed to other domains, there typically is a large knowledge gap between the questioner (a layperson) and the responders (lawyers), participating in the same conversation. We collect and release a large-scale real-world dataset called LegalConv with nearly one million legal conversations from a legal community question answering (CQA) platform. We address the particular challenges of processing legal conversations, with our novel **C**onversational **L**egal L**o**ngformer with Expertise-Aware Res**pons**e **R**anker, called CLosER. The proposed method has two main innovations compared to state-of-the-art methods for ConvPR: (i) Expertise-Aware Post-Training; a learning objective that takes into account the knowledge gap difference between participants to the conversation; and (ii) a simple but effective strategy for re-ordering the context utterances in long conversations to overcome the limitations of the sparse attention mechanism of the Longformer architecture. Evaluation on LegalConv shows that our proposed method substantially and significantly outperforms existing state-of-the-art models on the response selection task. Our analysis indicates that our Expertise-Aware Post-Training, i.e., continued pre-training or domain/task adaptation, plays an important role in the achieved effectiveness. Our proposed method is generalizable to other tasks with domain-specific challenges and can facilitate future research on conversational search in other domains.

CCS CONCEPTS

• **Information systems** → **Learning to rank**; **Novelty in information retrieval**.

This work is licensed under a Creative Commons Attribution International 4.0 License.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3614812

KEYWORDS

Conversational Legal Search, Conversational Search for Long context, Response Ranking in Legal Domain

ACM Reference Format:
Arian Askari, Mohammad Aliannejadi, Amin Abolghasemi, Evangelos Kanoulas, and Suzan Verberne. 2023. CLosER: Conversational Legal Longformer with Expertise-Aware Passage Response Ranker for Long Contexts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614812

1 INTRODUCTION

The development of information retrieval (IR) systems that can converse with people naturally and continuously is a popular research topic [3, 18, 73]. There are two main approaches to designing conversational search systems: response selection (our focus) and response generation (e.g., ChatGPT). In this paper, we investigate *conversational passage response retrieval (ConvPR) for the legal domain.* We propose a novel method for passage response selection called CLosER that takes into account the long conversation contexts and different levels of expertise of the conversation participants. To empirically study the problem and evaluate our model, we collect a large-scale real-world dataset from a legal CQA platform, called LegalConv.

ConvPR is a passage response selection task in which the next utterance is retrieved from a pool of candidate responses, given the conversation context. In ConvPR, each utterance is a passage, thus the next utterance is a longer text than typical utterances in chit-chat dialogues [73]. Examples of ConvPR datasets are MSDialog [70], AliMe [26], TREC CAsT [42], and UbuntuDialogue [35], which are mainly synthetic or crawled human–human conversations on CQA web sites. On CQA platforms, users pose questions or problems and other users respond with answers or solutions.

An effective Legal ConvPR system could benefit lawyers, questioners, and service providers. Given a new utterance by the questioner, and the conversation context, the Legal ConvPR system would retrieve the most likely next responses from the CQA collection. This way, questioners would receive preliminary information by receiving an instant ranked list of the most relevant lawyer responses which could even satisfy questioners' information need before the lawyers respond to their questions. This is also beneficial for service providers in reducing response time in conversations and boosting customer satisfaction. Moreover, the questioners can

quickly get familiarized with the lawyers, based on the retrieved responses which could then result in hiring lawyers by questioners.

Our novel dataset is a large-scale real-world collection of lawyer-questioner conversations, crawled from the open CQA website Avvo[1]. The collection contains over 7.6 million context–response candidate pairs that were extracted from ~ 360K questions and ~ 780K responses. Our data analysis reveals that there are three particular challenges to conversational search in the legal domain compared to other domains: (i) **Length:** The context and responses in the legal domain tend to be substantially longer than in other ConvPR datasets. (ii) **Legal language:** The content of conversations in the legal domain is different from other domains. The legal text has distinct characteristics and is usually classified as a sub-language [20, 57, 65]. The legal text has terms that are uncommon in generic corpora (e.g., 'restrictive covenant,' 'tort') and terms that have different meanings than in general language (e.g., 'alien'). (iii) **Expertise gap:** Legal CQA contains a wide range of categories, from bankruptcy to criminal defense, and a single lawyer has deep expertise only in a few categories while having shallow expertise in the other categories. As a result, in a conversation, we have text written by lawyers with diverse levels of expertise and by a questioner, who is a layperson [9, 30]. Figure 1 illustrates this with an example of a conversation in the legal domain.

While a legal ConvPR model could offer many benefits to legal CQA platforms, as well as other conversational search tasks with long contexts, to the best of our knowledge, there is neither scientific work that studies ConvPR in the legal domain nor available datasets for studying this task. Moreover, there are no pre-trained publicly available Transformers-based models in the legal domain for long legal texts with more than 512 tokens. Furthermore, the existing ConvPR models do not take into account the *expertise gap* between users of legal CQA platforms, as described above.

To address these limitations, in this work, we propose a novel conversational legal response ranker called CLosER, with which we address the above-mentioned challenges of the domain, as well as the limitations of existing work.

Aiming to address the *expertise gap* challenge, with CLosER, we propose a novel multi-task learning objective, called Expertise-Aware Post-Training (EA-PT). This training step after pre-training aims to model the expertise gap in the legal CQA content through an auxiliary utterance classification task. Secondly, we use Longformer to address the *length* challenge — modeling long conversational context. We pre-train and evaluate a Legal Longformer model on a dataset of 7 million case law documents to address the *legal language* challenge. This gives us the ability to feed up to 4096 tokens into the input (8 times longer than BERT's input). However, Longformer suffers from the local attention mechanism, which causes the input words to only attend to a local window of tokens, except for one global token. In order to overcome this limitation, we propose a novel utterance re-ordering strategy that adapts the order of the context's utterances and moves the utterances with a higher lexical matching score closer to the candidate response. This way, we provide more relevant content from the conversation context closer to the candidate utterance.

Our main contributions in this work are four-fold:

- We investigate and address the task of conversational search in the legal domain and release a large-scale real-world test collection for the task consisting of 7.6M context and response candidate pairs that are extracted from ~ 360K questions and ~ 780K responses.[2]
- We benchmark the dataset with state-of-the-art passage response selection methods to the legal domain and highlight their limitations.
- Motivated by our data analysis, we propose CLosER to address the limitations of existing methods by incorporating domain characteristics using a novel expertise-aware learning objective, followed by an utterance re-ordering strategy.
- We evaluate CLosER in extensive experiments against the state-of-the-art passage retrieval methods. Furthermore, we conduct a thorough analysis of the performance of CLosER to demonstrate the effectiveness of the expertise-aware learning objective and our proposed utterance re-ordering strategy.

2 RELATED WORK

Legal IR tasks and data. Research in legal IR mainly focuses on high-recall, full document retrieval tasks such as case law retrieval [1, 4–8, 24, 37, 47, 53] and eDiscovery [45, 61, 62]. Both case law retrieval and eDiscovery are tasks of legal professionals [51] that are very different from conversational legal search: In case law retrieval, a lawyer aims to identify all relevant prior judicial decisions for a new legal case [34]. eDiscovery is an evidence-finding task in the context of civil litigation [41]. The limited prior work addressing conversational search in the legal domain focuses exclusively on case law data [31, 32], which is of a completely different nature than online CQA data. Commonly used datasets for legal IR are COLIEE (Canadian case laws) [47, 47–49], AILA (Indian case laws) [12, 43], CAP, and CASELAW (United States case laws) [33, 46]. However, all of these are case law data. Work on user-generated content in the legal domain is sparse. Recently, Askari et al. [9] address this gap by investigating the task of expert finding in legal community question answering (i.e., finding the right lawyer to answer a given question). However, their goal is different from ours (identifying experts) and the dataset used is limited to only 10,000 legal questions from one category (bankruptcy). We aim to fill this gap by proposing a large-scale, diverse, and user-generated legal dataset — users being both professional lawyers and laypersons.

Furthermore, to the best of our knowledge, no studies have addressed conversational search in the legal domain based on real-world data. The most relevant work to ours in the same domain is the work by John et al. [23], which proposes a response generator bot to advise users on their everyday legal questions [56] by utilizing a small number of training instances (2400) of question-answer pairs taken out of short, formal legal information from an online law textbook. They use a sequence-to-sequence model from the pre-transformer age aimed at utterance generation. It cannot deal with longer context dependencies and does not address any domain-specific challenges. We instead tackle the response selection task by taking into account domain challenges and using roughly *one million real-world legal dialogues* from a legal CQA platform as

[1] https://www.avvo.com/

[2] For privacy reasons, we do not store the real lawyer names.

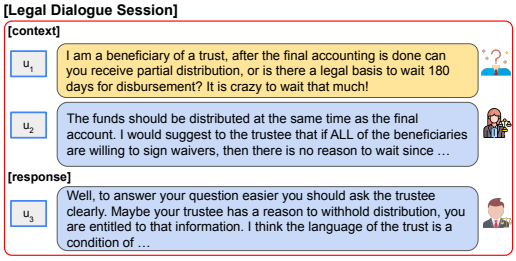

Figure 1: An example of a conversation in the legal domain. We show 3 out of the original 20 turns in the conversation in this example. u_i refers to i–th utterance. u_1, is written by the questioner. u_2 is written by a lawyer with deep expertise on the topic, and u_3 by lawyer who has shallow expertise on this topic.

Table 1: Data statistics for our dataset (LegalConv). C and R refer to Context and Response respectively, and C-R pairs refer to Context-Response pairs.

Items	Train	Valid	Test
# C-R pairs	5,592,330	11,312,900	10,671,500
# cand. per C	10	100	100
# + cand. per C	1	1	1
Min # turns per C	2	2	2
Max # turns per C	38	22	22
Avg # turns per C	5.5	6.1	5.3
Max # words per C	3875	3060	5031
Avg. # words per C	494	518	422
Max. # words per R	3853	3060	3060
Avg. # words per R	486	499	409

training data. The dialogues in our dataset are on average 8 times longer than the dataset that is created by [23].

ConvPR methods. Conversational passage response retrieval is the task of retrieving the actual next utterance (i.e., response) after the given context from a pool of candidate utterances. In ConvPR, each utterance is a passage, i.e., a longer text than is common for utterances in chit-chat dialogues' [73]. There is a variation on ConvPR called Open Retrieval ConvPR (OR-ConvPR) that defines an end-to-end ConvPR [73], retrieving the next utterance from a large collection of passages rather than a sampled set from the pool [39, 71, 73]. However, in this work, we focus on the re-ranking stage of ConvPR and leave OR-ConvPR for future work.

In prior work, ConvPR has been approached from various directions. Lu et al. [36] propose RoBERTa-SS-DA, a participant segmentation approach, discriminating different participants, and applying a dialogue augmentation based on RoBERTa. Whang et al. [63] propose BERT-DPT, a model that applies domain-adaptive post-training (DPT). Post-training is an auxiliary learning task between pretraining and finetuning that helps the model pick up domain-specific characteristics of the downstream task [63]. Gu et al. [19] propose SA-BERT, which incorporates speaker-aware embedding in the model; therefore, it is aware of the speaker change information. Additionally, there are prior works that use Multi-Task Learning to improve the effectiveness of multi-turn Response Selection. Whang et al. [64] propose $\text{UMS}_{\text{BERT+}}$, a multi-task learning framework consisting of three tasks (i.e., utterance insertion, deletion, and search). Xu et al. [68] propose BERT-SL, which introduces four self-supervised tasks, including session-level matching, utterance restoration, incoherence detection, and consistency classification. They train the response selection model with these auxiliary tasks in a multi-task manner.

In terms of competitive baselines, the most relevant model for our task is the Fine-Grained Post-Training model (FP) [21] since it is the state-of-the-art on a task that is somewhat similar to ours: response ranking on the Ubuntu Dialogue dataset [35]. Our proposed method is closely related to works that focus on post-training and multi-task fine-tuning for response selection such as [21, 63, 64, 68], but as opposed to the prior work, our method leverages post-training as an auxiliary task to distinguish different levels of expertise, and contains an innovative mechanism to effectively use long contexts.

Transformers for long sequences. Transformer-based architectures [59], such as BERT, have yielded improvements in many IR and NLP tasks. However, the time and memory complexity of the self-attention mechanism in these architectures is $O(L^2)$ over a sequence of length L [11, 72], which makes them unsuitable for long texts. For that reason, the maximum input length of BERT-based models is commonly restricted to 512 tokens. Various attention mechanisms have been proposed to handle the large memory consumption of the attention mechanisms. Longformer [11] introduces a localized sliding window-based mask with few global masks to reduce computation; this allows for the encoding of longer sequences. BigBird [72] uses additional random attention to approximate full attention. Longformer has shown better effectiveness compared to BigBird on domain-specific tasks [27]. Xiao et al. [67] propose Lawformer, a pre-trained Longformer for Chinese legal long documents understanding and show its gains. Mamakas et al. [38] build on top of Longformer and propose a model to handle even longer texts up to 8096 tokens. They add a [*SEP*] token at the end of each paragraph and fine-tune Longformer on the LexGLUE legal benchmark tasks without pre-training it. In this work, in order to overcome the input length limitation of BERT, we pre-train Longformer on the English Legal domain with 4096 tokens as the maximum input length. We call it *LegalLongformer*, before post-training and fine-tuning on the downstream task. To the best of our knowledge, our model is the first pre-trained Longformer on Legal English content as [38] only fine-tunes Longformer on the downstream tasks and [67] focus on the Chinese Language. In addition, we overcome the limitations of the local attention window in LongFormer through our Utterance Re-ordering method.

3 DATASET

In this section, we describe how we collect LegalConv and its characteristics. Moreover, we analyze the data to provide more insights into this problem. Our dataset is based on a legal online community question-answering website, based in the U.S., called Avvo. To comply with Avvo's terms of use while facilitating data access for the community, we make the hyperlinks publicly available, as opposed to directly publishing the dataset raw data. This approach aligns with our objective of honoring the necessary legal and ethical considerations while providing accessibility to the information.

Our data handling approach is similar to prior research on public online datasets such as Twitter [10, 13, 54]. To be more specific, we obtain the data from the public Internet Archive[3] using the Heritrix library [40]. Heritrix is designed to respect the 'robots.txt' exclusion directives and META 'nofollow' tags. We provide hyperlinks to Avvo web pages stored in the Internet Archive, along with a bash script to collect them using Heritrix, and a HTML parser that handles modifications made to the Avvo website over a span of 15 years. We store our data in two formats: raw HTML and structured JSON, enabling us to retain the flexibility of extracting additional information in the future. We found that five percent (around 350,000 question posts out of roughly 7 million questions) of Avvo's posts are preserved in the Internet Archive which is fairly large enough for our study compared to other datasets on ConvPR tasks such as MSDialog [70] with 35,000 questions and UbuntuDialogue [35].

Data collection. We crawl 359,708 dialogues and each legal conversation has a category that is related to the lawyers' expertise. A lawyer who has written the next utterance could have a deep or shallow expertise in the category of conversation [9, 16]. Examples of categories are "Bankruptcy," "Child Custody," and "Business." We filter out dialogues that have fewer than three utterances. After that, we split the data into training, validation, and test partitions chronologically. Specifically, the training data contains 166,166 dialogues from 2007-11-07 to 2015-05-20; the validation data contains 30,002 dialogues from 2015-05-20 to 2016-03-04; the test data contains 34,617 dialogues from 2016-03-04 to 2022-07-15. In case of duplicate questions, we kept the one that received the most responses and removed the other duplicates. The temporal split ensures that we avoid data leakage and that the evaluation setting reflects the real-world scenario in which we use historical data to train models that we can apply to more recent data.

Next, we follow previous research in other domains to generate the conversation (i.e., dialogue) context and response candidates [35, 60, 70]. For each dialogue in Avvo, we have one layperson user — the questioner— who poses the questions that lead to the information-seeking conversation, and one or multiple professional lawyers who provide responses. The *next utterance* in our data refers to a lawyer's response. *Context* refers to the previous utterances before the *next utterance*; each of the utterances from the context could be written by a lawyer or the questioner.

Data preparation for next utterance prediction. We keep the conversations with more than three utterances, and for each utterance by the questioner i, q_i^t, we collect the previous c utterances as the dialog context, where $c = t - 1$ and is the total number of utterances before q_i^t. In constructing the data we follow established approaches from prior works that have constructed datasets based on online CQA platforms for ConvPR tasks in different domains, such as UbuntuDialog and MSDialog [35, 76]. The true *next utterance* by the lawyer following the questioner's question becomes the positive (true) response candidate. For the negative (false) response candidates, we adopt negative sampling following previous work [70]. For each dialogue context, we first use the true *next utterance* as the query to retrieve the top 1,000 results from the whole candidate response set with BM25. Then we randomly sample n responses from them to construct the negative response candidates, For training, we use $n = 9$, while for validation, and test set we use $n = 99$. We select a larger number of negative responses for the test and validation set in order to make the task closer to the real-world setting by setting a larger re-ranking depth on the test set. This is in contrast to some of the prior work in which only 9 negative responses for both the train and test sets are selected [35, 70].

Lawyers' profiles on Avvo are identified with their real names, as opposed to questioners, who are anonymous. The questions are organized into categories and each category (e.g., 'bankruptcy') includes questions with different category tags (i.e., 'bankruptcy homestead exemption'). For Expertise-Aware Post-Training, we classify lawyers based on their level of expertise. We define two levels of expertise, namely, shallow and deep. We mark lawyers to have deep expertise on a category tag when two conditions are met [9]: (i) *Engagement filtering:* similar to the definition proposed in [9, 58], a lawyer should have ten or more of their answers marked as accepted by the questioner on a category and more than the average number of best answers among lawyers on that category tag. A 'best' answer is either labeled as the most useful by the question poster or more than three lawyers agree that the answer is useful. (ii) Following the idea proposed in [69], the acceptance ratio (the count of best answers divided by the count of answers) of an expert lawyer's answers should be higher than the average acceptance ratio on a category in the test collection. If a lawyer's profile does not satisfy either of these conditions, it is classified as a lawyer with shallow expertise in a certain category. As mentioned, we consider questioners to be users with zero expertise. Based on these two conditions, out of 25,931 active lawyers, 2,307 have a deep expertise in at least one category and collectively have provided 398,861 answers out of 782,165.

Data statistics and comparison to other datasets. In order to statistically analyze the length of the conversations, we compare the conversation context and response length of LegalConv with two additional comparable passage response selection datasets: UbuntuDiaog and MSDialog [35, 76]. MSDialog and UbuntuDiaog are automatically created by extracting dialogues from questions and answers that are respectively about Microsoft products and Ubuntu issues. The average length of responses of LegalConv is about 10 times longer than UbuntuDialogue and MSDialog corpora.

4 PROPOSED METHOD: CLosER

We address the task of Legal ConvPR in two main steps: (i) Expertise-Aware Post-Training; and (ii) Utterance Re-Ordering Fine-tuning. In the following, we first define the task and then describe the two aforementioned steps in detail.

4.1 Task Definition

We assume that the dataset $D = (c_i, r_i, y_i)_{i=1}^{N}$ is made up of N triples, each of which including the context c_i, response r_i, and ground-truth label $y_i \in 0, 1$. Here, the context is a sequence of previous utterances, that is $c_i = \{u_1, u_2, ..., u_M\}$, and M is the number of utterances in the context. The j_{th} utterance $u_j = \{w_{j,1}, w_{j,2}, ..., w_{j,L}\}$ where L is the number of words in the utterance u_j. Each candidate response (i.e., candidate next utterance), r_i, is a single utterance. $y_i \in 0, 1$ denotes the relevance label of a given triple where $y_i = 1$ indicates that r_i is the true (actual) next response for the context c_i;

[3] http://archive.org

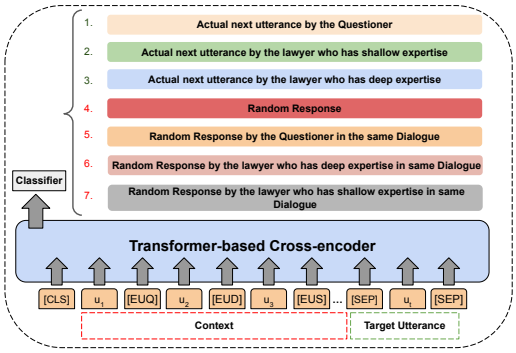

Figure 2: Expertise-Aware Post-Training Architecture.

otherwise, $y_i = 0$ which means the response is a random response. Our main objective is to train a model to predict y_i given c_i and r_i correctly. We then create the ranking list by sorting the candidate responses based on their corresponding probability for $y_i = 1$.

4.2 Expertise-Aware Post-Training

EA-PT in a multi-task optimization setting to address the expertise gap in the legal ConvPR task (see Section 1 and Figure 1). EA-PT aligns with studies on lawyer performance that show the information provided by lawyers has the highest quality when they have deep expertise in the field, while they can provide useful but less accurate information in other fields, as they are not experts in those fields [52]. The EA-PT component of CLoSER enables it to learn these differences in the level of expertise of a lawyer as a domain-level characteristic.

With EA-PT, the model learns to classify the relationship between the target utterance u and the input context into seven classes (see Figure 2). The distinction between these seven classes is motivated by the different levels of expertise: if we fine-tuned ConvPR models only on a binary classification task between random and actual next utterances, it would be challenging to distinguish the actual next utterances by a lawyer with shallow expertise from random (unrelated) utterances. We make two assumptions here: (i) the next utterances by lawyers with shallow expertise are less related to the given conversation context than the utterances by lawyers with deep expertise; and (ii) the utterances by lawyers with shallow expertise are more related to the conversation than random utterances. We argue that through the seven-class post-training phase, before fine-tuning, the model first learns to differentiate the expertise level of lawyers, and better capture random responses from the next utterances that are written by lawyers with shallow and deep expertise during fine-tuning.

Input format. We use transformer-based cross-encoders, in which the input format is (*[CLS], sequence A, [SEP], sequence B, [SEP]*), where [CLS] and [SEP] are separator tokens.[4] To do the post-training, we construct the input using sequence A as the context and sequence B as a candidate response. Moreover, inspired by prior work introducing utterance separator tokens [21, 63], we add three new tokens (Figure 2) in order to differentiate within utterances of the context. With our new separator tokens, we assist the model to better differentiate between different levels of expertise of the writers of utterances that appear in the context and compare them to the candidate utterance for predicting the expertise level of the response. Below, we describe these tokens:

- *[EUQ]*: End of utterance by questioner. This indicates that the previous utterance has been written by the questioner who is considered a layperson user with zero expertise in any legal conversation.
- *[EUD]*: End of utterance by a lawyer who has deep expertise on the category of conversation.
- *[EUS]*: End of utterance by a lawyer who has shallow expertise on the category of conversation.

Figure 2 shows our proposed post-training method, which takes into account data characteristics by classifying a pair of context and response into seven classes. The choice of classes motivated as follows:

- **Distinguishing expertise levels:** By classes $1-3$, the model differentiates the expertise level of utterances' writers: zero expertise for questioner, and shallow or deep expertise for the lawyer, given a conversation category. An utterance by the questioner could be the target utterance only during post-training in order to model the legal conversational language. During fine-tuning and inference, the actual next utterance can only be a response by lawyer. Moreover, classes $5-7$ emphasize differentiating the level of expertise even for a random utterance (from the *same dialogue*). By the *same dialogue*, we mean that the utterance is not the actual next utterance, but it appears in the same dialogue in later turns.
- **Distinguishing relevance levels:** Classes $4-7$ represent different random classes which give the model the ability to distinguish within the different levels of semantic relevance, e.g., an utterance that is not the actual next utterance but is in the same dialogue is less random than a random utterance from a different dialogue.

Training setup. Given a pair of target utterances u and the context c in the input (Figure 2), the model classifies the relationship between the context and the target utterance into seven classes. As input for the classification, the embedding of the *[CLS]* token is used. The final score $y'(c,u)$ is obtained by feeding *[CLS]* through a single-layer perception, and the degree of relevance between the context and target utterance is obtained through the score. To calculate the EA-PT loss, we use the cross-entropy loss, which is formulated as follows:

$$loss_{\text{EA-PT}} = -\sum_{i}^{7} y_i log(y'(c,u)_i)\,, \tag{1}$$

where 7 is the number of classes. To train the proposed model in the multi-task optimization setting, we use the Masked Langauge Modeling (MLM) and EA-PT together in a multi-task optimization setup with equal weight for loss of each task as follows:

$$loss = loss_{\text{EA-PT}} + loss_{\text{MLM}}\,. \tag{2}$$

The goal of MLM is to adapt the model for legal language conversations compared to other domains.

[4]We utilize the notation of BERT input throughout the paper for consistency while the separator of Longformer is <s> and </s> that corresponds to [CLS] and [SEP]

4.3 Fine-tuning with Utterance Re-Ordering

As described in Section 3, the long contexts in legal conversations cause more than 50% of LegalConv not to fit in the input of BERT (maximum input of BERT is 512 tokens). Therefore, we utilize the Longformer architecture to be able to encode 4096 tokens in the input. However, Longformer suffers from local attention as it has only one global token, <s>, to which all tokens attend, and each of the rest of the tokens only attends to a specific window of previous and next tokens. The size of this window is 256 [11], which means that each token only has full attention to 256 next and previous tokens. Therefore, the candidate response can fully attend only the utterances that are closer to it.

We turn this weakness of Longformer, having only local attention, to advantage with utterance re-ordering. The intuition is that we bring the context's utterances that are likely to be more relevant to the candidate response closer to the candidate response so that the self-attention mechanism can relate tokens from the candidate response to the most relevant responses in the context. To estimate what the most relevant prior responses are, we use BM25 [50] in the same setting of BM25 baseline explained in Section 4.4: we compute the lexical relevance score between each utterance in the context and the candidate response and sort the utterances based on their BM25 score.

Mathematically, we compute the BM25 relevance score between a candidate response r and an utterance u_i as follows:

$$\text{BM25}(r, u_i) = \sum_{t \in r \cap u_i} rsj_t . \frac{tf_{t,u_i}}{tf_{t,u_i} + k_1\{(1-b) + b\frac{|u_i|}{l}\}}, \quad (3)$$

where t is a term of response r, $tf_{t,d}$ is the frequency of t in utterance u_i, rsj_t is the Robertson-Spärck Jones weight [50] of t, and l is the average utterance length. k_1 and b are parameters with default values [28, 29]. By dynamically reordering the utterances in the context based on their BM25 scores, we effectively prioritize the most relevant information for each candidate response.

This strategy, while simple, proves to be effective in directing the self-attention mechanism toward the most relevant information. After post-training, we further fine-tune the model on the downstream task using a binary classification objective, while incorporating the utterance re-ordering methodology.

4.4 Baselines

We compare our proposed model, CLosER, with the following previous models: BERT-based and Longformer-based, Fine-Grained Post-Training (FP) [21] which is state-of-the-art on the Ubuntu conversational passage response selection dataset, and Domain Post-Training (DPT) [63]:

- **BM25:** We use BM25 with the default parameter values $k = 1.2$ and $b = 0.75$ for lexical retrieval. We concatenate utterances of context and use them as queries for BM25.
- **BERT-based and Lonformer-based models [11, 14, 17]:** we fine-tune BERT-based and Longformer-based models with a linear combination layer stacked atop the [CLS] token with cross-entropy loss and the Adam optimizer using classic input format: ([CLS], sequence A, [SEP], sequence B, [SEP]), where sequence A is context and sequence B is candidate response. We concatenate utterances of context.
- **Fine-Grained Post-Training (FP):** Han et al. [21] applies Fine-Grained Post-Training with an Utterance Relevance Classification Post-Training Method that classifies the target utterance into three classes given a context: (i) next utterance, (ii) random utterance, (iii) random utterance from the same dialogue. Next, they fine-tune the post-trained model to the response selection task.
- **Domain Post-Training (DPT):** Whang et al. [63] propose a model that applies Domain Post-Training (DPT). The model is post-trained with BERT's pre-training methods, MLM, and Next Sentence Prediction (NSP), and then fine-tuned to the response selection task.

Both DPT and FP models differentiate the utterances of context with a [*EOU*] token as the separator. In Section 6, we refer to FP and DPT models by adding them as suffixes to the initial checkpoint name, e.g., Longformer FP refers to the Longformer base model that is trained with the FP method.

5 EXPERIMENTAL SETUP

Pre-training Longformer. We pre-train Longformer on legal case laws, which we call Legal Longformer, with the MLM objective in advance of EA-PT. We do so to: (i) have a fair comparison between Longformer and BERT since Legal BERT is publicly available; and (ii) train Longformer to better process the legal domain compared to the general web domain. We start from the *Longformer-base-4096* [11] released checkpoint. For pre-training the Longformer on legal language, we do not use our dataset, LegalConv, as we want to train Longformer to understand general legal language before training CLosER. It also allows us to fairly compare Legal Longformer with Legal BERT as none of them have been pre-trained on LegalConv. As data for pre-training, we randomly select 250K, which covers our 65K training steps, case laws out of 7 million case laws in 75 categories of the Caselaw Access Project (CAP) dataset [46]. We select the CAP dataset because it is larger and more diverse than the other publicly available datasets (see Section 2). We pre-train Legal Longformer for 65,000 steps following the original Longformer paper and employ early stopping, based on the Masked Language Modeling Bits per Character (MLM BPC) value on a validation set of 10,000 case laws. MLM BPC measures the average number of bits needed to encode a character in an unsupervised manner and a lower MLM BPC for a model means that the model captures the text. We use the same hyperparameters as the original Longformer [11]: we set the batch size 64 (2^{18} tokens), the maximum learning rate 3×10^{-5}, a linear warmup of 500 steps, followed by a power 3 polynomial decay, and a local attention window size of 512 [11]. Table 2 shows the evaluation result of Longformer on 100,000 case laws that are selected randomly from [46] as the test set, in terms of MLM BPC before (Longformer) and after (Legal Longformer) pre-training. The results show that after pre-training, Legal Longformer could capture the legal language much better than Longformer. Furthermore, the achieved BPC is slightly higher than the MLM BPC reported in the original Longformer paper on the general web dataset, which is 1.71. This shows that processing legal language is more challenging than general web data.

Post-training dataset. In order to Post-Train CLosER on Legal content, we balance our classes. Our aim is to make a relatively large

Table 2: MLM BPC for Longformer and Legal Longformer with Legal and General (original) Tokenizer.

	Steps	MLM BPC
Longformer [11]	65k	9.82
Legal Longformer **(ours)**	65k	**2.55**

train, evaluation, and test set to train and evaluate post-training. The number of negative instances of train, validation, and test set is: 5,033,097, 11,199,771, and 10,564,785 in our data for response selection, which we round up and under-sample to 450,000, 100,000, and 150,000 instances for post-training. Next, we perform over-sampling for classes with fewer instances (classes 2 – 7). This results in 3,150,000, 700,000, and 1,050,000 instances for all classes in the training, validation, and test set of our post-training dataset. For oversampling, we randomly select from the instances of the target class until we reach the number of equal instances across classes. The post-trained CLosER achieved an accuracy of 63% for the utterance classification task on the post-training test set.

Training configuration and model parameters. We use the Huggingface library [66], and PyTorch [44] for training and inference of Transformer-based models. Following prior work [22] we use the Adam [25] optimizer with a learning rate of $7 * 10^{-6}$ for all cross-encoder layers, regardless of the number of layers trained. We truncate the longer context and candidate response from the left and right sides respectively. We cap context and candidate response to 256 tokens in BERT. For Longformer, we cap context and candidate response to 3586 and 512 tokens respectively. We employ early stopping, based on the sum of all evaluation metrics on the validation set. We use a training batch size of 32.We utilize the publicly available original implementation of Fine-Grained Post-Training (FP) [21] and DPT [63] baselines in order to replicate them with Legal BERT and Legal Longformer.[5]

Evaluation. We evaluate the returned ranking of utterances, based on the rank of the true next utterance. For the evaluation metrics, we use mean average precision (MAP), Recall@1, Recall@2, and Recall@5 following prior work [35, 70] in addition to NDCG@5. MAP is equivalent to the mean reciprocal rank (MRR) since there is only one positive response candidate per dialog context [70].

6 EXPERIMENTAL RESULTS

In this section, we answer the following research questions, assessing the effectiveness of our proposed method, CLosER, from different perspectives:

- **RQ1:** What is the effectiveness of CLosER compared to the existing state-of-the-art Longformer-based and BERT-based models?
- **RQ2:** How robust are response ranker models (ours and the state-of-the-art) against the different levels of expertise of the lawyers producing the response utterances?
- **RQ3:** To what extent is CLosER's performance affected by the number of training steps of the EA-PT phase?
- **RQ4:** What is the impact of Utterance Re-ordering on the effectiveness of CLosER and how does this impact relate to the context length?

[5]The code is available on https://github.com/arian-askari/CLosER

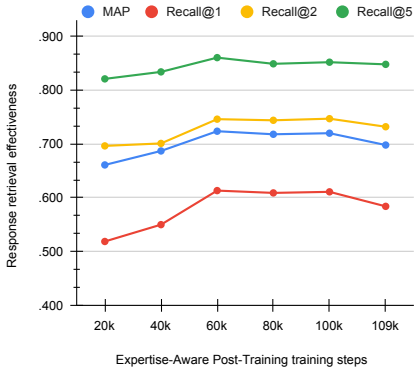

Figure 3: Impact of training steps for Expertise-Aware Post-Training on the CLosER effectiveness on the test set.

Table 3: The effectiveness of our proposed method, CLosER, on LegalConv for conversational legal search task compared to the previous state-of-the-art models. 'R' refers to recall and 'N' refers to NDCG. 'FP' refers to the fine-grained post-training approach proposed in [21], the current state-of-the-art model on the Ubuntu corpus benchmark. 'DPT' refers to the domain post-training approach proposed in [63]. Significance is shown with † for our proposed method, CLosER (row *j*), compared to the best baseline (row *i*). Statistical significance was measured with a paired t-test ($p < 0.05$).

		MAP	R@1	R@2	N@5
(a)	BM25	.213	.151	.208	.225
(b)	BERT	.557	.471	.574	.578
(c)	Legal BERT	.565	.485	.583	.585
(d)	Longformer	.580	.490	.594	.597
(e)	Legal Longformer	.604	.498	.636	.637
(f)	Legal BERT-DPT [63]	.571	.488	.604	.601
(g)	Legal Longformer-DPT [63]	.624	.501	.654	.657
(h)	Legal BERT-FP [21]	.603	.492	.610	.635
(i)	Legal Longformer-FP [21]	.651	.512	.688	.681
(j)	CLosER **(ours)**	**.724**†	**.613**†	**.746**†	**.749**†
(k)	CLosER w/o UR (ours)	.705	.569	.722	.710
(l)	CLosER w/o EA-PT (ours)	.652	.515	.687	.684

Performance comparison (RQ1). The top part of Table 3 shows the performance of the baseline models, namely, BM25, several cross-encoder models (rows *b*–*e*), the domain post-training method by Whang et al. [63] (rows *f* and *g*), and the fine-grained post-training method proposed by Han et al. [21] (row *h* and *i*). We find that Longformer-based models achieve better results compared to the BERT-based models. Row (*a*) shows that BM25 achieves the lowest performance on all metrics by a large margin, which could be partly due to the large lexical mismatch between candidate responses and the context, making the task challenging for BM25. Comparing rows *b* with *c*, and *d* with *e*, we see that regardless of BERT or Longformer, for both models, using the legal pre-trained version leads to higher effectiveness than the models that have been pre-trained on general data. Among the baselines, the two models

Table 4: Analyzing retrieval difficulty for shallow-expertise lawyer utterances. 'Average of Δ%' is the reduction in model effectiveness for retrieving the next responses from shallow vs. deep expertise lawyers. A lower 'Average of Δ%' indicates that the model achieves robust effectiveness between the next utterances that are written by lawyers with shallow and deep expertise.

Model	Lawyer Expertise Level	MAP	Recall@1	Recall@2	Recall@5	Average of Δ%
BM25	deep	.241	.175	.237	.328	5.838 ↓
	shallow	.184	.126	.179	.259	
Legal Longformer	deep	.629	.519	.662	.742	5.015 ↓
	shallow	.579	.477	.610	.686	
Legal Longformer-FP	deep	.673	.532	.714	.843	4.715 ↓
	shallow	.629	.491	.663	.791	
CLosER (ours)	deep	.727	.616	.749	.865	0.687 ↓
	shallow	.720	.610	.743	.856	
CLosER w/o EA-PT (ours)	deep	.674	.536	.710	.842	4.450 ↓
	shallow	.631	.495	.665	.794	

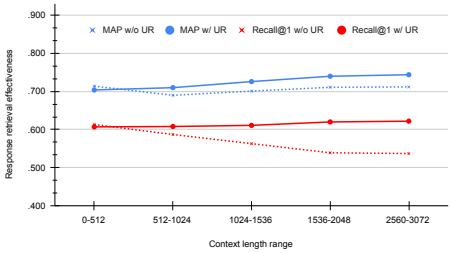

Figure 4: Effectiveness difference within CLosER without Utterance Re-Ordering (UR) and CLosER with UR for different ranges of context length on average. Utterance Re-ordering is effective when the context is longer.

with the highest effectiveness are the replicated FP models [21] with Legal BERT and Legal Longformer checkpoints (row *h* and *i*). The results show that our proposed method, CLosER, significantly outperforms all the baselines in terms of various evaluation metrics. Specifically, we see in row *j* that CLosER beats the state-of-the-art Legal Longformer-FP by a large margin, demonstrating the effectiveness of our post-training learning objective and utterance re-ordering strategy.

Effect of expertise awareness (RQ2). We hypothesize that retrieving the next response written by lawyers with shallow expertise is more difficult than by lawyers with deep expertise because a lawyer with shallow expertise probably writes less exact answers which are more difficult to distinguish from random responses.To investigate this effect on different models, we categorize the test dialogues' context into two categories based on the expertise level of the lawyer who is the writer of the ground truth next utterance. We then evaluate all models separately on the two subsets. Table 4 shows that all of the baselines suffer substantially in effectiveness when the ground-truth next utterances are written by shallow-expertise lawyers. This suggests that the shallow-expertise lawyers respond less accurately when it is out of their expertise domain, making it more difficult for the models to retrieve their next utterances. Nevertheless, the table shows that CLosER exhibits a more robust behavior where the model's performance is less affected by the expertise level (4.715 Δ% vs. 0.687 Δ%). We attribute this to the expertise-aware post-training learning objective, helping the model to better differentiate between utterances that are written by shallow-expertise lawyers and irrelevant utterances. This is more evident if we compare the performance of CLosER and CLosER w/o EA-PT in Table 4. We see that when the expertise-aware post-training learning objective is not used, the performance degrades to a greater extent (0.687Δ% vs. 4.450Δ%), reaffirming our hypothesis.

Table 5: The effectiveness of BM25 given different context selections (utterances used as query). Significance is shown with † for comparing BM25 with AllPrevUtterances context representation compared to all of the other representations. Statistical significance was measured with a paired t-test ($p < 0.05$) with Bonferroni correction for multiple testing.

Context	MAP	R@1	R@2	R@5	N@5
FirstUtterance	.166	.110	.157	.237	.175
PrevUtterance	.161	.107	.153	.229	.170
FirstPrevUtterance	.199	.137	.192	.278	.210
AllPrevUtterances	**.213**†	**.151**†	**.208**†	**.294**†	**.225**†
RandomUtterance	.166	.112	.158	.235	.175

Table 6: Ablation study on the impact of each task of Post-Training on the effectiveness of CLosER.

Post-Training	MAP	Recall@1	Recall@5
Only MLM	.678	.579	.821
Only EA-PT	.715	.601	.847
MLM + EA-PT	.724	.613	.861

Effect of training steps for expertise-aware post-training (RQ3). Figure 3 depicts the effect of the number of training steps of post-training on the effectiveness of CLosER. Our goal is to examine how sensitive CLosER is to the effectiveness of the post-training phase. We see that more post-training steps lead to improved performance in general. However, CLosER converges at 60K steps, where we see diminishing results with more training steps. Since the same effect occurred on the validation set during post-training, we stopped training after 60K steps and picked that post-trained model for fine-tuning CLosER. Next, we investigate the classification effectiveness to see if it also agrees with our end-to-end results. Interestingly, we observe the same trend with the highest classification accuracy (63%) after the same number of post-training steps.

This confirms the effect of effective post-training on end-to-end performance.

Effect of utterance re-ordering (RQ4). We analyze the impact of Utterance Re-ordering (UR) by comparing the effectiveness of CLosER w/o UR and CLosER w/ UR on different ranges of context lengths. Figure 4 shows for the shorter contexts (length range from 0 to 512 words), the CLosER w/o UR achieves higher effectiveness. However, by increasing the length of context, we observe a consistent improvement (for all context length ranges from 512 to 3072 words) by CLosER w/ UR compared to CLosER w/o UR. Therefore, our analysis confirms that utterance re-ordering aligns with the lengthy context of the Legal ConvPR task and local attention architecture of Longformer. This could maybe suggest that re-ordering the shorter context effectively is more challenging and need further study in future works.

7 DISCUSSION

To further analyze the legal ConvPR task and the behavior of the proposed CLosER model, we address three additional questions.

Do we need all the previous utterances?. One of the challenges of legal conversations is that the conversations consist of multiple utterances — much more than the conversations in other domains. Given the fact that each utterance is on average 8 times longer than other domains, it is very easy to reach the maximum-token limit of Transformer-based models. Therefore, in this experiment, we aim to study how much deep we should go back into the conversation's context in the legal domain. In our experiments, we use the complete conversation context for all models that can deal with long input texts. In additional analyses, we investigate the effect of using the full vs. only a part of the context: Inspired by Aliannejadi et al. [2], we design five heuristic approaches to select utterances from the context: (i) *FirstUtterance*, which means we only use the first utterance of context as the query; (ii) *PrevUtterance*, which means we only use the last utterance of the context (the one before the correct response) as the query; (iii) *FirstPrevUtterance*, which means we concatenate the first and previous utterances as the query; (iv) *AllPrevUtterances*, which means we use all the utterances of the context as query; and (v) *RandomUtterance*, which means we randomly select an utterance from context as the query. We then use the baseline model, BM25 (with no length limitation), to perform the ConvPR task. Table 5 lists the results. It shows that the highest effectiveness is achieved by exploiting all the utterances and that the effectiveness in terms of all metrics drops significantly if any other utterance selection method is used. This indicates that the conversations in LegalConv are complex, making it challenging for such simple heuristic models to model context relevance. Considering the successful results of our utterance re-ordering technique, we can conclude that predicting relevant utterances in the conversation context is a key task in conversational legal search, benefiting the model even more in cases where the whole conversation context exceeds the model's input limit.

Ablation study on EA-PT. We do an ablation study on the multi-task optimization of EA-PT to analyze to what extent the MLM loss and Classification Cross Entropy loss have an impact on the effectiveness of CLosER on the next response selection task. As shown in Table 6, the effectiveness of CLosER is the highest when we utilize both losses. We see that the MLM loss has less impact on the effectiveness of legal ConvPR than EA-PT. This suggests that this can partly be due to the fact that, prior to the EA-PT, we first pre-train the Longformer on the legal cases, and consequently, the MLM loss cannot have a huge additional impact when we post-train it on LegalConv.

Importance of added tokens. In order to analyze more in-depth what is the effect of the three newly added tokens, namely, *[EUD]*, *[EUQ]*, and *[EUS]*. We show their contributions to the matching score compared to other words in the context and candidate response using the Integrated Gradient (IG) [55] which has been proven to be a stable and reliable interpretation method in many different applications including IR [15, 74, 75]. To perform our analysis we randomly sample 1000 queries from LegalConv. For each query, we take the 100 candidate responses from the pool and feed all pairs of query and their corresponding candidate response ($100k$ pairs) to CLosER. Then we compute the attribution scores over the input at the word level. We rank tokens based on their importance using the absolute value of their attribution score and take the mode of the rank of the added tokens over all samples. The modes are 8, 5, 7 for *[EUQ]*, *[EUD]*, and *[EUS]*, respectively. This shows that CLosER relatively highly attributes the added tokens in order to tackle the next response selection task.

8 CONCLUSION

In this paper, we address the task of conversational search in the legal domain, more specifically conversational passage response retrieval, a form of next utterance selection for passage utterances. We propose a novel model called CLosER with two methodological contributions: (i) *Expertise-Aware Post-Training*, an auxiliary task that helps the model to distinguish utterances from users with different levels of expertise; and (ii) *Utterance Re-ordering*, a strategy to overcome the limitations of the sparse attention mechanism of the Longformer architecture. We collect and release a large dataset of conversations from a legal CQA website and evaluate CLosER against probabilistic, BERT-based, Longformer-based state-of-the-art ConvPR baselines. Our experiments show that the proposed method substantially outperforms all the baselines. Our additional analyses indicate that Expertise-Aware Post-Training indeed enables the model to capture the different expertise levels expressed in utterances and that the Utterance Re-ordering component improves the effectiveness of CLosER by bringing the potentially relevant utterances in the attention context of the target utterance.

While our proposed method is designed for response selection in the legal domain, it is generalizable to other tasks with domain-specific challenges. Our method for Expertise-Aware Post-Training can be adapted to other auxiliary classification tasks for conversational search. Our Utterance Re-ordering method can be applied to other tasks with long contexts, as a way to overcome the limitations of the sparse attention mechanism in the Longformer architecture. For future works, our data facilitate other tasks such as legal question answering and legal response generation.

ACKNOWLEDGMENTS

This work was supported by the EU Horizon 2020 ITN/ETN on DoSSIER (H2020-EU.1.3.1., ID: 860721).

REFERENCES

[1] Amin Abolghasemi, Suzan Verberne, and Leif Azzopardi. 2022. Improving BERT-based query-by-document retrieval with multi-task optimization. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II*. Springer, 3–12.

[2] Mohammad Aliannejadi, Manajit Chakraborty, Esteban Andrés Ríssola, and Fabio Crestani. 2020. Harnessing evolution of multi-turn conversations for effective answer retrieval. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 33–42.

[3] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 475–484.

[4] Sophia Althammer, Arian Askari, Suzan Verberne, and Allan Hanbury. 2021. DoSSIER@ COLIEE 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. *arXiv preprint arXiv:2108.03937* (2021).

[5] Sophia Althammer, Sebastian Hofstätter, Mete Sertkan, Suzan Verberne, and Allan Hanbury. 2022. PARM: A paragraph aggregation retrieval model for dense document-to-document retrieval. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*. Springer, 19–34.

[6] Arian Askari, Georgios Peikos, Gabriella Pasi, and Suzan Verberne. 2022. Leibi@ coliee 2022: aggregating tuned lexical models with a cluster-driven bert-based model for case law retrieval. *arXiv preprint arXiv:2205.13351* (2022).

[7] A Askari and S Verberne. 2021. Combining lexical and neural retrieval with longformer-based summarization for effective case law retrieva. In *Proceedings of the second international conference on design of experimental search & information REtrieval systems*. CEUR, 162–170.

[8] Arian Askari, Suzan Verberne, Amin Abolghasemi, Wessel Kraaij, and Gabriella Pasi. 2023. Retrieval for Extremely Long Queries and Documents with RPRS: a Highly Efficient and Effective Transformer-based Re-Ranker. *arXiv preprint arXiv:2303.01200* (2023).

[9] Arian Askari, Suzan Verberne, and Gabriella Pasi. 2022. Expert Finding in Legal Community Question Answering. In *European Conference on Information Retrieval*. Springer.

[10] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2, 3 (2021), 315–324. https://doi.org/10.3390/epidemiologia2030024

[11] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[12] Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. FIRE 2019 AILA track: Artificial intelligence for legal assistance. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*. 4–6.

[13] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*. Washington DC, USA.

[14] 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*. Association for Computational Linguistics, Online, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[15] Lijuan Chen, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2021. Toward the Understanding of Deep Text Matching Models for Information Retrieval. *arXiv preprint arXiv:2108.07081* (2021).

[16] Raymond D'Amore. 2004. Expertise community detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 498–499.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[18] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176* (2022).

[19] Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2041–2044.

[20] Rupert Haigh. 2018. *Legal English*. Routledge.

[21] Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained Post-training for Improving Retrieval-based Dialogue Systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1549–1558.

[22] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666* (2020).

[23] Adebayo Kolawole John, Luigi Di Caro, Livio Robaldo, and Guido Boella. 2017. Legalbot: A deep learning-based conversational agent in the legal domain. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 267–273.

[24] Mi-Young Kim, Juliano Rabelo, Kingsley Okeke, and Randy Goebel. 2022. Legal information retrieval and entailment based on bm25, transformer and semantic thesaurus methods. *The Review of Socionetwork Strategies* 16, 1 (2022), 157–174.

[25] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[26] Feng-Lin Li, Minghui Qiu, Haiqing Chen, Xiongwei Wang, Xing Gao, Jun Huang, Juwei Ren, Zhongzhou Zhao, Weipeng Zhao, Lei Wang, et al. 2017. Alime assist: An intelligent assistant for creating an innovative e-commerce experience. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2495–2498.

[27] Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838* (2022).

[28] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2356–2362.

[29] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: BM25 Baseline for MS MARCO Document Retrieval. https://github.com/castorini/pyserini/blob/master/docs/experiments-msmarco-doc.md

[30] Shuyi Lin, Wenxing Hong, Dingding Wang, and Tao Li. 2017. A survey on expert finding techniques. *Journal of Intelligent Information Systems* 49, 2 (2017), 255–279.

[31] Bulou Liu, Yiran Hu, Yueyue Wu, Yiqun Liu, Fan Zhang, Chenliang Li, Min Zhang, Shaoping Ma, and Weixing Shen. 2023. Investigating Conversational Agent Action in Legal Case Retrieval. In *European Conference on Information Retrieval*. Springer.

[32] Bulou Liu, Yueyue Wu, Fan Zhang, Yiqun Liu, Zhihong Wang, Chenliang Li, Min Zhang, and Shaoping Ma. 2022. Query Generation and Buffer Mechanism: Towards a better conversational agent for legal case retrieval. *Information Processing & Management* 59, 5 (2022), 103051.

[33] Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1261–1264.

[34] Daniel Locke and Guido Zuccon. 2022. Case law retrieval: problems, methods, challenges and evaluations in the last 20 years. *arXiv preprint arXiv:2202.07209* (2022).

[35] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. (Sept. 2015), 285–294. https://doi.org/10.18653/v1/W15-4640

[36] Junyu Lu, Xiancong Ren, Yazhou Ren, Ao Liu, and Zenglin Xu. 2020. Improving contextual language models for response retrieval in multi-turn conversation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1805–1808.

[37] Yixiao Ma, Qingyao Ai, Yueyue Wu, Yunqiu Shao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. Incorporating Retrieval Information into the Truncation of Ranking Lists for Better Legal Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 438–448.

[38] Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer. *arXiv preprint arXiv:2211.00974* (2022).

[39] Kelong Mao, Zhicheng Dou, and Hongjin Qian. 2022. Curriculum Contrastive Context Denoising for Few-shot Conversational Dense Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 176–186.

[40] Gordon Mohr, Michael Stack, Igor Rnitovic, Dan Avery, and Michele Kimpton. 2004. Introduction to Heritrix. In *4th International Web Archiving Workshop*. Citeseer, 109–115.

[41] Douglas W Oard, William Webber, et al. 2013. Information retrieval for e-discovery. *Foundations and Trends® in Information Retrieval* 7, 2–3 (2013), 99–237.

[42] Paul Owoicho, Jeff Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R. Trippas, and Svitlana Vakulenko. 2022. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *Proceedings of the Thirty-First Text REtrieval Conference*.

[43] Vedant Parikh, Upal Bhattacharya, Parth Mehta, Ayan Bandyopadhyay, Paheli Bhattacharya, Kripa Ghosh, Saptarshi Ghosh, Arindam Pal, Arnab Bhattacharya, and Prasenjit Majumder. 2021. AILA 2021: Shared task on artificial intelligence for legal assistance. In *Forum for Information Retrieval Evaluation*. 12–15.

[44] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[45] Alina Petrova, John Armour, and Thomas Lukasiewicz. 2020. Extracting Outcomes from Appellate Decisions in US State Courts.. In *JURIX*. 133–142.

[46] The President and Fellows of Harvard University. 2022. Caselaw Access Project. https://case.law/about/

[47] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and Discussion of the Competition on Legal Information Extraction/Entailment (COLIEE) 2021. *The Review of Socionetwork Strategies* 16, 1 (2022), 111–133.

[48] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2019. A Summary of the COLIEE 2019 Competition. In *JSAI International Symposium on Artificial Intelligence*. Springer, 34–49.

[49] Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. COLIEE 2020: Methods for Legal Document Retrieval and Entailment. https://sites.ualberta.ca/~rabelo/COLIEE2021/COLIEE_2020_summary.pdf

[50] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*. Springer, 232–241.

[51] Tony Russell-Rose, Jon Chamberlain, and Leif Azzopardi. 2018. Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management* 54, 6 (2018), 1042–1057.

[52] Rebecca L Sandefur. 2015. Elements of professional expertise: Understanding relational and substantive expertise through lawyers' impact. *American Sociological Review* 80, 5 (2015), 909–933.

[53] Yunqiu Shao, Yueyue Wu, Yiqun Liu, Jiaxin Mao, Min Zhang, and Shaoping Ma. 2021. Investigating User Behavior in Legal Case Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 962–972. https://doi.org/10.1145/3404835.3462876

[54] Vered Shwartz, Gabriel Stanovsky, and Ido Dagan. 2017. Acquiring predicate paraphrases from news tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*. 155–160.

[55] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[56] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27 (2014).

[57] Peter M Tiersma. 1999. *Legal language*. University of Chicago Press.

[58] David van Dijk, Manos Tsagkias, and Maarten de Rijke. 2015. Early detection of topical expertise in community question answering. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 995–998.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[60] Heyuan Wang, Ziyi Wu, and Junyu Chen. 2019. Multi-turn response selection in retrieval-based chatbots with iterated attentive convolution matching network. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1081–1090.

[61] Zihan Wang, Hongye Song, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Xiaozhong Liu, Hongsong Li, and Maarten de Rijke. 2021. Cross-Domain Contract Element Extraction with a Bi-Directional Feedback Clause-Element Relation Network. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1003–1012. https://doi.org/10.1145/3404835.3462873

[62] Sabine Wehnert, David Broneske, Stefan Langer, and Gunter Saake. 2018. Concept Hierarchy Extraction from Legal Literature.. In *CIKM Workshops*.

[63] Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2019. An Effective Domain Adaptive Post-Training Method for BERT in Response Selection. In *Interspeech*.

[64] Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14041–14049.

[65] Christopher Williams. 2007. *Tradition and change in legal English: Verbal constructions in prescriptive texts*. Vol. 20. Peter Lang.

[66] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[67] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open* 2 (2021), 79–84.

[68] Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14158–14166.

[69] Jie Yang, Ke Tao, Alessandro Bozzon, and Geert-Jan Houben. 2014. Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In *International conference on user modeling, adaptation, and personalization*. Springer, 266–277.

[70] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *The 41st international acm sigir conference on research & development in information retrieval*. 245–254.

[71] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval*. 829–838.

[72] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for Longer Sequences.. In *NeurIPS*.

[73] Hamed Zamani, Johanne R Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808* (2022).

[74] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Interpreting Dense Retrieval as Mixture of Topics. *arXiv preprint arXiv:2111.13957* (2021).

[75] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. An analysis of BERT in document ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1941–1944.

[76] Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1118–1127.