A Multi-LexSum release

A.1 Accessing Multi-LexSum

The dataset source files are stored in JSON format, and they are uploaded to Amazon S3 and can be downloaded publicly. Following the HuggingFace datasets library,[7] we develop one Python script (the `multi_lexsum.py` file) that handles both the downloading of the source files and loading them into easily usable format:

```
from datasets import load_dataset
multi_lexsum = load_dataset("multi_lexsum.py", name="v20220616")
# Download multi_lexsum locally and load it as a Dataset object
example = multi_lexsum["test"][0] # The first instance of the test set
example["sources"] # A list of source document text for the case
for sum_len in ["long", "short", "tiny"]:
    print(example["summary/" + sum_len]) # Summaries of three lengths
```

Currently, the `multi_lexsum.py` file can be retrieved from Multi-LexSum's GitHub repository: `https://github.com/multilexsum/dataset`. The authors are working on incorporating the script as part of the HuggingFace datasets library to further streamline the downloading and usage of Multi-LexSum. We include a similar instruction on the project website, `https://multilexsum.github.io`, which can be regularly updated to reflect the latest changes and future updates or erratum to Multi-LexSum.

A.2 Multi-LexSum distribution and maintenance

License Multi-LexSum is distributed under the Open Data Commons Attribution License (ODC-By). The case summaries and metadata are licensed under the Creative Commons Attribution License (CC BY-NC), and the source documents are already in the public domain. Commercial users who desire a license for summaries and metadata can contact `info@clearinghouse.net`, which will allow free use but limit summary reposting. The authors bear all responsibility in case of violation of rights, and confirm the dataset licenses. The corresponding code for downloading and loading the dataset is licensed under the Apache License 2.0.

Hosting and maintenance The authors are committed to providing long-term support for the Multi-LexSum dataset. At present, Multi-LexSum files are hosted on Amazon S3 by the authors themselves. In the event when the authors are not able to host the data, we will migrate the data to common dataset repositories (e.g., HuggingFace Datasets) and update the documentation and code. The authors will closely monitor the usage of the dataset, and develop necessary updates of bug fixes when needed.

For additional details, we refer readers to the dataset documentation (or the "datasheet" for datasets [19]) attached in Appendix H.

B Multi-LexSum summary writing and reviewing guidelines

B.1 Reading source documents

Different from multi-document summarization tasks in other domains [14, 16], Multi-LexSum summary writers are required to read several long documents and distill key facts therein. Strategically reading these documents saves time and effort and also improves the chances of successfully extracting the important information.

When summary writers first begin summarizing work on a case, they can orient themselves in several ways:

- *Web searches.* Summary writers are instructed to do a quick web search when starting a case. They may find news articles or blog posts that help explain what the case is about and developments in the case. They may also find websites that are updated with documents filed in the case.

[7] `https://github.com/huggingface/datasets`

Table 6: Rubrics for whether to include a source document in Multi-LexSum.

Always Include	The first complaint The last amended complaint Settlement agreements, consent decrees, and litigated decrees Opinions Orders granting temporary restraining orders or preliminary injunctions Orders granting or denying class certification Orders awarding or denying attorneys' fees Monitor reports
Sometimes Include *Based on Writers' Judgment*	Amicus briefs Orders that settle a contentious issue Motions/briefs if there's no order/opinion explaining their resolution
Rarely to Never Include	Answers Amended complaints that are not the latest one Orders that say nothing more than what's on the docket Motions/Briefs resolved by an opinion or order Attorney appearances Other orders

- *Recent judicial opinions.* At the beginning of opinions, judges often summarize developments in the case up to that point.
- *Notes from other summary writers.* When someone takes on a case picked by another person, that person usually includes notes, documents, or links to websites about the case. These can provide context or background, provide part of the narrative, or indicate some important events in the case.

Summary writers next skim their case's trial court docket to get an overview. A docket contains a chronological list of every document that is filed with a court in a given case, whether the filer is a party, the judge(s), or someone else. Each row or entry of a docket typically contains a number (its position in the list), a date, a title and short description, and—in many systems, including most federal court cases since 2003—a link to the filed document. The description conveys what kind of document has been filed and who filed it.

Based on the description, summary writers determine whether a docket entry relates to something important and warrants a deeper dive into the linked document. While there can be up to hundreds of entries in a docket, the writers are required to whittle the long list down to typically a dozen or fewer of the documents most essential for understanding the case, which constitutes the source documents in Multi-LexSum.[8] Table 6 describes the rubric for whether to include a document based on its type.

B.2 Writing summaries

The written summaries generally follow the order of events, as presented by the docket. The best summaries tell the story of the court proceedings. The student writes about the case's developments, progressing through the most important docket entries. If an entry's document is also important, the writer may also summarize the contents of the document as part of the narrative.

When writing the summary, writers also have a checklist of facts that they need to include, as illustrated in Table 1.

B.3 Reviewing summaries

After a summary is written, a reviewer with additional training then checks the summaries for factual accuracy. Reviewers may elect to go through a docket and verify that the summary includes all the important entries, but more often they just read the summary and keep an eye out for potential gaps in the narrative, and for events that are confusing or seem implausible.

[8] Some documents might be helpful for understanding the case and writing the summary but do not need to be added to the CRLC.

Table 7: Different Types of Source Documents in Multi-LexSum

Document Type	Avg. Docs Per Case	Description
Common Document Types		
Complaint	1.517 (0.88)	The document that starts a case and will usually be the first thing filed. Plaintiffs can also file amended complaints to add or subtract parties or claims.
Opinion/Order	3.300 (4.78)	Created by judges, opinions or orders memorialize rulings in the case.
Pleading/ Motion/ Brief	1.72 (7.05)	This broadly covers documents filed by the parties in order to make requests or explain their arguments.
Monitor/ Expert/ Receiver Report	0.221 (1.85)	Reports created by non-parties to help with the litigation in various ways. A "monitor" is a court-appointed expert, usually superintending compliance with a court order; an "expert" works for one or the other side, during the litigation; a "receiver" is an entity appointed by the court to run defendant operations because the defendant has somehow demonstrated incapacity.
Settlement	0.501 (0.83)	An agreement among parties that resolves some or all of the issues in the lawsuit.
Press Release	0.113 (0.46)	What it sounds like—-a press release.
Dockets	1.089 (0.41)	The docket is the court's index to everything that has happened in a case, in that court.
Less Common Document Types		
Correspondence	0.03 (0.25)	Letters NOT directed to the court. (In some jurisdictions (particularly in New York City), parties will conduct lots of litigation through letters to the court or "letter motions"—these are classified as motions or briefs, not as correspondence,.)
Declaration/ Affidavit	0.065 (0.8)	Documents in which someone provides information under penalty of perjury.
Discovery/ FOIA Material	0.018 (0.3)	Discovery material is evidence turned over by one party to another. The Clearinghouse usually doesn't collect it, so this document type is rare. FOIA materials are documents produced in response to a Freedom of Information Act (FOIA) request. These are also uncommon in the Clearinghouse.
FOIA Request	0.003 (0.09)	A request for information under the Freedom of Information Act or a state equivalent. This doesn't come up much in the Clearinghouse.
Internal Memorandum	0.01 (0.12)	An organization's internal memo (different from litigation documents with "memorandum" in the title). This is a rare category.
Magistrate Report/ Recommendation	0.044 (0.34)	Decisions from magistrate judges.
Statute/ Ordinance/ Regulation	30.018 (0.56)	A law or rule of government entity—federal, state, city or county, or agency. This document type includes policies created by prisons, school districts, police departments, immigration authorities, etc.
Transcripts	0.027 (0.32)	Verbatim transcripts of court proceedings or depositions.
Other	0.050 (0.38)	

Reviewers also ensure that the writing style conforms to the general practices as described below. In addition to checking spelling and grammar, they look for ways to keep the writing concise and somewhat free of legal jargon. Some examples of specific improvements:

- *Overall flow.* A summary tells a story, and so writers are encouraged to avoid too much repetition in terms of sentence structures. Following the chronology presented by the docket, beginning writers often start every paragraph with "on x date, the party did y," but that can make a summary less interesting and more difficult to read. In addition, while by default summaries present events

in chronological order, there are circumstances in which it makes sense for the narrative to tell pieces of the story in a different order.

- *Level of detail.* Documents that lay out the parties' initial arguments, the court's reasoning, and the outcomes of the case can be lengthy, so students need to make sure they are including enough detail for readers to understand what happened while still summarizing the documents in a concise manner.
- *Party descriptions.* Summary writers are instructed to describe the parties beyond just their role in a case. This tends to result in including the names of organizations or a description of individuals regarding why they would be involved in the case (e.g., "individuals incarcerated in prison" for a case about prison conditions).
- *Accurate terminology.* As with any discipline, words can have a particular meaning in the legal field, and so it is important with these summaries to accurately convey the events of a case. These types of fixes include using the right verbs around motions being "filed" and "granted," as well as including the full names of courts.
- *Avoiding legal jargon.* These summaries are read by people other than attorneys and law students, such as policymakers and reporters. This includes avoiding the idiosyncratic capitalization sometimes found in legal writing and court documents. Part of the task of summarizing a case is to translate legal technicalities into a story more suitable for a general audience.
- *Adding references.* If a summary discusses a judicial opinion that has a formal citation (e.g., 123 F.2d 456—meaning volume 123 of the case reporter Federal Reporter, second series, page 456), writers should include the citation, so that lawyers and researchers are able to find and cite the opinion for their own purposes. Writers may also choose to add links to news articles or blog posts that help add detail to a summary should a reader want to learn more.
- *Grammar and spelling.* In addition to correcting syntactical mistakes, reviewers also ensure that acronyms are either avoided or spelled out the first time. In addition, our style guideline requires reviewers to ensure that the summaries are written in past tense, which beginning writers may overlook because court documents and news articles may describe some events in present tense. (Writing summaries in the past tense avoids revisions in later years to change the tense.)

C Usability study system design

For our human evaluation of automatic summarizers applied to Multi-LexSum, we performed a usability study as described in Section 5. We initially presented the end-to-end generated long case summary to the legal experts, and found that it was far from helpful. As illustrated in Table 8, the generated text may contain factual incorrect or suspicious information that, according to expert users, needs more time to verify and correct than writing from scratch. We then designed our usability study system based on iterative feedback from CRLC experts to maximize practical helpfulness.

Illustrated in Figure 2, it breaks down case summarization into four steps: (a) docket reading and important entry selection, (b) summary outlining and content grouping, (c) source document reading and extraction, and (d) summary selection/rating/editing.[9] The system is designed to enable users to select relevant text snippets from the massive source documents to aid the model in salient information selection, and to decrease the model generation length to one paragraph at a time to reduce difficulty in both model generation as well as human editing burden. We detail each step:

(a) Docket reading and important entry selection The trial court docket is the governing document in a case: it contains a chronological list of all documents filed with the court in the case, including a brief text description of each document. It's important for human summary authors to read through the docket and arrive at a small subset of important docket entries and documents that should be included in the summary and may warrant further reading. This subset of candidate documents for in-depth reading are a superset of the "source documents" included in Multi-LexSum. The interface in Figure 2(a) supports this step by providing a tabular interface: each row ② corresponds to an entry in the docket, and users can click the checkbox in the first column ① when the referenced document is an important event that may also need in-depth reading.

[9] We implement the system except for the interface in (a) and (b), which is developed based on Notion: `https://www.notion.so/` and `https://github.com/lolipopshock/notion-df`

Table 8: Comparing the human-written long summary between the end-to-end generated summary from a BART model. Another expert annotates the incorrect (in red) and uncertain (in blue) spans of text in the model generation after comparing it with the ground-truth.

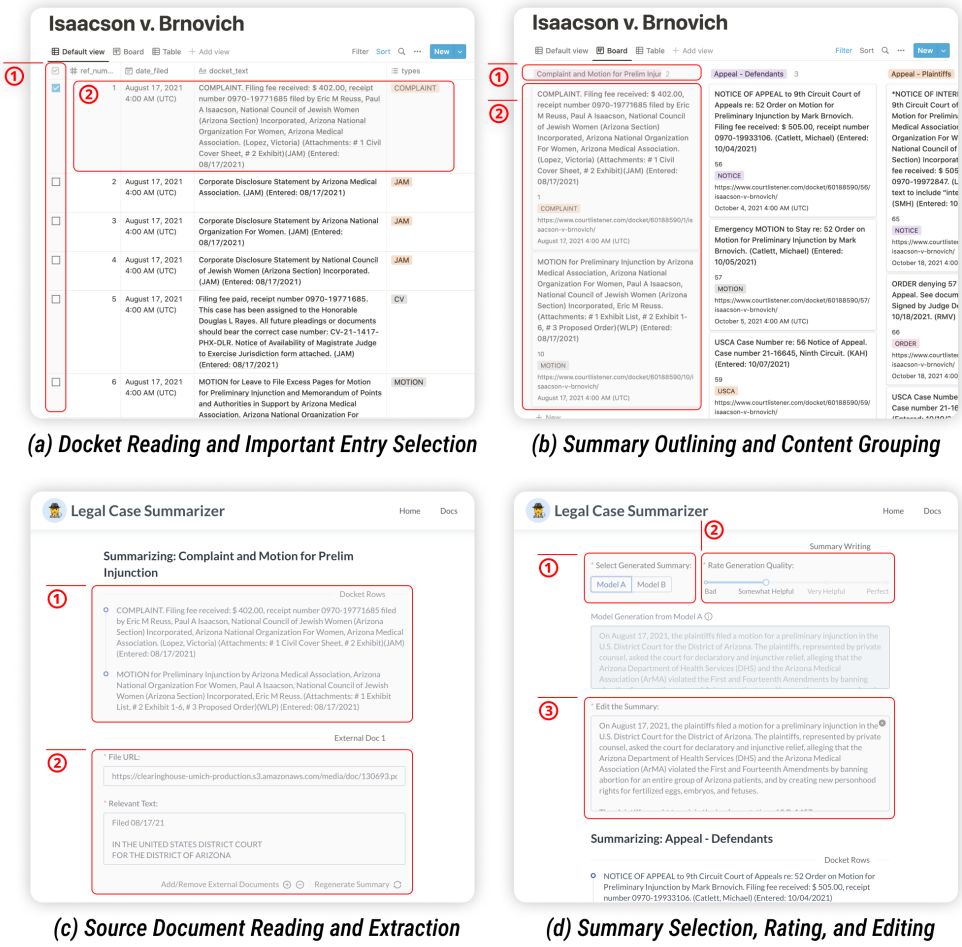
Ground-Truth Long Summary	**End-to-End Generated Summary**
In September 2004, the Los Angeles District Office of the EEOC filed suit against Casino Columbia Sussex Corporation and Gold River Operating Corporation, both of whom were doing business as River Palms Resort, a casino/resort in Laughlin, Nevada. According to the complaint, which was filed in U.S. District Court for the District of Nevada, the casino subjected female employees to sexual harassment, violating Title VII of the Civil Rights Act, and attempted to force out the older female employees so that they could be replaced by younger, more attractive women, in violation of the Age Discrimination in Employment Act. One of the female employees intervened as a plaintiff. In August of 2006, the EEOC moved for partial summary judgment, and the defendants moved to dismiss the case, alleging procedural irregularities in the EEOC's investigation. The EEOC's motion was granted in part and denied in part; the defendants' motion was dismissed. A three-year consent decree was entered April 25, 2008. The settlement request pay $200,000 to the plaintiff intervener and claimants ($5000 to plaintiff intervenor, and $195,000 to be divided among the claimants). It also enjoins defendants from discrimination and retaliation. Defendants must undertake training, must keep records of discrimination complaints, and must provide semi-annual reports to the EEOC. A month later, the parties jointly moved to dismiss all remaining claims with prejudice, except as to the consent decree enforcement. As no further activity is recorded, presumably this case closed in 2011.	On September 29, 2004, the Los Angeles District Office of the EEOC filed this lawsuit against Gold River Operating Corporation and Columbia Sussex Corporation in the U.S. District Court for the District of Nevada alleging discrimination on the basis of sex, female, in violation of Title VII of the Civil Rights Act of 1964. Specifically, the complaint alleged that the defendants subjected the charging party, a female employee, to a sexually hostile work environment and retaliated against her when she complained about the harassment. Following some discovery, the parties settled the lawsuit through a consent decree on May 6, 2008. The two-year decree, containing non-discrimination and non-retaliation clauses, required the defendants to: post a notice of compliance with Title VII, provide Title VII training for all employees, and pay $1,000,000 to be distributed among four individuals. The docket sheet does not show any further enforcement took place; the case was presumably closed in 2009.

(b) Summary outlining and content grouping After reading and selecting key source documents, human authors develop an overall understanding of the case's major events, each of which will generally correspond to one paragraph in the long summary, as per the CRLC's writing guidelines. The interface instructs writers to create "event groups" via the functions shown in Figure 2(b); only the candidate documents selected in the previous step are displayed in individual "cards". Writers specify a group name ① and drag the relevant cards into that group ②. The cards can also be assigned to multiple groups when appropriate.

(c) Source document reading and extraction The writers next begin in-depth reading for documents in each event group. In a typical unassisted workflow, authors would manually take notes consisting mainly of copied text snippets of key information from these source documents. Our interface, shown in Figure 2(c), includes the selected document and description from the docket in ①, and ② allows the human writers to perform this snippet-extraction workflow.

(d) Summary selection, rating, and editing Assisted by human selection of important documents and manually extracted text snippets, we run two summarizers, BART [36] and DistlBART [53], to produce the draft summaries. In Figure 2(d), the writer can pick the preferred model generation via ①, provide the 4-point rating[10] for the selected summary in ②, and edit the summary text in ③.

[10]Rating scale described in Section 5.



(a) Docket Reading and Important Entry Selection *(b) Summary Outlining and Content Grouping*

(c) Source Document Reading and Extraction *(d) Summary Selection, Rating, and Editing*

Figure 2: Illustration of the key components and functions in the user study system.

Because the models are provided human-selected salient text snippets, and are only required to generate a single paragraph at a time, the summaries generated in our setting are higher quality than those produced by end-to-end systems; this was verified via feedback from the study participants. Despite these efforts to improve generation quality, a mean rating of 0.43 for model output can be viewed as an upper-bound on the ability of modern end-to-end summarization models to produce usable summaries for this important task.

D Negative social impact

We believe that release of the Multi-LexSum dataset will have positive scholarly and societal impact. However, there are some possible negatives:

One intended use case of Multi-LexSum is to support training automatic summarizers for court documents. However, current summarization models are known to often make up facts in the generated text [33, 40], and it is difficult to differentiate between the "hallucinated" and faithful information in the outputs. If such summarizers are deployed at scale without having solved the hallucination problem, the factually incorrect summaries could lead to misinterpretation of the case by anyone using the application. In addition, the possibility of factual errors could undermine trust in the resource even if automatic summarization is only used sparingly. Though this concern could be resolved by future improvements of summarization models, we highlight this risk to encourage particular care when deploying such summarizers.

Table 9: Multi-LexSum train-dev-test splits.

	Source D	Long L	Short S	Tiny T	Total
Train (70%)	28,557	3,177	2,210	1,130	6,517
Test (20%)	7,428	908	616	312	1,836
Dev (10%)	4,134	454	312	161	927

Table 10: The average and standard deviation of BART models' performance on different test splits.

Strategy	Test Split	R-1$_{f1}$	R-2$_{f1}$	R-L$_{f1}$	BS$_{f1}$
Hold-out	**10%**	47.21(0.34)	23.05(0.17)	28.21(0.34)	38.72(0.30)
	15%	47.00(0.13)	22.89(0.13)	27.94(0.14)	38.30(0.26)
	20%	47.25(0.02)	23.00(0.08)	27.99(0.10)	38.29(0.12)
K-fold	**20%**	47.24(0.49)	23.12(0.25)	28.23(0.29)	38.79(0.33)

Moreover, as we discussed in Section 6, the cases in Multi-LexSum are drawn from a non-representative subset of all (civil rights) cases in U.S. courts. Models trained on this dataset will tend to adapt the language to the cases appeared in the dataset, which could be problematic when applied to other types of cases, or to cases in different legal systems from other countries. In the short term, we acknowledge this could lead to an "unfair" development of NLP methods that work only for certain types of cases (though we note that the cases in question are of both particularly high public interest and, because they are non-commercial, are unlikely to spur private profit-driven development). We strongly endorse efforts to increase the transparency of the court system, including free release of court documents and case summaries for other types of lawsuits in the U.S. and for different legal systems.[11]

E Multi-LexSum train-test split

We randomly split all cases in Multi-LexSum into Train (70%), dev (10%) and test (20%) sets, with detailed statistics reported in Table 9. The split strategy is verified in terms of (1) whether the test split is large enough for robust evaluation and (2) whether models are sensitive to a specific splitting of the data. We train and evaluate BART models under different split settings with the same hyperparameters, and we report the average and standard deviation of the ROUGE scores across different splits.

Determining the optimal test split size We test the same model (trained on 70% of the data) on different test split sizes (10%, 15%, and 20%), which we refer to as the "hold-out" strategy. Table 10 shows that the increased test split size leads to more stable test results (lower standard deviation). We decide to use 20% as the optimal test split size, while keeping 70% of the dataset for training.

Verifying model stability on different splits We conduct a 5-fold cross-validation experiment, where each time the model is tested on one fold and trained on the other folds. Given our decision from the earlier step to use 70% of the total data for training, for the purposes of this experiment, we also cap the amount of training data used in the cross validation experiments to this amount. Shown in Table 10, the low standard deviation indicates the models are not sensitive to specific random splits. Note that we used slightly different hyperparameters for this experiment than those for the main results reported in Table 3.

[11] For example, the SCALES-OKN project [45].

Table 11: Comparing LED performances when certain types of documents are removed in the input.

Model	Removing Documents				Summary Quality						
	Complaint *1.5 docs/case*	Opinion *3.3 docs/case*	Docket *1.1 docs/case*	Random Docs	R-1$_{f1}$	R-2$_{f1}$	R-L$_{f1}$	BS$_{f1}$	P-value[1]	Input Words[2]	Pred Words
LED-16384					49.07	25.17	29.40	40.05	-	9617	310
LED-16384				1 Doc	49.06	24.76	29.01	39.63	0.299	8919	326
LED-16384	✗				47.94	24.24	28.74	39.83	0.087	8305	296
LED-16384		✗			48.09	24.47	28.81	39.48	0.027	9306	298
LED-16384			✗		49.11	24.87	29.14	40.09	0.455	9303	317
LED-4096					47.75	24.13	28.89	39.10	-	2793	295
LED-16384				2 Docs	47.61	23.79	28.50	38.88	0.001	8201	303
LED-16384	✗	✗			46.36	23.21	28.15	38.46	0.000	7388	277
LED-16384	✗		✗		47.99	23.97	28.67	39.19	0.003	7559	299
LED-16384		✗	✗		47.18	23.39	28.24	39.14	0.000	8688	299
LED-16384	✗	✗	✗		46.38	22.44	27.81	37.72	0.000	6371	294

[1] We conduct two sample t-test for R-2$_{f1}$ scores against the baseline LED-16384 model result, where none of the documents are removed. For each experiment, the two populations are the individual R-2$_{f1}$ scores for each case.
[2] The number of actual words in the input might be smaller than 16384, which is the maximum number of Byte-Part Encoding tokens in the model input.

F Model Training Details

Text Preprocessing We use the pyxpdf tool [12] to extract the text from the PDF files of the court documents. A small percentage (11%) of these documents are not born-digital, meaning they are scanned and contain OCR'd text. We do not perform any special processing for unique aspects of our corpus (such as citations or legal terminology), but these are interesting possibilities for future improvements.

Model and Hyperparameter Choice For our baseline models and settings, we choose models that are state-of-the-art in similar summarization dataset papers (e.g. BookSum [35], Multi-XScience [39], GovReport [28]) and additionally add models like PRIMERA [59] and LED [1] that can handle long input context which is a feature of our dataset. For our training settings, we searched different hyperparameter configurations on BART and found the model validation performance converged after 6 epochs for training. We then used the same set of training hyperparameters on all the experiments for both BART and PEGASUS. The hyperparameters for "long models" like LED and PRIMERA are slightly different, based on recommendations in the respective papers. We release our model training scripts for reproducibility.

G Additional experiments

G.1 Understanding the importance of input documents based on their types

Detailed in Table 7, multiple types of documents are created and collected during the course of a case. As they are written and formatted distinctively and may detail different aspects for the lawsuit, they might contribute to the final summaries in different ways. For example, dockets include all major events that happened during the case, while other documents may only contain specific details; complaints and opinions may summarize certain parts of a case, which can be directly of use in the final case summary.

To understand the how different types of document contribute to the final summary, we perform an ablation study to remove a specific type(s) of document from the summary input and train and evaluate the models. Shown in Table 11, we report model performance after removing the complaint(s), option(s), or docket(s), as they are typically the most important documents according to the summary writers. As a baseline, we also evaluate when random documents (other than the aforementioned three categories) are removed.

[12] https://pypi.org/project/pyxpdf/

We find that firstly removing any document from the source input will negatively impact the output summary quality. Comparing removing a complaint, docket, opinion, or a random document (of other types), it appears complaint and opinion have a stronger influence on the summary quality. We see that the ROUGE score after removing a complaint or an opinion appears to be lower ($p<0.10$, two-sample t-test) than the full-document ROUGE, whereas it is not significantly lowered by removing a random other document or even (perhaps surprisingly) the docket.

When we remove multiple document types from the input, there's also a more significant drop of ROUGE scores when complaint and opinion are removed together. Interestingly, despite the test model being a LED-16384 with longer input context, the summary quality is worse than the LED-4096 model when either complaint or opinion is not present. Taken together, the experiments demonstrate the unique challenges in this dataset for processing legal documents, and suggest the potential of domain-specific modeling techniques.

H Multi-LexSum datasheet

Please see next page.

Multi-LexSum Dataset Sheet

We develop the dataset sheet based on the template (v7) from Gebru et al.[1] The Multi-LexSum dataset can be accessed via the ***project website*** https://multilexsum.github.io or the ***Github Repo*** https://github.com/multilexsum/dataset.

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The Multi-LexSum dataset was curated to facilitate the development of automatic summarization methods for civil rights lawsuits.

Recent advances in document summarization have led to impressive results in generating a short description for passages typically in hundreds of words. However, the source inputs for summarizing civil right lawsuits are considerably longer: they can contain up to 70k words on average. It's still a crucial challenge for existing models to handle such long input context. Multi-LexSum is constructed to serve as a benchmark for this "long" document summarization scenario.

Additionally, human readers have different needs for summaries---ranging from one sentence to a paragraph or multi-paragraph narrations. Existing datasets only provide summaries of one granularity for a given input source, while Multi-LexSum contains summaries of up to three different levels of detailedness for one case, enabling novel research in this direction.

We also consider the summaries in Multi-LexSum to be "gold" summaries: each summary is written and reviewed by legal experts following a detailed instruction (detailed in Appendix B in the paper). In contrast, the reference summaries in previous datasets are usually obtained via automatic linking of contents, e.g., using the first sentence or summary bullets as the target summary for a piece of news, or automatically extracting and linking scientific paper abstracts and citing sentences.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset is created by the collaboration between Civil Rights Litigation Clearinghouse (CRLC, from University of Michigan) and Allen Institute for AI. Multi-LexSum builds on the dataset used and posted by the Clearinghouse to inform the public about civil rights litigation.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The underlying Civil Rights Litigation Clearinghouse data has been funded by numerous entities over its 15 year history, including:

- University of Michigan Law School

[1] Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92.

- Washington University in St. Louis School of Law, Center for Empirical Research in Law
- Arnold Ventures, “Improving Criminal Justice Reformers’ Use of Litigation Information, Documents, and Insights” (2021-2023), $400,000.
- Vital Projects Fund, “Revamping the Civil Rights Litigation Clearinghouse” (2021), $100,000.
- Proteus Fund, “Revamping the Civil Rights Litigation Clearinghouse” (2021), $50,000.
- National Science Foundation SES-0718831, “The Litigation Process in Government-Initiated Employment Discrimination Suits” (2007), $213,999.

The construction of the Multi-LexSum dataset was also funded in part by NSF Convergence Accelerator Award ITE-2132318.

Any other comments?

None.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance in the dataset represents a lawsuit and contains a set of source documents (extracted from a collection of public PDF files from U.S. federal courts), the corresponding summaries manually written by legal experts, and metadata describing attributes about the lawsuit.

How many instances are there in total (of each type, if appropriate)?

There are a total of 4,539 instances in this dataset, along with 40,119 source documents and 9,280 summaries. A detailed breakdown can be found in Appendix E of the paper.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances, i.e., sampled from all civil rights lawsuits. It is not random; the CRLC includes cases only if they are (a) injunctive (that is, seeking court-ordered behavior/policy change) or (b) class-actions, or effectively similar to class actions (that is, adjudicating the rights of groups of people), and only in certain topics (for a list, see https://clearinghouse.net/case-types). In addition, for this project, the dataset is limited to (a) federal cases with (b) computerized dockets and documents (in a small number of cases, docket coverage may begin mid-case; these will be flagged in a planned update to the dataset). In addition, because there is no reliable way to locate every case that meets the above criteria, the sample is non-representative even among cases that fit CRLC's inclusion rules: CRLC is more likely to include cases where the plaintiff wins because such cases typically last longer and receive more attention. (As such, we clarify in the paper that performance might not generalize to under-represented cases [e.g., where the defendant wins].) We additionally provide case metadata to facilitate future diagnosis of this bias.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance contains the following data:

1. Source documents text for a case. The text is extracted from the source PDF documents. We include the title and the type of the documents as well.
2. Summaries for a case, which can come in up to three lengths:

a. Long summaries typically contain multiple paragraphs, covering the case background, parties involved, and proceedings. Major case events and outcomes typically receive a paragraph each.
b. Short summaries have only one paragraph with a shorter description of the background, parties involved, and the outcome (so far) of the case.
c. Tiny summaries are one-sentence summaries intended to appear on Twitter to describe the case at a specific point in its history.

3. We contain the metadata for each, including but not limited to:
    a. The author(s) of the summaries
    b. Case type
    c. Case name
    d. Filing year
    e. Court & judge
    f. Plaintiff information
    g. Plaintiff attorney information
    h. Defendant information
    i. Causes of action
    j. Issue tags
    k. Prevailing party
    l. Relief information, including source and form

Is there a label or target associated with each instance? If so, please provide a description.

N/A.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Each instance may contain summaries up to three different granularities (long, short, tiny). All the instances have a long summary, but some lack short or tiny summaries (or both). Which instances are missing short or tiny summaries is mostly a function of the date on which the summarization was done at CRLC; the summary writers were not always required to produce shorter summaries. In addition, while key metadata is available for every case, there is some missing data on less important fields, for similar reasons.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Each individual instance is considered to be independent in our dataset, though some writers may have contributed to summaries for different instances. We include an ID for the summary writer(s) for each instance.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes. The dataset is split into train/dev/test in the released version. We detail the statistics in Appendix E of the paper.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Sometimes the source document PDFs were scanned and not digital-born. OCR was required to extract the text in these documents, which can be a source of noise. There is human error in the human summaries and human-coded metadata.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is entirely self-contained, but it also links to CRLC, using a case-specific id. This link is not necessary (or even useful) to use the dataset. CRLC is a long-term project of the University of Michigan, funded into the future, and its pages are archived at the Internet Archive.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

There is no confidential information in our dataset; all the source documents are posted (albeit some behind a paywall) by the federal courts, available in courthouses for public inspection, and uncopyrighted and fully public.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Yes. Civil rights documents describe alleged (and eventually perhaps proven) violations of the law that harm the plaintiffs who bring the lawsuits. For example, some might concern medical neglect in prison, police violence, race or sex discrimination by employers, harmful results of abortion restrictions, and the like. In other words, the documents may contain offensive content, as they describe case allegations; these are often the central topic of the lawsuits. The data, however, should not cause any additional anxiety because every document is already made public by the federal courts, and every document and summary is already posted by CRLC.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The underlying lawsuits relate to people. In addition, the summaries and metadata were written by people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

For cases that allege sex, race, or national origin discrimination, the affected sex, race, or national origin is coded.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Yes. Many of the underlying cases were filed by and/or against one or more natural person; the documents name these individuals, and their lawyers. Again, all the information is already public.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Yes, where the underlying cases relate to these issues, the documents disclose them. For example, in a case alleging race or religious discrimination, the race or religion of the plaintiff will be described in the documents (especially in documents that were filed on the plaintiff's behalf). In cases addressing criminal justice issues or jail or prison conditions, the criminal history of the plaintiffs may be relevant and discussed. However, the federal courts have rules against posting social security numbers and the CRLC has also done automated checks of the documents, as a backup, to ensure that social security numbers are not posted.

Any other comments?

Every case in the dataset has a docket pulled from the federal court's electronic docketing system. A very small number of cases *also* have scanned PDFs of earlier, non-digitized dockets. These are not currently tagged, but will be so in a future correction. In addition, for 118 cases, identified by metadata, the posted dockets are incomplete.

COLLECTION PROCESS

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

For each instance in the dataset, it contains the source documents (a collection of public PDF files from U.S. federal courts) and the target summary(s). The source documents are directly observable and we developed the PDF parsers to extract the raw text from the PDF documents (detailed in PREPROCESSING/CLEANING/LABELING section). The target summaries were written and the metadata entered by legal experts summarizing the source documents following instructions.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Manual human curation, verified by legal experts, was used. We detail the manual and summary writing guidelines in Appendix B of the paper.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

N/A. See our discussions above in the COMPOSITION section.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Legal experts from University of Michigan (and for earlier cases, from Washington University in St. Louis), including legal scholars, attorneys, and students. The CRLC's director is faculty and paid as such–she has not received additional compensation for her work on the Clearinghouse. CRLC has sometimes hired part-time attorneys to assist in the project; they are paid under University staff contracts. Law students have three different compensation methods: some do the work for credit; some are paid; some volunteer. For those who are paid, law students at the University of Michigan are currently paid $15/hour; the rate was a little lower in prior years. Some metadata coding of documents was performed by undergraduates; they are currently paid $12/hour, but, again, the rate was a little lower in prior years. For volunteers, the project qualifies as one of many "pro bono" projects; law students are encouraged to volunteer for such projects, and some states require several dozen hours of pro bono volunteer time as a prerequisite for attorney licensure.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Multi-LexSum contains court documents from from the 1950s to 2021, heavily concentrated from 2000 to present. The contained case summaries were written between 2005 and 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Schlanger explained the project to staff of the University of Michigan Institutional Review Board, and they responded that it did not need IRB approval.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The dataset does not relate to people as human subjects. The cases that underlie the dataset relate to people. And people wrote the target summaries and entered the metadata.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

All data were obtained from the CRLC, which in turn obtained documents from the federal court system and assigned legal scholars, lawyers, and law students to write the summaries.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No, individuals whose names and circumstances appear in the court documents have not been notified; the documents are public and posted by the federal courts, and are used and reposted by many sites, including CRLC. It is available to parties in court to request that documents be "sealed"--that is, made non-public. But the standards for sealing are quite stringent, because the public has a First Amendment right to know what happens in court. Similarly, parties or participants can request to proceed by pseudonym (Jane Doe, for example), but the First Amendment and policy commitments to government transparency limit such permission to cases with significant and unusual privacy interests, and to cases involving minors. In any event, the decision to file a document under seal or to seek to proceed by pseudonym is made by the affected party in court, and then adjudicated by that court. CRLC and by extension this dataset do not undertake further review.
The individuals who wrote the summaries and entered the metadata were notified that their summaries and work would be publicly posted (using their names) by the CRLC. This is part of appropriate acknowledgement of their work and authorship; they are not data subjects but research collaborators. Their names have long been posted at https://clearinghouse.net/people.
The individuals who wrote the summaries and entered the metadata were not and cannot be notified that their work are included in this dataset, because that summarization/coding was

done over a period of more than 10 years, nearly all of it long before this dataset/development was contemplated. In any event, this dataset does not include their names, just an identifier.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Individuals whose names and circumstances appear in the court documents have not consented separately to inclusion at CRLC or in this dataset. The documents are public and posted by the federal courts, and are used and reposted by many sites, including CRLC. However, for any who requested it over the past 15 years (by easily available email), the documents in question were either flagged using a robots.txt notice to guard against crawling, or redacted. The dataset does not include any of the flagged documents. For any individuals identified as a summary writer by the CRLC, they agreed to participate in the project and have their authorship appropriately acknowledged and their contribution recognized during training. In any event, this dataset does not include that identification, just an id code.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

The authorship acknowledgement is not experimental data, but appropriate recognition of intellectual contribution. If one of the former CRLC research assistants or researchers wanted their name removed from a summary, they could reach out to CRLC and it would quickly accommodate that request.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No; the authors of summaries are not data subjects but collaborators. They are listed as such at CRLC: https://clearinghouse.net/people.

Any other comments?

None.

PREPROCESSING/CLEANING/LABELING

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes. We developed software to extract the text from the raw PDF files from court documents, and we store the method for extracting the document text in the released data as well.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes. The raw data will be released on a later date than the dataset release date (specified in the DISTRIBUTION section). Because the PDFs files combined are significantly larger, we are working on the best solution for long-term hosting and maintenance.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes. The PDF extraction code will be released in the project's Github Repo.

Any other comments?

None.

USE

Has the dataset been used for any tasks already? If so, please provide a description.

In our paper, we demonstrate that the dataset can be used for training models to that can:

1. produce summaries for a legal case from the source documents,
2. perform controlled automatic summarization that can produce summaries of different granularities.
3. condense a long summary to a short version.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

N/A. If papers are produced, links will be posted at https://clearinghouse.net/search/resources/?resource_types=6132

What (other) tasks could the dataset be used for?

We envision the dataset can also be used for the following scenarios, including but not limited to:

1. Large-scale pre-training for legal document understanding models. As we provide a massive collection of documents, they can be used as the pre-training corpus for large language models for understanding legal text.
2. Information extraction models from legal documents. As we provide metadata for each case (e.g., causes of actions, outcomes), and they are based on the source documents, a potential use case might be training an information retrieval model for these fields from the source documents.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

As discussed above and in the paper, the cases in the dataset are a non-representative sample of all civil right lawsuits, with various inclusion criteria (electronic availability, topic area, injunctive and class litigation) and also some practically-produced selection bias: CRLC is more likely to include cases where the plaintiff wins because such cases typically last longer and receive more attention. As such, we clarify in the paper that performance might not generalize to under-represented cases (e.g., where the defendant wins). The dataset should not be used for training models to predict the outcome of a lawsuit.

Are there tasks for which the dataset should not be used? If so, please provide a description.

Given the constraints mentioned above, the dataset should not be used for training models to predict the outcome of a lawsuit.

Any other comments?

None.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset will be publicly available on the internet.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be uploaded to an Amazon S3 bucket on AWS, and people can download it publicly via the provided link. In addition, we release a (Python) script for loading and using the dataset files in https://github.com/multilexsum/dataset. We plan to incorporate the dataset to the Huggingface Datasets library for easy access in the future.

When will the dataset be distributed?

Since June 16, 2022.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The Multi-LexSum dataset is distributed under the Open Data Commons Attribution License (ODC-By). The case summaries and metadata are licensed under the Creative Commons Attribution License (CC BY-NC), and the source documents are already in the public domain. Commercial users who desire a license for summaries and metadata can contact info@clearinghouse.net, which will allow free use but limit summary reposting.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

N/A.

Any other comments?

None.

MAINTENANCE

Who will be supporting/hosting/maintaining the dataset?

Zejiang Shen is supporting the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Zejiang Shen will be the main contact for the dataset, and the up-to-date contact information can be retrieved at www.szj.io. Questions pertaining to the Allen Institute for AI's involvement in curating or maintaining this dataset should be directed to dougd@allenai.org. Questions pertaining to the CRLC's involvement in curating or maintaining this dataset and/or the CRLC's past and ongoing efforts to produce and disseminate these summaries should be directed to info@clearinghouse.net.

Is there an erratum? If so, please provide a link or other access point.

Currently we haven't found any errors in the version (to be released). If we do, we will post the erratum and update information in the dataset website and Github repo.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Though we have no concrete plans for dataset updates, we envision there will be updated versions for error corrections and inclusion of additional data. If there are any updates, the updated information will be posted on the dataset website and Github repo.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes. The public links of the dataset files contain version information. We will keep the links for older versions available after the release of newer versions.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Given that we released the dataset under the Creative Commons (CC BY-NC) license, others should feel free to extend and build upon the dataset. Any contributions to the dataset can happen in the form of Github Pull Requests / Issues: the contributors can submit the changes or suggestions, and we will monitor and moderate them monthly.

Any other comments?

None.