LePaRD: A Dataset of 4+ Million Examples of Judges Citing Precedents

Anonymous ACL submission

Abstract

We present the Legal Passage Retrieval Dataset, LePaRD. LePaRD is a massive collection of over 4 million U.S. federal judicial citations to precedent in context. The dataset aims to facilitate work on legal passage retrieval, a challenging practice-oriented legal retrieval and reasoning task. Legal passage retrieval seeks to predict relevant passages from precedential court decisions given the context of a legal argument. We extensively evaluate various retrieval approaches on LePaRD, and find that classification appears to work best. However, we note that legal passage retrieval is a difficult task, and there remains significant room for improvement. By publishing LePaRD, we provide a large-scale and high quality resource to foster further research on legal retrieval. Legal passage retrieval is a practice-oriented NLP task that promises to help expand access to justice by reducing the burden associated with legal research via computational assistance. The LePaRD dataset and code will be made freely available upon publication.

anonymized github repo

1 Introduction

A third of the global population lives in a common law jurisdiction where legal arguments are based on prior decisions, known as *precedents* (Fathally and Mariani, 2008). Judges and lawyers use citations to build on these precedents and frequently quote passages directly from prior cases. The U.S. legal system is an example of a common law system and U.S. federal courts have produced around 1.7 million judicial opinions, giving rise to tens of millions of passages containing legal rules, standards, and explanations, which could potentially be cited in new cases.

As part of their everyday work, lawyers and judges frequently cite such passages. These citations serve as the basis for future arguments, such as the example in Figure 1. As a result, identifying appropriate precedents relevant to a given argument represents a fundamental component of legal practice. This is a complicated and time consuming endeavour: The sheer volume of passages which could potentially be cited adds to the complexity of case law, which is exacerbated by subtle rules about the contexts in which a given passage is legally binding. We provide the large-scale dataset *LePaRD*, which can be used to develop computational retrieval methods that facilitate the retrieval of precedent in the U.S.. LePaRD was constructed by focusing on how U.S. federal judges actually used precedential passages and as such it builds on millions of expert decisions.

In practice, highly paid attorneys spend significant time on legal research and finding relevant precedent – and they routinely bill up to $100 per individual search (Franklin County Law Library, 2023). Meanwhile, in the U.S., around 90% of civil legal problems encountered by low-income individuals do not receive adequate legal help (Slosar, 2022) and access to such services is also limited for small businesses (Baxter, 2022). The cost of precedent retrieval and legal research contributes to this access to justice gap. Thus, the complexity of legal research may be partially responsible for the high cost of litigation and the associated access to justice gap.

Legal NLP promises to be a powerful equalizer in the legal profession (Mahari et al., 2023b), but many areas of legal practice have been slow to adopt relevant technologies that increase efficiencies and reduce costs for clients. While this may be partially driven by a lack of incentives and risk-aversion from legal community, legal NLP research also seems to be disconnected from the needs of legal practitioners (Mahari et al., 2023b). This in turn is partially driven by the lack of large-scale resources for practice-oriented legal NLP tasks.

To help address the high costs of legal research,

***Diamond v. Chakrabarty*, 447 U.S. 303, 315 (1980)**

It is, of course, correct that Congress, not the courts, must define the limits of patentability; "the province and duty of the judicial department to say what the law is." Marbury v. Madison, 1 Cranch 137, 177 (1803).

Figure 1: A simple example of how judges use quotations to precedent taken from the *Diamond v. Chakrabarty*. In LePaRD, preceding context is extracted ahead of a quotation from the destination opinion (*Diamond v. Chakrabarty*). Quotations are matched to the corresponding target passage from the source opinion (*Marbury v. Madison*) by using the citations contained in judicial opinions. The goal of legal passage retrieval is to predict the correct target passage given the preceding context.

and the resulting access to justice issues, and to foster more legal NLP research on practice-oriented tasks, we release the Legal Passage Retrieval Dataset LePaRD. LePaRD represents a large set of previously cited U.S. federal precedents, containing millions of argument contexts and the relevant target passage. In this work, we document the construction of LePaRD and describe relevant dataset statistics. We also extensively evaluate various retrieval approaches from the NLP literature (see e.g., Yang et al., 2017; Reimers and Gurevych, 2019; Mahari, 2021; Tay et al., 2022), some of which have been applied to other legal IR tasks (e.g., Ma et al., 2021a; Rosa et al., 2021). Our most accurate method achieves a recall@10 of 60% on the LePaRD testset, indicating that there is still ample room for improvement to be made on legal passage retrieval. No large-scale resources for legal passage retrieval exists and so we address this gap by constructing and releasing LePaRD.

The LePaRD dataset contains citations to relevant precedents paired with the contexts in which they have been cited by judges. We also provide relevant meta-data, such as the court and year of an opinion, which may be relevant for future work on legal retrieval. Retrieving relevant passages with computational assistance has the potential to reduce the time and cost associated with legal research and thus to reduce the overall cost of litigation. In publishing the dataset, we seek to catalyze practice-oriented legal NLP, and ultimately, we hope that models trained on LePaRD will reduce the burden associated with legal research for judges and lawyers, thus helping to expand access to justice.

2 Related Work

Retrieval of relevant legal passages or cases is a fundamental task in legal practice. Most existing search tools are closed-source and the usage of such tools can cost up to $100 per search (Franklin County Law Library, 2023). Under these conditions, it is noteworthy that relatively little academic work has focused on legal retrieval.

Legal retrieval has been explored in some prior work. Mahari (2021) introduces the legal passage retrieval task, however, no corresponding dataset was released and the paper focused on just 5,000 target passages (in contrast to 1.8 million in LePaRD). This is a general problem in legal NLP where large-scale professionally annotated data sources remain proprietary[1]. Moreover, creating such resources remains costly due to the intricacies of legal language, which complicate the creation of large-scale resources without expert annotators and the high cost of expert annotators. The lack of data has in turn made it challenging for legal NLP research to focus on tasks aligned with the needs of legal practitioners.

Other related work includes the COLIEE shared task series related to legal case retrieval (e.g., Rabelo et al., 2022; Kim et al., 2023). In this setting, a system is given a query and has to retrieve the most related case (or statute) from a pre-defined knowledge base. Compared to these information retrieval tasks using synthetic queries, our dataset construction is more closely aligned with actual legal practice. Furthermore, the COLIEE datasets remain limited in size, containing around 4,400 cases which could potentially be retrieved[2], whereas our dataset allows us to investigate legal passage retrieval methods at scale, containing the universe of all cited legal passages in U.S. federal courts. This setting more closely resembles how a practicing attorney would perform legal research. Finally, lexical overlap seems to play a significant role in COLIEE datasets (Rosa et al., 2021), making BM25 a strong baseline in that setting. In contrast, we find

[1]For example Lexis, Westlaw, and Bloomberg.

[2]We acknowledge and greatly appreciate the continued effort in constructing and expanding the COLIEE datasets. They are increasing in size each year, however, we believe there is room for other, complementary larger-scale legal retrieval datasets.

that this does not hold in LePaRD.

A growing body of work investigates legal citation prediction (Dadgostari et al., 2021; Huang et al., 2021) or the retrieval of relevant cases given a query (Sansone and Sperlí, 2022; Ma et al., 2021b). Based on the preceding context from a legal document, the goal in legal citation prediction is to identify the citation that supports the context in question. By contrast, in legal passage retrieval, the aim is to identify a specific *passage* of precedent rather than a citation to a whole case (which is usually tens or even hundreds of pages long). We believe there are several reasons to focus on legal passage retrieval over legal citation prediction. Legal citation prediction accuracy numbers seem very strong (see e.g., Huang et al., 2021). We attribute these results to the long-tailed distribution of citations and believe that models take shortcuts to determine a topic for a snippet and then return the most cited cases for these topics – whereas legal passage retrieval inherently requires more involved legal reasoning. This also connects to *relevance* in legal search, i.e., finding the appropriate target (Van Opijnen and Santos, 2017). We believe legal *relevance* is more strongly captured by searching for short passages, instead of predicting citations to entire cases since a case is likely to deal with multiple independent arguments.

Some passages may not be semantically linked to the concepts they stand for, making it difficult to identify them using lexical overlap or semantic search.[3] Instead, the link is established via frequent citations. By contrast, sometimes there exists an entailment relation (see e.g. Dagan et al., 2005; Bowman et al., 2015) between the context and the cited source passage, where the two passages are connected via legal reasoning. However, we find that this entailment in legal reasoning manifests differently in practical legal settings than in other NLP contexts. Thus, models trained on e.g, natural language inference (Bowman et al., 2015; Reimers and Gurevych, 2019) fail to recognize such relations in LePaRD. Hence, our specially curated dataset may better facilitate the approximation of legal reasoning by NLP models. Finally, from the perspective of practitioners, we believe that it is more useful to predict specific passages than citations to cases that may be hundreds of pages long.

[3]For example the phrase "play in the joints" is commonly used by courts to refer to a category of state actions that are permitted by the Establishment Clause but not required by the Free Exercises Clause of the First Amendment to the U.S. Constitution, see *Locke v. Davey*, 540 U.S. 712 (2004).

3 Legal Passage Retrieval Dataset

U.S. federal courts are bound by the doctrine of *Stare Decisis*, which means that they must abide by past decisions. As a result, judges and lawyers build their arguments on citations to precedents. Often these citations will be accompanied by quotations. When performing legal research, frequently cited passages of precedent are often displayed prominently by research platforms and serve as quasi-summaries of judicial opinions. In this work, we leverage the quotations contained in judicial opinions to assemble a large dataset of precedential passages.

3.1 Case Law Access Project

Harvard's Case Law Access Project (CAP) has scanned almost seven million published judicial opinions from U.S. federal and state courts.[4] CAP provides access to raw opinion texts along with opinion metadata (which includes the relevant court, citations contained in the opinion, and the decision date). Here we focus on judicial opinions published in U.S. federal courts including the U.S. Supreme Court, 13 federal appellate courts, and 94 district courts. Our study focuses on the 1.7 million published federal judicial opinions contained in CAP.

3.2 Dataset Construction

The LePaRD dataset is assembled by identifying quoted passages in judicial opinions, matching these passages to source opinions, and extracting the context within which the passages occur. This procedure is summarized in Figure 2. In general, our construction process aims to construct a large dataset that covers as many legal contexts as possible while minimizing the amount of noise introduced by e.g., OCR errors. Given the large volume of data available, we made some design decisions that removed training examples (for example, very long passages), because including these special cases led to other issues, e.g., noisier data. We will re-visit such issues in future work.

Preprocessing. For each CAP case, we retain the case id, opinion text, citations, court, and decision date. To facilitate downstream tasks, each opinion text was split into sentences using a Roberta

[4]https://case.law

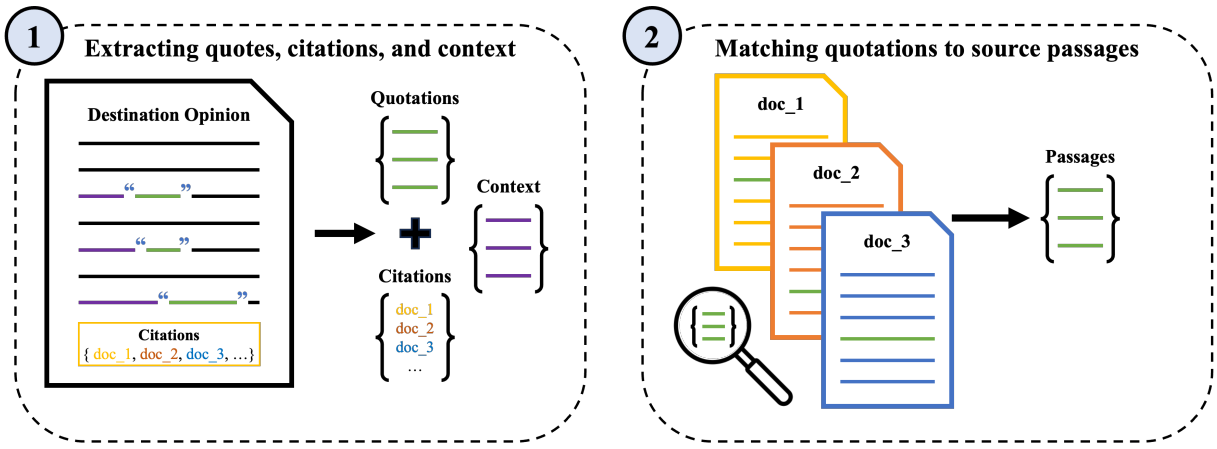

Figure 2: Flowchart of how LePaRD was constructed. Quotations, citations, and context were extracted from all 1.7 million opinions and then matched against the text of the opinions cited (source opinions). If a match was found, then a training example was constructed using the relevant preceding context and the passage.

model (Liu et al., 2019) trained to predict sentence boundaries in legal text, using the legal sentence tokenization dataset introduced by Sanchez (2019). The model was trained using the transformer library (Wolf et al., 2020) with the standard hyperparameters found in the *Trainer* library. No further text preprocessing is performed.

For all case citations, we drop duplicated citations as well as erroneous self-citations. We convert citations to case ids by mapping each possible case citation to the relevant id. For example, *Marbury v. Madison* may be cited as "1 Cranch 137", "5 U.S. 137", "2 L. Ed. 60", "SCDB 1803-005", or "1803 U.S. LEXIS 352". We map all of these to case_id = 12121622.

Extracting quotations and context from destination opinions. In each sentence of each opinion, we search for text in quotation marks (either straight or left/right quotation marks) using a regular expression. We retain quotations longer than five words and less than 150 words (long quotations often are the result of a missing quotation mark due to OCR errors).

We extract one or more sentences of "preceding context" before the quotation up to a maximum of 300 words or until we reach the end of the last quotation to avoid "overlapping contexts" where we would have to predict multiple precedential passages from the same context. For multi-sentence contexts, we impose a word limit as sentences vary drastically in length.

Matching quotations to source passages. We refer to the opinions from which quotations have been extracted as "destination opinions" and we seek to match these quotations to the relevant "source opinion". Based on the previous steps, we have a list of quotations and citations for each destination opinion. Using these citations, we check whether each quote appears in each of the cited opinions (using fuzzy string matching to account for OCR errors and modifications judges might make to the quotation to match verb tenses and capitalization). Specifically, we match the quoted text against each sentence in the source opinion. This means that source passages will always be a single sentence long, potentially excluding very long quotations. In practice, we find that courts usually quote fairly short portions of longer passages (see Table 1). To avoid many versions of the same passage we retain the entire passage sentence as the target (see Appendix A for some examples.). If a quoted passage is found to exist in a cited opinion, then this opinion is treated as the "source" of the passage. Each passage thus has one source but it may have many destinations (two on average, see Table 1). While most of the unsuccessful matches are quotations that do not come from other opinions, our approach does not tend to match multi-sentence quotations or ellipsized quotations. We leave the treatment of these outliers for future work.

The LePaRD dataset contains the preceding context, target passage, destination court, source court, destination decision date, and source decision date for each quotation that could be matched to a passage. Ultimately, we extract and validate 1.8 million unique target passages that have appeared in approximately 4.3 million contexts.

4 Dataset Statistics

Quotations serve several purposes in legal writing: they may be used for emphasis, refer to case documents and exhibits, introduce information from witness or expert testimony, cite supporting materials like treatises or academic publications, or they may reference precedential court opinions. Across all federal judicial opinions, we identify 15,800,000 quotations over five words long and we successfully match 4,300,000 of these to passages of precedent. While quotations and citations in judicial opinions offers several interesting avenues for legal NLP and legal passage retrieval, we focus on quotations that can be mapped to single sentences from another opinion. Future work could also examine the retrieval of longer passages or move beyond quotations to general citations (many of which are associated with a "pincite" or page number). In this section, we present several summary statistics about LePaRD (see Table 1) and we highlight some key observations.

First, we note that citations in judicial opinions obey a long-tailed distribution, with the top-1% accounting for 18% of all citations and 64% of all passages receiving just 1 citation. This results in an inherent imbalance in the dataset, raising unique challenges for legal precedent retrieval.

Second, the sentence lengths vary substantially and this results in passages and contexts of varying lengths (the longest passage is over 18,000 characters long). This means that many passages and contexts will be truncated by standard text retrieval approaches.

Third, most destination opinions contain several passages (around 7 on average, but occasionally tens or hundreds). This suggests that there are multiple contexts that occur within a single opinion—something that will be familiar to legal practitioners. In our view, this validates the approach of using local context before a quotation rather than searching for more remote context that may be less relevant (for example, many opinions will discuss factors related to jurisdiction or venue early on but these will not come up anywhere else in the opinion).

Fourth, the average source opinion is represented 11 times in our data. While we treat passages from the same source as separate, it appears likely that they would be conceptually linked (since the portions of an opinion that are cited tend to be somewhat novel or unique and it is uncommon, though not impossible, for there to be multiple such passages in the same opinion). Future work could thus attempt to group passages by source.

Finally, we find that there is a tremendous amount of variance in the training data by source court. We include courts to allow future users of LePaRD to narrow predictions by court in order to consider the role of binding precedent. However, it appears that for most courts, there is insufficient data to train independent models.

5 Expert Evaluation

A legal expert (Massachusetts licensed attorney) was asked to review 100 randomly sampled training examples. For each example, the expert determined whether (1) the example was generally clean and free of errors and (2) the preceding context provided sufficient information to determine that the target passage is relevant to the context. Based on this evaluation, all examples were clean and free of errors other than preexisting errors stemming from the OCR—we leave addressing these as an opportunity for future work. In 99% of these examples, the expert determined that there was enough information in the context to determine the relevance of the target passage. In the problematic case, the destination context spans two footnotes, the former a series of citations to unrelated memoranda, and the latter an explanatory footnote containing a quotation. Due to the CAP processing, these unrelated consecutive footnotes appear as adjacent sentences. Further investigation showed that this type of explanatory footnote with a quotation is very uncommon in the data.

6 Experiments

Problem Definition. Legal passage retrieval seeks to identify passages of precedent given a legal context. In total, we have around 1.3 million unique candidate passages that have been quoted at least once. Hence, given a legal context x_i, the task is to retrieve the relevant cited passage y_i from the set of all possible passages $\{y_1, y_2, ..., y_n\}$.

Experimental Setup. We release three sets of passages mapped to precedent in LePaRD. In these different sets, we vary the numbers of potential target passages from 10K to 50K (containing the most cited n passages). In future work, we plan to also release a dataset with *all* cited passages which will enable research on zero- and one-shot retrieval.

Feature	Mean	Std	Min	25%	50%	75%	Max
Length of passage text (chars)	306	225	24	179	253	362	18,342
Length of preceding context (chars)	562	216	5	430	516	638	14,062
Training examples per passage	2.22	8.45	1	1	1	2	4,136
Training examples per destination	6.73	12.2	1	1	3	7	1,375
Training examples per source	11.3	51.8	1	2	4	10	14,450
Training examples per source court	9,831	57,236	1	14	295	1,335	865,594

Table 1: Summary statistics of dataset features

Number of cited passages	Train	Dev	Test
10'000	520K	26K	26K
20'000	701K	35K	35K
50'000	928K	52K	52K

Table 2: Number of examples in different splits of LePaRD.

This is an important extension as the majority of passages have only been cited once. Note that the labels from the 10K to the 20K version increase by a factor of 2, but the number of training examples only by a factor of 1.3. This is because citation frequency of passages obey a long-tailed distribution where a few passages are cited with disproportionate frequency, while most are rarely cited.

We split the dataset into training, development and test sets, with 90% of the data being in the training set, 5% in the development set, and 5% in the test set. We show dataset statistics in Table 2.

We compare a variety of well-established retrieval algorithms from the NLP literature on LePaRD. These results are intended to serve as a baseline for follow-up work to build upon. Our experiments also indicate that there remains substantial room for improvement and our experiments allow us to highlight some of the key challenges related to legal passage retrieval.

We specifically experiment with (1) a sparse lexical retrieval approach via BM25 using the Anserini package (Yang et al., 2017), (2) a dense embedding-based retrieval approach using generic SBERT embeddings[5] (Reimers and Gurevych, 2019), followed by maximum dot-product similarity retrieval via the FAISS package (Johnson et al., 2019), (3) a fine-tuned SBERT variant where we fine-tune SBERT[6] on our training set using the Multiple Negatives Ranking Loss (Henderson et al., 2017), and (4) passage retrieval as a text classification task where each target passage is mapped to a unique label which is the prediction target for its preceding context (Mahari, 2021; Tay et al., 2022). We provide results in this setting for a DistilBERT model (Sanh et al., 2020), and LEGAL-BERT (Chalkidis et al., 2020), a domain-adapted BERT model trained on vast amounts of legal documents. The classification models have been trained using the huggingface transformer library (Wolf et al., 2020) with the standard hyper-parameters found in the *Trainer* class.

The aim of our experiments is to include results for established information retrieval methods. These methods have been used extensively in all NLP domains. Our experiments are all implemented using their respective libraries and standard hyperparameters described above.

6.1 Results

We observe that there is only limited lexical overlap between the context and the cited passage, reflected in rather poor performance of the BM25 retrieval. This is in strong contrast to e.g., the COLIEE shared tasks where BM25 remains one of the most competitive retrieval methods (Rosa et al., 2021). We then take a pre-trained SBERT variant, which seems to transfers poorly to the legal passage retrieval task. We attribute this finding to the domain shift (the model was not trained on legal data), and the particular challenges of legal language and entailment present in the legal passage retrieval task.

We find, however, that results improve noticeably as soon as we start to fine-tune models on the LePaRD training set. We see at least double the recall for dense SBERT-based retrieval after domain-specific fine-tuning.

Recall results improve even further if we turn legal passage retrieval into a supervised classifica-

[5]We use the all-mpnet-base-v2 model which at the time of experimenting was the best overall SBERT model across 14 benchmarks.

[6]using the code from the SBERT github repository with the already set hyper-parameters: https://github.com/UKPLab/sentence-transformers/.

Approach	Number of labels	Development Set			Test Set		
		rc@1	rc@5	rc@10	rc@1	rc@5	rc@10
BM25	10K	4.94	13.94	19.66	5.01	14.0	19.52
	20K	4.18	11.61	16.48	4.13	11.51	16.67
	50K	3.41	9.22	13.25	3.41	9.43	13.53
SBERT	10K	3.78	11.94	17.89	3.8	11.9	17.65
	20K	3.23	9.61	14.13	3.08	9.57	14.2
	50K	2.33	7.09	10.73	2.3	7.1	10.67
fine-tuned SBERT	10K	8.65	29.74	43.95	8.99	30.07	44.79
	20K	7.84	25.0	38.39	7.53	25.23	38.5
	50K	5.12	16.82	26.07	5.0	16.54	25.9
Classification LEGAL-BERT	10K	16.38	37.15	47.54	16.66	37.13	47.65
	20K	12.52	28.51	36.61	12.8	28.81	36.84
	50K	8.42	19.3	24.99	8.77	19.61	25.22
Classification DistilBERT	10K	19.39	45.9	58.24	19.67	46.17	59.12
	20K	16.32	39.91	51.75	16.74	40.51	52.16
	50K	12.11	29.94	38.78	12.32	30.01	39.34

Table 3: Main results: Recall at *1, 5, 10* results for development and test set using various baselines. Best results were obtained using classification and (relatively) few labels.

tion task: Instead of trying to have the embeddings of a source and target passage close in some representation space, we instead assign a unique class label to each source passage, and the goal becomes to predict that label from the legal context (see e.g., Mahari, 2021; Tay et al., 2022). We experiment with two different models, and observe that the DistilBERT model achieves the best overall recall in all settings. Our best performance in the 10K label setting means in practice that the required target passage would be found among the top 10 search results of this model in 6 out of 10 cases.

Surprisingly, a domain-specific LEGAL-BERT model achieves worse performance than the more generic DistilBERT model. We speculate that LEGAL-BERT has been pre-trained on vast amounts of legal text from various judicial systems – and some of this pre-training data does not seem to be beneficial to retrieving relevant U.S. precedent.

Although a supervised classification approach seems to work best in our experiments, this approach comes with major limitations. Firstly, updating models in case of new precedents being introduced requires either updating existing models or re-training them from scratch (Tay et al., 2022). Secondly, LLMs have shown to exhibit biases (Abid et al., 2021; Lucy and Bamman, 2021) and resulting classification of passages in our application might potentially perpetuate these biases. Lastly, zero- and few-shot retrieval for the long tail of the distribution (e.g., all precedents in LePaRD) will not be solvable in this setting, and requires other approaches.

We show additional metrics (mean average precision and normalized discounted cumulative gain) of all experiments in Appendix Table 5. These metrics lead to the similar conclusions as the recall at various positions discussed above.

Our experiments showcase how LePaRD is a large-scale yet challenging legal retrieval dataset. We believe there is ample room for improvement, for example by considering re-ranking approaches or late interactions (Khattab and Zaharia, 2020). Nevertheless, our experiments help us make sense of the dataset, by e.g., highlighting how there is only limited lexical overlap between context and the target passage. All experiments showcase consistent behavior across dataset splits and metrics—and might be useful as baselines in future work.

7 Discussion

We highlight that the legal passage retrieval task is non-trivial, complicated by the long-tailed distribution of cited precedent and the sheer size of the corpus. In publishing LePaRD, we aim to encourage NLP work on a set of problems that are closely aligned with the needs of the legal profession. More broadly, our aim is to offer an example of how NLP can be used to broaden access to justice and to catalyze similar work in other legal

domains.

One of the challenges of legal research is that not all case law content carries the same weight. On the one hand, the structure of court systems means that precedent that is binding in one court may not be binding in another court, even if they are part of the same system (e.g., precedent from the U.S. District Court for the District of Massachusetts is not binding in the U.S District Court for the District of Oregon because these district courts are part of different judicial circuits within the U.S. federal judiciary). Similarly, old precedents may be overturned and thus lawyers must be careful to cite "good law" (although we find that passages tend to be cited for an average of ten years, see Appendix B). On the other hand, not everything that is said in a judicial opinion has the status of *precedent*: only the elements of a court's reasoning that are essential to the decision bind future courts while other content contained in a judicial opinion is known as *obiter dictum* and is not legally binding. As a result, methods that focus on lexical overlap or semantic search create a large risk of retrieving content that is not binding precedent. The LePaRD dataset addresses these issues in two ways. First, we include the court and date associated with each precedent to facilitate the identification of precedents that are valid in a certain court and time. Second, only passages that have been previously cited by judges are included in the dataset, which significantly reduces the probability of retrieving non-binding dicta. While we note that requiring a passage to be cited at least once restricts our dataset, we believe this limitation is far outweighed by the value of knowing that the passage has been selected for citation by a federal judge.

One particularly promising application of precedent prediction is its potential to serve as the basis for retrieval augmented generation using large language models (RAG). RAG has been put forward as a method of allowing models to generate text based on information that is not contained in the training data (Lewis et al., 2020; Karpukhin et al., 2020; Izacard et al., 2022). In the context of legal research and writing, RAG appears to have several key advantages. First, RAG is likely to increase the correctness of citations by allowing practitioners to ensure that only real precedents are cited (i.e., reducing the risk of hallucinations), the cited precedent is relevant to the particular court, and the cited precedent remains good law (it has not been overturned). The importance of this capability was highlighted by the recent *Mata v. Avianca Airlines* case where an attorney relied on ChatGPT to write a brief that turned out to rely on non-existent references (Weiser, 2023) Second, RAG is more easily updatable than fine-tuned models and thus allows case law to be quickly updated as new cases come out and old cases are overturned (Mahari et al., 2023a). Third, RAG is auditable in the sense that practitioners see the factual basis for generated outputs, allowing them the remove any irrelevant precedent before text is generated. While rules of professional responsibility around lawyers' use of generative AI continue to evolve, some proposals highlight an attorney's "duty to supervise" the technologies they use (Greenwood et al., 2023) and the ability to evaluate what precedent will be used as a basis for a brief appears to be a likely prerequisite for "supervising" brief writing models.

8 Conclusion

We introduce LePaRD, a large-scale dataset for predicting a target precedential passage given a legal argument context. We believe legal passage retrieval to be an important task for legal practitioners, and an interesting NLP retrieval task. From a legal perspective, searching for relevant case law consumes significant resources and contributes to the cost of litigation and the associated access to justice gap. From an NLP perspective, legal passage retrieval is a retrieval task with little lexical overlap between queries and targets, which makes it a particularly challenging retrieval problem.

We present various experiments using widely used retrieval algorithms but acknowledge that there is still significant room for improvement. There are several approaches toward better legal precedent retrieval, some of which we outline here, and the experiments we present are intended as baselines rather than optimal solutions. One example approach is to combine citation and passage retrieval to first find relevant cases and then identify specific passages within them—which can be thought of as a retrieve and re-rank approach. Alternatively, one could also retrieve the top-N passages, and re-rank those with a more powerful re-ranker. LePaRD provides a large-scale resource for such experiments and other retrieval research in the legal domain.

9 Limitations

We discussed a few limitations of this work throughout the paper. In this section, we expand on some of these points, add other limitations and outline avenues for future work.

Noise in the CAP data. Opinions are usually published in a PDF format. CAP converted these PDFs into text, which at times results in errors and the resulting text can contain errors typical in such conversion efforts at scale. While the data is clean enough to provide a valuable NLP dataset for retrieving relevant legal passages, and works well to explore legal information retrieval methods, it would need to be corrected for submission in a legal document.

Fuzzy Matching. LePaRD is created by heuristically leveraging quotations and the case law citation to retrieve the source passage form the source opinion. Due to our heuristics, OCR errors and fuzzy matching, not all examples in the dataset are actual examples of citing a source passage. In particular, if Opinion A quotes Opinion B which quotes Opinion C, then it is possible that a passage will be matched to both B and C although it originates from C. However, after expect evaluation and several experiments, we believe that LePaRD is a high quality dataset that can form the basis of impactful NLP research.

Focus on the U.S. legal system. LePaRD contains only U.S. precedent. In future work, we plan to explore whether we can create similar datasets for other jurisdictions or even for civil law contexts where citations to regulations, laws, and statues predominate.

Experiments. In the experiment section, we show experiments for, by today's standards, small transformer models such as DistilBERT and SentenceBERT. We believe that using larger and more recent models such as LLama 2 (Touvron et al., 2023) will result in better performance. However, the experiments we show are intended to be generally accessible, including for researchers with limited compute budgets. In particular, we highlight well-established retrieval algorithms, like BM25 and dense retrieval, and believe that these provide valuable baseline experiments and insights. We think of this contribution as a resource paper where we provide appropriate baseline results. Thus, we leave exploration of bigger and more recent models to future work.

10 Ethical Considerations

Intended Use. This work presents a legal information retrieval dataset—it is not intended to be a resource for anyone engaged in a legal dispute. LePaRD is intended to further practice-oriented legal NLP and it could also form the basis for real-world systems that help litigants and their attorneys with legal research. We hope that these types of technologies can help to alleviate the access to justice crisis.

Misuse Potential. We recognize that the legal context is especially sensitive, and caution researchers to think carefully about how they use LePaRD and other legal datasets. In particular, efficient legal research could help under-resourced litigants, but it can also facilitate frivolous filings.

Model Bias. Although the reported performance of NLP models is often very high, it is widely known that ML models suffer from picking up spurious correlations from data. Furthermore, it has been shown that pre-trained language models such as DistilBERT and LegalBERT suffer from inherent biases present in the pre-training data (Abid et al., 2021; Lucy and Bamman, 2021). This in turn leads to biased models—and it is thus likely that the models we present also suffer from such biases. This is especially troubling if legal passage retrieval methods work particularly poorly for certain areas of law or certain categories of litigants; we highlight the exploration of these biases and their mitigation as an important area for future work.

Data Privacy. The data used in this study is exclusively public textual data provided by CAP. It contains legal opinions from the U.S. which are public records. There is no user-related data or private data involved, which would not have been public prior to our work.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Ralph Baxter. 2022. Dereliction of duty: State-bar inaction in response to America's access-to-justice crisis. *Yale Law Journal Forum*, 132:228.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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.

Faraz Dadgostari, Mauricio Guim, Peter A. Beling, Michael A. Livermore, and Daniel N. Rockmore. 2021. Modeling law search as prediction. *Artificial Intelligence and Law*, 29(1):3–34.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Jabeur Fathally and Nicola Mariani. 2008. *Percentage of the World Population, Civil Law and Common Law Systems*. Wilson & Lafleur.

Franklin County Law Library. 2023. Lexis & westlaw pricing - cost-effective electronic legal research.

Dazza Greenwood, Shawnna Hoffman, Olga V. Mack, Jeff Saviano, Megan Ma, and Aileen Schultz. 2023. Task force on responsible use of generative AI for law. *MIT Computational Law Report*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E. Ho, Mark S. Krass, and Matthias Grabmair. 2021. Context-aware legal citation recommendation using deep learning. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 79–88, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *ArXiv*, abs/2208.03299.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert.

Mi-Young Kim, Juliano Rabelo, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2023. Coliee 2022 summary: Methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence: JSAI-IsAI 2022 Workshop, JURISIN 2022, and JSAI 2022 International Session, Kyoto, Japan, June 12–17, 2022, Revised Selected Papers*, page 51–67, Berlin, Heidelberg. Springer-Verlag.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55.

Yixiao Ma, Yunqiu Shao, Bulou Liu, Yiqun Liu, M. Zhang, Shaoping Ma, and yiqunliu. 2021a. Retrieving legal cases from a large-scale candidate corpus.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021b. Lecard: A legal case retrieval dataset for chinese law system. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2342–2348, New York, NY, USA. Association for Computing Machinery.

Robert Mahari, Tobin South, and Alex Pentland. 2023a. Transparency by design for large language models. *Computational Legal Futures, Network Law Review*.

Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex Pentland. 2023b. The law and NLP: Bridging disciplinary disconnects. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3445–3454, Singapore. Association for Computational Linguistics.

Robert Zev Mahari. 2021. AutoLAW: Augmented legal reasoning through legal precedent prediction.

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):111–133.

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*. Association for Computational Linguistics.

Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, BM25 is a strong baseline for legal case retrieval.

George Sanchez. 2019. Sentence boundary detection in legal text. In *Proceedings of the natural legal language processing workshop 2019*, pages 31–38.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal information retrieval systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Mary C. Slosar. 2022. The justice gap: The unmet civil legal needs of low-income Americans. Technical report, Legal Services Corporation.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *SIGIR*. ACM.

Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, 25:65–87.

Benjamin Weiser. 2023. Here's what happens when your lawyer uses ChatGPT. *The New York Times*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1253–1256, New York, NY, USA. Association for Computing Machinery.

A Data Sample

Table 4 shows a sample of five training examples from the LePaRD dataset. Note that only a small portion of each target passage is actually quoted in the destination opinions.

B Further Dataset Statistics

Here we provide some additional insights derived from LePaRD. In contrast to the details provided in Section 4, we will explore interdisciplinary insights that may catalyze future research.

We find that passages are cited for a long time after initial publication with a mean of 10 years and a maximum of over 150 years between the first and last citation (see Figure 4). This is relevant insofar as it highlights that a legal passage dataset will be a valuable contribution with a lasting impact for legal precedent retrieval. We further observe that a majority of quotations are to passages produced by another court, especially by the U.S. Supreme Court or by appellate courts (see Figure 3). In particular, district courts appear to cite very little of their own passages, which is unsurprising given that they are bound by the relevant higher courts and thus are more likely to cite precedent from such higher courts. These observations provide some evidence that LePaRD represents a fairly representative sample of precedential passage usage.

Clustering passage co-occurrence based on whether passages appear in the same destination context reveals interesting patterns (see Figure 5). We observe three clusters: First, a very small cluster (just two cases, *Anderson v. Liberty Lobby, Inc.* and *Celotex Corp. v. Catrett*) which pertain to summary judgement, when a judgement is entered without a full-trial which happens very frequently in many different civil disputes. Second, a small cluster of bankruptcy court cases, which are brought in a subset of specialized federal courts. Third, a large cluster containing all other passages. This clustering highlights an alternative approach to legal passage retrieval that uses a pre-existing set of

citations to predict missing ones, as explored by (Huang et al., 2021).

C Additional Results

In Table 5 we report additional evaluation metrics for our experiments. Overall, these results are quite similar to those we report in Table 3.

Meta-Data	Preceding Context	Target Passage
Destination Court: E.D.N.Y **Destination Date:** 2001-03-28 **Source Court:** Supreme Court **Source Date:** 1974-12-23	In order to satisfy this requirement, a plaintiff must establish a "sufficiently close nexus between the State and the challenged action. See American Mfrs. Mut. Ins. Co. v. Sullivan, 526 U.S. 40, 50, 119 S.Ct. 977, 985, 143 L.Ed.2d 130 (1999). Alternatively, if the government has	There where a private lessee, who practiced racial discrimination, leased space for a restaurant from a state parking authority in a publicly owned building, the Court held that the State had **so far insinuated itself into a position of interdependence with the restaurant that it was a joint participant in the enterprise.**
Destination Court: D.D.C. **Destination Date:** 2012-02-13 **Source Court:** Supreme Court **Source Date:** 2005-04-19	He filed no opposition. That Order was also mailed to Plaintiff on Sept. 14. The Court again informed Plaintiff that he must respond on or before Sept. 30 or face dismissal. Although the notice pleading rules are	We concede that ordinary pleading rules are **not meant to impose a great burden upon a plaintiff.**
Destination Court: 5th Circuit **Destination Date:** 1971-10-21 **Source Court:** Supreme Court **Source Date:** 1966-06-20	That petitioners seek to commence an immediate appeal of that portion of the courts order entered on May 28, 1971. The motives of the officers bringing the charges may be corrupt, but that does not show that the state trial court will find the defendant guilty if he is innocent, or that in any other manner the defendant will be	Against any person who is **denied or cannot enforce in the courts** of such State a right under any law providing for the equal civil rights of citizens of the United States, or of all persons within the jurisdiction thereof;"(2) For any act under color of authority derived from any law providing for equal rights, or for refusing to do any act on the ground that it would be inconsistent with such law.
Destination Court: 9th Circuit **Destination Date:** 1980-03-28 **Source Court:** Supreme Court **Source Date:** 1911-02-20	In this case there is even a stronger possibility of recurrence since the police have not offered to discontinue the practice. Id. at 43, 65 S.Ct. at 14-15. (Citations omitted). Some might read De Funis v. Odegaard, 416 U.S. 312, 94 S.Ct. 1704, 40 L.Ed.2d 164 (1974), the equal protection challenge to the University of Washington's "quota" system in admissions as authority for the proposition that the W. T. Grant or the	The questions involved in the orders of the Interstate Commerce Commission are usually continuing (as are manifestly those in the case at bar) and their consideration ought not to be, as they might be, defeated, by short term orders, **capable of repetition, yet evading review**, and at one time the Government and at another time the carriers have their rights determined by the Commission without a chance of rédress.
Destination Court: 11th Circuit **Destination Date:** 2000-03-08 **Source Court:** 10th Circuit **Source Date:** 1994-11-22	Section 1512, however, applies to attempts to prevent or influence testimony not only in federal courts but also before Congress, federal agencies, and insurance regulators. Moreover, § 1512(b) subsumes but is significantly broader than the provision of § 1985(2) making it illegal to	Section 1985(2) creates a cause of action against those who **"conspire to deter, by force, intimidation, or threat**, any party or witness" from attending or testifying in a federal court.

Table 4: Sample from the LePaRD dataset. For readability, only the last few sentence of preceding context are displayed. The portion of the target passage that appears in quotations in the destination opinion is in bold.

Approach	Number of labels	Development Set		Test Set	
		NDCG@10	MAP	NDCG@10	MAP
BM25	10K	11.38	8.85	11.37	8.88
	20K	9.53	7.40	9.56	7.39
	50K	7.66	5.96	7.79	6.04
SBERT	10K	9.79	7.34	9.75	7.35
	20K	7.92	6.03	7.85	5.92
	50K	5.91	4.45	5.88	4.42
fine-tuned SBERT	10K	26.31	19.88	26.27	19.84
	20K	20.8	15.47	20.72	15.32
	50K	13.98	10.31	13.78	10.12
Classification LEGAL-BERT	10K	30.66	25.42	30.75	25.52
	20K	23.49	19.43	23.77	19.71
	50K	15.92	13.12	16.25	13.48
Classification DistilBERT	10K	37.26	30.73	37.73	31.09
	20K	32.45	26.46	32.91	26.94
	50K	24.21	19.69	24.57	20.0

Table 5: Additional results: NDCG@10 and Mean Average Precision for development and test set using various baselines. Best results were obtained using classification and (relatively) few labels. Metrics calculated using the (Van Gysel and de Rijke, 2018) package.

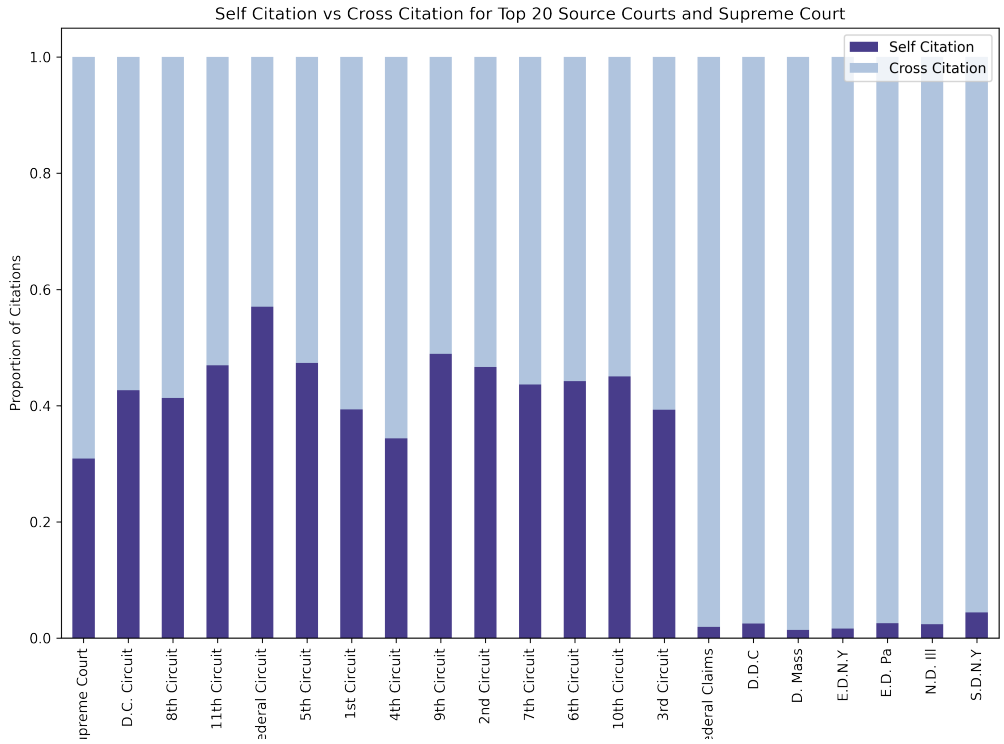

Figure 3: Comparing citations to judicial opinions from the same court ("self citation") to citations to other courts ("cross cite"). We find that appellate courts are most likely to cite themselves, while district courts only rarely cite their own precedent.

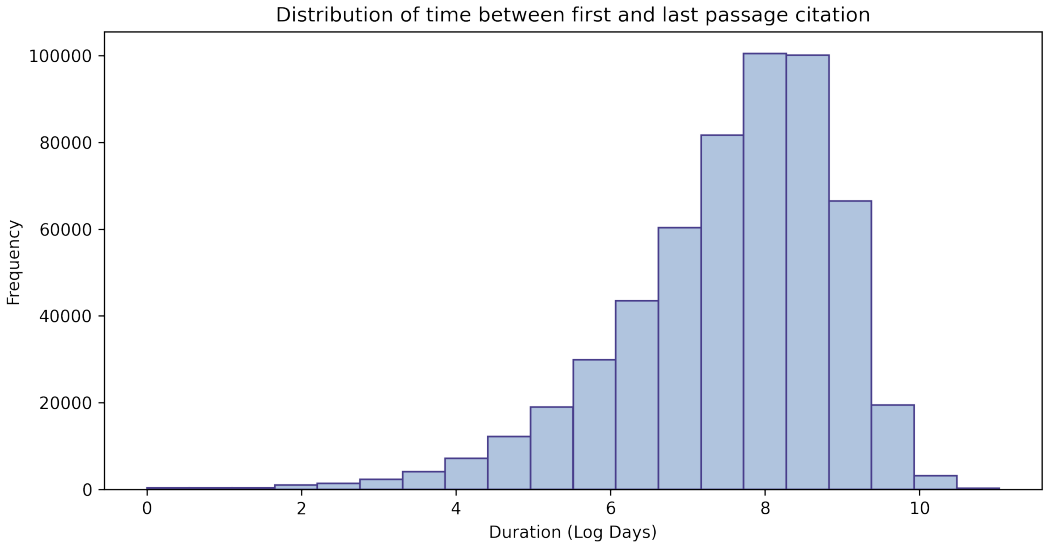

Figure 4: Distribution of time in units of log days between the first and last citation of a passage in our data.

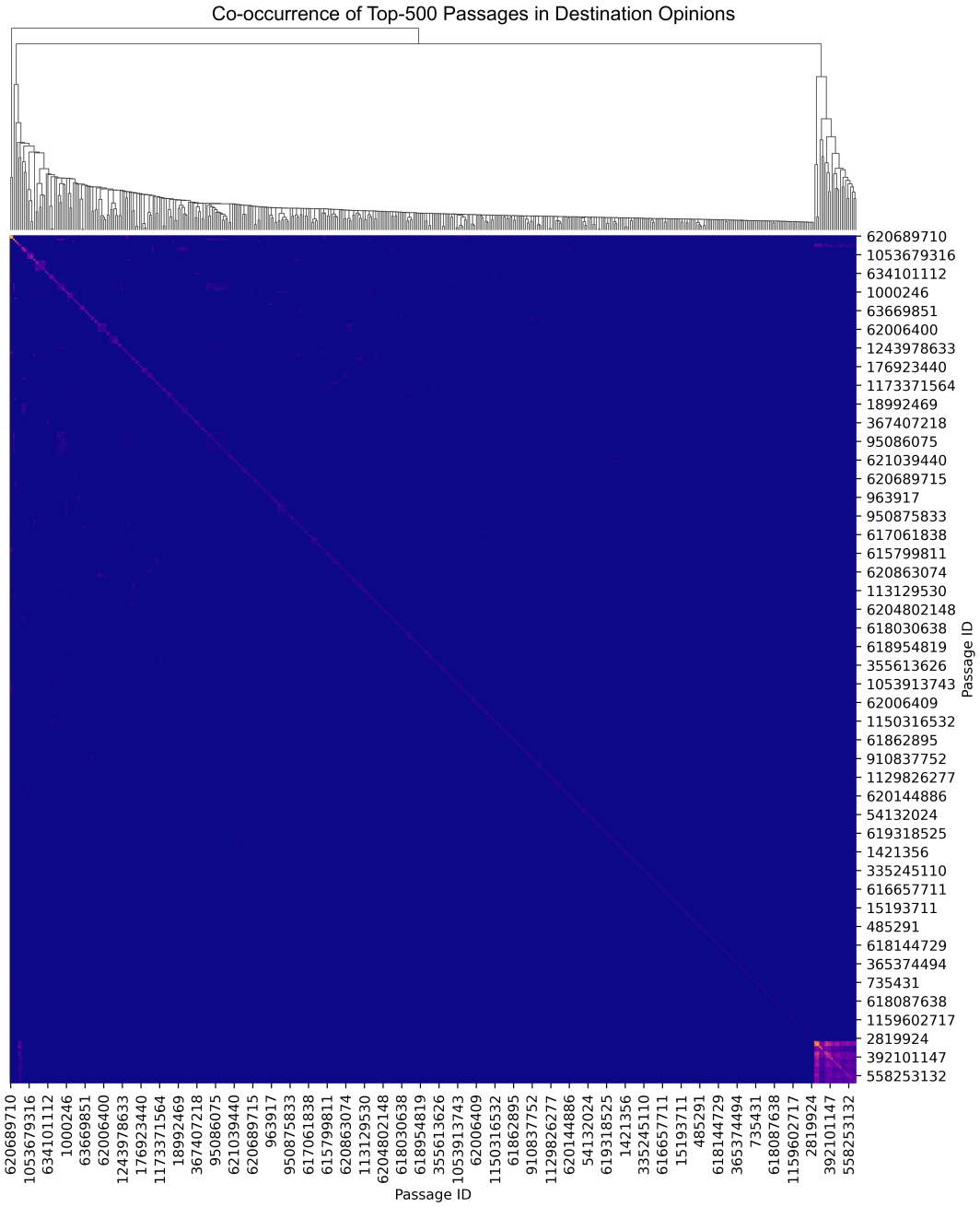

Figure 5: Hierarchical clustering of passage co-occurrence.