LegalText+: Advancing the Data Frontier for CoT Extraction and Reasoning in Legal NLP

Jesse Woo[1] **Lawrence Leung**[1] **Nikhil Ghosh**[1] **Elizabeth Kim**[1] **Gregory Hunter**[1]
[1]Columbia University
{jw4202,lsl2162,nrg2156,ek2935,geh2129}@columbia.edu

Abstract

We present a pilot dataset of legal arguments extracted from court briefs filed in U.S. Supreme Court cases. These legal arguments represent an explicit human chain of thought reasoning for legal NLP, an area that is generally underrepresented in terms of useful training data. Briefs are a unique source of legal arguments because by convention the section headings are a semi-structured summary of the brief's argument. Therefore, a well-written brief effectively offers a concise and inherently semi-structured summary by way of its table of contents. We sourced briefs from the pile-of-law dataset, which entailed significant data cleaning and a fairly complex extraction process. We developed a step-by-step workflow, and tasked a pool of human annotators with extracting arguments from the table of contents, labeling the argument's conclusion, and rating the salience of the arguments. In parallel, we also tested a model-assisted approach to extract arguments and classify conclusions. In practice, we found that the models did not perform well enough to be relied on at scale, but were nonetheless able to evaluate their standalone performance while also doing a detailed analysis of their annotations compared to our human-annotated dataset. Legal NLP is an understudied field that, despite the large amount of text generated by lawyers, lacks good data. By providing a substantive and high-quality legal NLP dataset, this work could open up entirely new lines of research in the fields of legal NLP and human reasoning— with and without the aid of LLMs.

1 Context

As shown by ChatGPT, large language models (LLMs) have become quite capable of generating text that has both high fluency and faithfulness, but often fail to operate in a way that aligns with human reasoning or chains of thought. Chain of thought (CoT) reasoning is an important cutting edge area of research that has shown to improve model performance and accuracy on tasks that have previously been difficult, such as solving mathematical word problems [9].

While human reasoning is often implicit in an NLP corpus, this dataset would be a source of explicit human reasoning on a specific task: legal argumentation. The convention of legal brief writing is to have the heading of each section and subsection state a concise version of the argument that would be contained in that section, with each section building toward the ultimate argument. So in a good brief, the table of contents ought to serve as a fairly structured form of human legal reasoning. This approach was partially inspired by Zhong and Litman's approach of exploiting the inherent semi-structured nature of legal documents in their unsupervised extractive approach for legal case summarization [13]. As the legal field produces and consumes massive amounts of text data, legal NLP is poised to benefit significantly from advances in LLMs and related datasets [12].

There are a few existing datasets related to law and NLP, though none that explicitly capture structured legal reasoning or chain of thought. The CaseHOLD [10] dataset is a short excerpt from a judicial opinion paired with a set of multiple choice answers which purport to represent the holding of the case.

Preprint. Under review.

A holding is a short summary of the case's rule and facts that forms the core of the precedential value of the case. LexGLUE [5] is a law-themed General Language Understanding Evaluation benchmark. It consists of seven datasets and tasks: the European Court of Human Rights cases tasks A and B, the Supreme Court of the U.S. (SCOTUS) cases, European Union Legislative text, American contract provisions, Terms of Service from several large web platforms, and the CaseHOLD dataset. The Caselaw Access Project offers free, programmatic access to millions of U.S. cases from several federal courts and all 50 states. Zhong et. al published JEC-QA, a question/answering dataset based on the National Judicial Examination of China [11].

Hugging Face has a dataset called pile-of-law [6] that includes some briefs filed in federal courts from the RECAP archive of the Free Law Project, as well as Supreme Court filings that are housed on scotusblog.com. Multi-LexSum is a dataset of expert annotated legal summaries of telescoping granularity [8]. The authors had lawyers and law students read multiple longform documents for civil rights legal cases from the Civil Rights Litigation Clearinghouse, and summarize them in long (multiple paragraph), short (single paragraph), and tiny (single sentence) summaries. The dataset contains 9280 summaries in total. This is an impressive feat notable for its use of human experts, but the summaries are not explicitly meant to capture CoT reasoning.

One previous chain-of-thought rationale dataset is the CoT Collection [7]. Building upon 9 previous NLP datasets, this dataset comprises 1060 NLP tasks/prompts with 1.88 million CoT rationales. The format of the CoT collection is primarily three chunks of text starting with a source, a target, and a rationale on how to reason from the source to the target. For example, on a math related task, the source will have a word problem followed by a question, the target would be a numerical number, and the rationale are the steps taken to derive the solution. Another related dataset is LogiCoT [**?**], where researchers gave GPT-4 a set of instructions (usually request a derivation or reasoning) and an input (a problem with a solution). From here, GPT-4 returned a reasoning blurb. This instruction tuning dataset is similar to this project, but the domains do not overlap as LogiCoT is not based on legal reasoning.

2 Format

The inputs for our dataset are legal briefs for cases in front of the Supreme Court of the United States (SCOTUS). The briefs are sourced from the pile-of-law dataset available on HuggingFace. We used regular expressions to truncate and extract only the table of contents and conclusions from each brief.

We relied on briefs of Supreme Court cases for two main reasons. First, SCOTUS briefs are the most widely available. Second, we believed that the legal reasoning in the briefs were likely to be of higher quality, as typically only the most accomplished attorneys are admitted to practice in front of the Supreme Court. As we will discuss in the Collection and Limitation sections, this is generally true, although we encountered some data quality issues with writs of certiorari. We included briefs on merit cases (those argued in front of the Court on the merits), as well as Amicus briefs and briefs for writs of certiorari. An Amicus brief is a supplemental brief filed by groups or individuals who are not themselves litigants, but have some interest in the outcome of the case. A writ of certiorari is an argument about whether the Supreme Court should hear a case. All three types of briefs contain arguments and are thus useful data, although qualitative inspection reveals that briefs of merit cases are generally of the highest quality.

The output is a set of legal arguments that are represented by the brief's section headings. Task 0 was for annotators to confirm that the argument and conclusion were properly identified in the snippet from the brief. Task 1 was then to extract the portion of the table of contents that represents the legal arguments, which is only inconsistently marked with a separate header in the table. Next, task 2 was to label the conclusion of the argument as exactly one of [Affirm, Reverse, Remand, Grant, Deny, Other]. Task 3 was to rate the salience of the reasoning behind the conclusion based on the arguments from the table of contents. We discussed but did not have time to implement masking the conclusions or individual aspects in the argument body to make the dataset more appealing to researchers, and also pairing the outcome of the brief with the outcomes of the related Supreme Court decision.

3 Data Collection

We originally intended to use a model-in-the-loop approach to extract the arguments from the tables of contents, but this approach proved not to be reliable enough to scale. Instead, we used simpler string parsing techniques and regular expressions to filter from the pile-of-law dataset. This approach also had challenges, but was reliable enough to provide sufficient data for the pilot. This section will describe the pile-of-law filtering, the model-in-the-loop approach, and our annotation methodology. Our unexpected difficulties related more to data extraction and cleaning, both from the pile-of-law dataset and with our model-in-the-loop.

3.1 pile-of-law filtering

We sourced our data from the pile-of-law/scotus_filings dataset, hosted on HuggingFace. The dataset contains over 63,000 entries from supremecourt.gov, the majority of which are not briefs but rather short filings (e.g. a motion to change a hearing date). We worked with the training set, which contains over 47,000 rows. Metadata is very sparse; it includes only the raw text of the filing as a single long string, the 'created_timestamp', 'downloaded_timestamp', and the url of the original document. Document type is not given, necessitating a significant search process to find briefs.

After downloading the SCOTUS filings from pile-of-law, the data were filtered to exclude documents containing 'writ of certiorari'. Unfortunately this did not filter all writs of certiorari however. Next, all documents that did not contain 'TABLE OF CONTENTS' were excluded. The specification of all capital letters in the phrase was selected because the actual table of contents had such a style in the data. This was useful in avoiding false positives of any lowercase 'table of contents' being referenced or discussed in the contents of the paper. At this point, it observed that some files had multiple tables of contents. For the purposes of this project, it was decided to focus on the documents with a single 'TABLE OF CONTENTS'. This process successfully yielded 545 entries.

As a sanity check, associated PDFs for the respective cases were downloaded from the supreme court website. This was done by cross reference the filename with the metadata from Pile of Law. After a certain number of requests, the website denied the downloads. This was worked around by logging onto a VPN before running more code to re-download those files which had been missed. The value of the PDFs served as an easy formatting of the contents to a more human-friendly layout, as opposed to being long strings with many new characters amongst other things. It also provided a good check to see whether there were differences between the actual PDF documents of the filings and the data in the Pile of Law dataset.

From here, the tables of contents were extracted using regex patterns. The first pass involves using 'TABLE OF AUTHORITIES' as a marker to signify the end of the table of contents. It was observed that this happens for a vast majority of cases. However, there was an issue regarding the capitalization of the phrase. In some tables of contents, the table of authorities was also listed. When all entries of that table of contents were capitalized, this created a false positive for 'TABLE OF AUTHORITIES' acting as a stopping. The second pass involved checking for the length of extracted text. If it was around the same length as 'TABLE OF CONTENTS', the output would be deemed to be incomplete. To remedy this, the other pattern to check for was "CONCLUSION" or "Conclusion". These two were almost always present in the table of contents. The reason why this second condition was not solely used is because it was seen that applying both yields better results.

As for extracting the conclusion, the procedure was to iterate backwards until the first instance of "CONCLUSION" or "Conclusion" was found. The fully lower-case variant was not used to avoid false positives. Once this was found, the program would iterate forward in the text until some variant of 'respectfully submitted' was reached. The contents in between would be the extracted conclusion. This was not a fool proof method, but it did relatively well, given the unstructured nature of the data. Many cases had cleanliness issues such as empty strings, extremely short content, or Null values, which were removed from the result. An entry for a brief might contain an empty string or null value because the PDF contained an image of the text, rather than actual text itself, which apparently rendered it unreadable to whatever scanning technique was originally used with pile-of-law. For example, one of the conclusions in the PDF and in the data itself was missing, yet if a human looked at the PDF, they would see the conclusion there as an image but would be unable to select the text. Cases such as these were also filtered. In addition, pile-of-law contains many duplicate entries with

identical data but different indices. After all cleaning and filtering, duplicates were manually removed, yielding 264 results.

3.2 Model-in-the-loop

Once the data had been extracted from pile-of-law, a copy was formatted for human annotators to label. Another copy was given to two large language models (Llama2 and Mistral). Both of these models are their 7 billion parameter version, and both were given tasks similar to the human annotators. The models were instructed to give a zero-shot effort of extracting the arguments from a table of contents and classifying the type of conclusion. While slightly different instruction tokens were used for each model the system prompts were very similar: "You are a legal expert who specializes in classifying the conclusion types of legal briefs. Write out your short and succinct!" vs "You are MistralOrca, a legal expert who specializes in classifying the conclusion types of legal briefs. Write out your short and succinct!". The instructions were the exact same, "Classify the conclusion into one of the following categories: affirm, deny, reverse, remand, other, or incomplete. Return only the classification. text" was the prompt used in conclusion classification. “Extract the explicitly stated arguments from the following table of contents. Return only the arguments or ’None’ if there are none. text” was the prompt used for argument extraction from the table of contents. When extracting from the table of contents, the system prompt only changed to telling the model it “specializes in extracting arguments of legal briefs.” Unfortunately, as we will discuss in the analysis, the models’ performance in this task was not reliable enough to be trusted without human supervision.

3.3 3.3 Annotation process

We had 12 volunteer annotators from the class. We asked each person to annotate 20 briefs. Annotators received detailed instructions (included in Appendix B) covering four tasks. Task 0 was to confirm valid extraction of the table of contents and conclusion text. This task was necessary given the difficulty of cleanly identifying briefs and tables of contents. Task 1 was to extract the arguments from the table of contents. The arguments were inconsistently marked within tables of contents, necessitating human extraction (or a better performing model, more on that below). Task 2 was to classify the outcome of the brief as exactly one of [Affirm, Reverse, Remand, Grant, Deny, Other]. Task 3 was to rate the salience of the reasoning behind the conclusion based on the table of contents. In other words, whether the logic of the outcome is clear solely from the arguments in the table of contents. We asked annotators to rate on a Likert scale from 1 to 5, where 1 is ‘Not Clear At All’, and 5 is ‘Extremely Clear’. We initially experimented with using a custom web portal to collect and store all annotations, but eventually settled on a simpler solution with google drive, where each annotator was given their own google sheet to fill out.

We asked annotators to complete a short questionnaire after completing their work and received responses from 8 of 12. The questionnaire asked how many minutes the annotator took to read and comprehend the instructions, complete all annotations, and complete a single brief (tasks 0-3). It also asked which task was quickest, which was slowest, and gave a free text option for additional feedback.

Most respondents said it took them 5 or 10 minutes to understand the instructions, with one saying it took up to 20 minutes. A minority of respondents (3) said it took 40 or 45 minutes to complete all annotations, and others said it took around 120 and up to 180 minutes. The reported average time to complete all annotations ranged from 2 to 9 minutes. Annotators said the quickest tasks were confirming good extraction and classifying the outcome, and the slowest task was rating the salience of the conclusion based on the arguments. In general in free text response, annotators felt that rating salience was very subjective, or that they did not have the qualified legal expertise to feel confident in their ratings.

4 Analysis

4.1 Human-Annotated Dataset

After human annotation we had a dataset of 240 rows. We ran python scripts to clean, standardize, and aggregate our data from the individual csv files given to each annotator, and after that process we had 238 annotated rows. There are some clear trends evident from the pilot dataset. First, the

Figure 1: Reported times in minutes for each respondent to complete all annotations

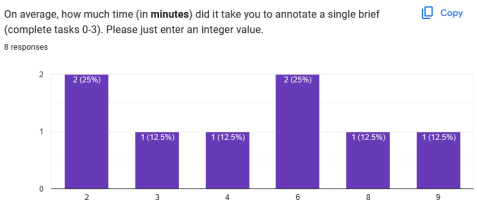

Figure 2: Reported average time in minutes for each respondent to finish annotations for a single row

majority of conclusion types (the requested relief from the court that should logically follow from the brief's argument) were "reverse" or "affirm". This makes sense, as a typical case will have two parties, and normally when one is asking the Court to reverse the lower court's ruling the opposing party is seeking to affirm it.

The arguments asking the Court to "grant" something, or classified as "other" could serve as interesting variants on argumentation to include in the distribution. For example, some parties asked the Court to enjoin (legal jargon for "order something to stop") a prohibition on church gatherings during the pandemic. Annotators likely used "other" when they could not easily classify the argument. This may indicate that the arguments would serve as a good challenge for a classification model. Salience ratings skewed heavily toward 4 or 5. This indicates either high salience and quality of the extracted arguments, or that annotators struggled to rate the arguments and therefore defaulted to high ratings (or both). Qualitatively however, low ratings do seem to correlate with poor argumentation. For example, one of the arguments that had a 0 salience rating was structured simply as follows:

> II. Summary of Argument
>
> a) ARGUMENT AND AUTHORITIES
>
> b) The district court made a plain mathematical error in determining the defendantâ€™s drug quantity

This indicates to us that annotators were paying attention to the task even though some found it difficult, and rated poor arguments appropriately.

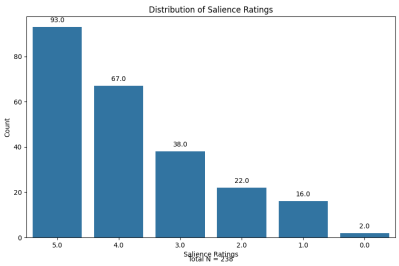

Figure 3: Distribution of salience ratings of whether the conclusion follows from the arguments

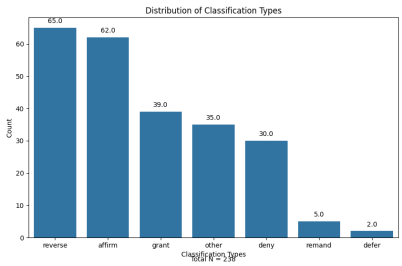

Figure 4: Distribution of classes for argument conclusions

The average word length of the extracted argument was approximately 643 words. There was however a long tail of argument word lengths into the thousands of words, as shown in the figure. Looking at the examples in Appendix A, we believe that longer arguments tend to include more argument steps and greater granularity, which would better capture the chain of thought reasoning.

4.2 Model-in-the-Loop

While the open-source models showed generally good performance on the extraction task, they had some deficiencies that made it difficult to rely solely on an automated process to extract the

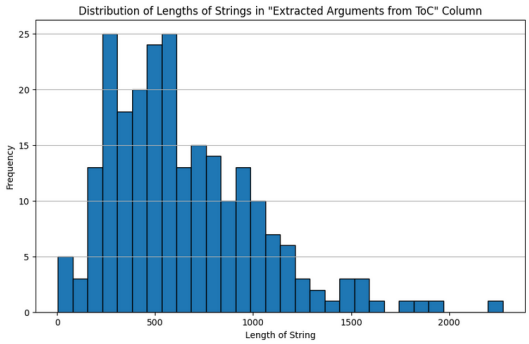

Figure 5: Distribution of the number of words in the extracted arguments (not the ToC as a whole)

arguments. For example, consider the argument from one of the briefs about athlete pay in the NCAA in Appendix A.6. This argument has three main parts and multiple sub-parts that are critical to understanding the chain of thought reasoning being put forth in the brief. However, in this case the arguments extracted by Mistral were missing the sub-arguments, a critical oversight. It is not only sub-arguments that were missing, other times the models would leave out individual top-level arguments, breaking the chain reasoning.

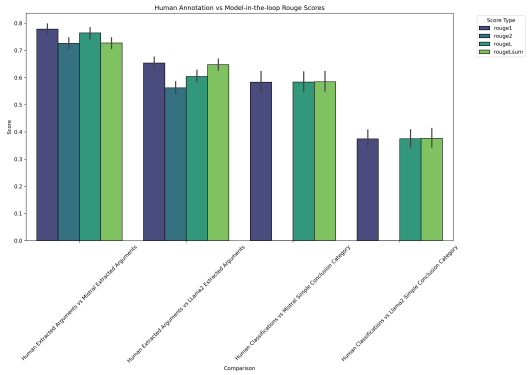

Figure 6: Rouge2 is empty for conclusions because human annotators had a single word classification for the conclusion, while the unfiltered zero-shot generations of the models tended to be more verbose.

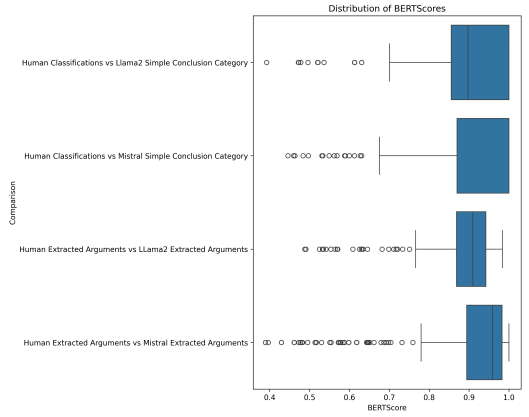

Figure 7: BERT scores generated with LegalBERT, which is finetuned on legal corpora

To better assess the models we calculated ROUGE and BERT scores for the extracted arguments of both models and compared them against our pilot dataset as a reference. ROUGE is a measure of n-gram overlap, and ROUGE-1 measures one-gram overlap between a generated/extracted text

and the reference. A ROUGE-1 score of 1 indicates perfect unigram overlap between the model extraction and reference. BERT score compares BERT embeddings, which is said to better capture semantic similarity between generated and reference text. For argument extraction, the F1 ROUGE-1 scores for the entire dataset were for .7644 for Mistral, and .6300 for Llama 2. The average BERT scores were .8916 and .8778, respectively. For conclusion classification, the ROUGE scores were .5723 for Mistral and .3686 for Llama 2.

Figures 6 and 7 display the distribution of scores for both models across the pilot dataset. Both perform well, with Mistral the clear winner, but Mistral also has a long tail of low scores. This level of performance would make it difficult to rely solely on a model, which means the process would be hard to scale.

5 Limitations

While sourcing only from SCOTUS cases has certain benefits as discussed above, it also has some limitations that are worth considering. SCOTUS cases represent only a small subset of the overall cases that are litigated in the United States. They also tend to skew toward certain topics in the law, such as administrative or criminal law. Further, some areas of law or legal reasoning will not be captured by litigation, e.g. transactional law. There are existing cultural and demographic biases in the legal field that may be reflected in the dataset. Lawyers at the highest level skew affluent, white, and male, and their reasoning may reflect the biases of these demographic classes.

In addition, because we are limiting our labels of the conclusions to a limited set of outcomes (affirm, reverse, remand, affirm-in-part/reverse-in-part), we will lose some nuance of the legal reasoning. However we think labeling like this will improve our annotation process significantly enough to make it worth it. Requiring more sophisticated extraction or labeling from the document would be beyond the capabilities of annotators without legal training. As discussed, we also encountered limitations with using a model in the loop to scale data extraction. The models we tested were not up to the task with a zero-shot approach. Future work may involve using the human-annotated examples to guide a model with a few-shot approach. Automated metrics from the model-extracted arguments are discussed in the Analysis section above.

In addition, the pile-of-law dataset had significant issues with data quality and cleanliness. There were many duplicated rows, unusual special characters, or incomplete inputs. We attempted to filter out writs of certiorari, which are often shorter or may be submitted by pro se litigants (individuals who are not represented by an attorney), but these data quality issues prevented us from filtering effectively. Internally we discussed a web scraping approach that extracts the data directly from the SCOTUS blog, combined with computer vision or some other method to read the pdf representations of the briefs. However we did not have sufficient time to implement this method.

Even after significant filtering data itself was sometimes lacking in terms of how well it represented explicit legal reasoning, especially for non-merit cases. An example of the arguments from a motion for the Court to deny a writ of certiorari (asking the Court not to hear a case on the merits) shows how these types of briefs can be quite limited (see Appendix A.2)

Some Amicus Curiae briefs could also be quite shallow in terms of argumentation, such as the following brief on California's COVID-19 measure to close public meeting spaces including churches (see Appendix A.3.1). However, others are more fully fleshed out, with multiple arguments and sub-arguments that represent structured legal reasoning (see Appendix A.3.2)

This further supports the idea of focusing primarily on merit cases, those that are actually argued in front of the Supreme Court. If we were to extend this work to a complete dataset, we might consider scraping or sourcing higher quality briefs from other sources. As the Analysis section showed, the average argument length is reasonably long but there are a number of very short arguments, which may indicate poor quality.

As discussed, in a post-annotation survey, some annotators also felt that the task to rate the salience of the arguments was too subjective, or they did not feel confident in their ability to understand legal terms well enough to make that determination. More work is probably necessary to furnish annotators with examples or to structure the task in such a way that they feel confident making judgements about arguments.

6 Impact

As discussed above, this dataset will be novel in the legal NLP and chain-of-thought reasoning literature. Our hope is that this dataset will allow researchers to train models with better performance on reasoning tasks specifically and text generation more broadly, as well as improve post-hoc explainability. Given the recent explosion in the use and availability of LLMs, there has been a corresponding increase in the desire to generate structured and more formulaic output — which directly relates to one of the challenges we ran into when investigating the model-in-the-loop workflow. Some examples of promising projects include jsonformer[2], LangChain output parsers [3], Llama.cpp grammar builder [4], and guardrails-ai [1]. Based on our investigation on Llama2 and Mistral's performance on classification and extraction, there is substantial reason to believe that refining their outputs to a more standardized or structured format would result in outsize improvements in performance, possibly replacing (or augmenting) human annotator capabilities. Furthermore, developing a more mature data processing life-cycle (using Airflow or something similar) rather than our one-off script-based approach could allow a future extension of this effort to grow the dataset by orders of magnitude. Though there is a large amount of meaningful legal data available, parsing and labeling remain two of the main obstacles. There is relatively little work on legal NLP in general and even less so on legal reasoning, in part because there are few good datasets. This work could open up entirely new lines of research in the fields of legal NLP and human reasoning.

References

[1] URL https://github.com/guardrails-ai/guardrails.

[2] URL https://github.com/1rgs/jsonformer.

[3] URL https://python.langchain.com/docs/modules/model_io/output_parsers/.

[4] URL https://grammar.intrinsiclabs.ai/.

[5] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras. Lexglue: A benchmark dataset for legal language understanding in english, 2022. URL https://arxiv.org/abs/2110.00976.

[6] P. Henderson, M. S. Krass, L. Zheng, N. Guha, C. D. Manning, D. Jurafsky, and D. E. Ho. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset, 2022. URL https://openreview.net/forum?id=3HCT3xfNm9r.

[7] S. Kim, S. J. Joo, D. Kim, J. Jang, S. Ye, J. Shin, and M. Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning, 2023. URL https://arxiv.org/abs/2305.14045.

[8] Z. Shen, K. Lo, L. Yu, N. Dahlberg, M. Schlanger, and D. Downey. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. 2022. URL http://arxiv.org/abs/2206.10883. Accessed: Oct. 20, 2023.

[9] J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models. 2023. URL http://arxiv.org/abs/2201.11903. Accessed: Dec. 04, 2023.

[10] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. *CoRR*, abs/2104.08671, 2021. URL https://arxiv.org/abs/2104.08671.

[11] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun. Jec-qa: A legal-domain question answering dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020. doi: 10.1609/aaai.v34i05.6519.

[12] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun. How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.466.

[13] Y. Zhong and D. Litman. Computing and exploiting document structure to improve unsupervised extractive summarization of legal case decisions. 2022. doi: 10.48550/arXiv.2211.03229. URL https://arxiv.org/abs/2211.03229.

A First Appendix Title

A.1 Example from a merit case argument:

I. This Case Presents a Live Case or Controversy That Satisfies the Requirements of Article III

A. The Memorandum Threatens Concrete and Imminent Harm to Government Appelleesâ€™ Representation and Federal Funding

B. The Memorandumâ€™s Harms to the Census Count Provided the District Court with Jurisdiction and Fall Within the Evading-Review Exception to Mootness

II. Appellantsâ€™ Reliance on Immigration Status Alone to Subtract Residents from the Apportionment Base Violates Both the Constitution and the Census Act

A. The Constitutionâ€™s Inclusion of All â€œPersons in Each Stateâ€ in the Apportionment Base Encompasses Undocumented Immigrants Who Reside in a State

B. The Census Act Independently Prohibits Appellants from Excluding Usual Residents from the Apportionment Base Due Solely to Their Immigration Status

C. Appellantsâ€™ Arguments in Support of the Memorandum Are Meritless

III. The Memorandum Violates the Constitutional and Statutory Requirements to Base Apportionment Solely on the Censusâ€™s Enumeration

A.2 Example from a writ of certiorari:

I. THIS CASE DOES NOT INVOLVE A FEDERAL CONSTITUTIONAL OR STATUTORY RIGHT AND NONE WAS RAISED IN THE COURTS BELOW

II. ACT 77 COMPLIES WITH THE PENNSYLVANIA CONSTITUTION

A.3 Examples from an amicus curiae brief:

A.3.1 Short

I. States Have Relied On The CDC Order When Crafting Their Own Responses To COVID-19

II. Sudden Vacatur Of The CDC Order Threatens Irreparable Harm To The States

A.3.2 Long

I. Mass Evictions Are Likely Nationwide Without the CDC Order

A. The COVID-19 Pandemic Significantly Increased Housing Insecurity

B. Without Legal Protections from Eviction, Filing Rates Increase

II. Eviction Moratoriums Slow the Spread of COVID-19 and Prevent Negative Short- and Long-Term Health Outcomes

A. Evictions Spread COVID-19, Thwarting Efforts to Contain the Virus

B. Eviction Increases the Rate of COVID-19 Among High-Risk Populations, Leading to Long-Term Complications or Death

C. Studies Suggest Eviction Moratoriums Prevent COVID-19 Deaths

D. Evictions Are Concentrated in Neighborhoods with the Lowest Vaccination Rates And Slowest Economic Recoveries

III. Eviction and COVID-19 Disproportionately Harm Marginalized Groups

A. Evictions Disparately Affect Groups Based on Race, Gender, and Sexual Orientation

B. COVID-19 Has Killed Black, Indigenous, and Latinx People at Higher Rates

A.4 Examples of a conclusion

A.5 Affirm

CONCLUSION

The District Court did not abuse its discretion. Far from it, it did what was necessary both to protect First Amendment rights and the health and safety of Ohioans. Appellees respectfully request that the decision below be AFFIRMED.

Respectfully submitted,

/s/ Jeffrey T. Green

A.6 Deny

CONCLUSION

Given the above facts and authorities, this Court should enforce the appellate waiver and dismiss the appeal. Should the Court deny this motion, the government requests an extension of time of 30 days from the denial to respond to Fuentesâ€™s brief.

Respectfully submitted,

A.7 Other

CONCLUSION

The Court should stay the injunction below and remedial proceedings pending appeal. Alternatively, the Court should stay the injunction and remedial proceedings pending its forthcoming Rucho and Benisek decisions, treat this stay application as a jurisdictional statement, and vacate and remand the opinion and injunction below once those decisions are issued for further consideration.

A.8 Full example

A.8.1 Raw extracted ToC

TABLE OF CONTENTS

CORPORATE DISCLOSURE STATEMENT i

TABLE OF AUTHORITIES vi

STATEMENT REGARDING ORAL ARGUMENT xv

STATEMENT OF THE ISSUES..... 1

STANDARD OF REVIEW 1

INTRODUCTION 2

STATEMENT OF THE CASE 3

A. THE COVID-19 CRISIS 4

1. The March 9, 2020 Ohio State of Emergency and Subsequent Events Essentially Close Ohio 4

2. The March 22, 2020 Shutdown Order Deepens Ohioâ€™s Restrictions.... 5

3. The April 30, 2020 Shutdown Order Extends Ohioâ€™s Sweeping Restrictions 8

B. THE DISTRICT COURT ENJOINS OHIOâ€™S BALLOT-ACCESS RESTRICTIONS 9

C. THE MOTIONS PANEL STAYS THE DISTRICT COURTâ€™S PRELIMINARY INJUNCTION 11

D. CONDITIONS IN OHIO WORSEN AS THE STATE ATTEMPTS REOPENING 12

E. THOMPSON SEEKS TO LIFT STAY 15

SUMMARY OF THE ARGUMENT 16

ARGUMENT 17

I. LAWS REGULATING BALLOT ACCESS FOR INITIATIVES IMPLICATE THE FIRST AMENDMENT IN THE SIXTH CIRCUIT AND IN COURTS ACROSS THE COUNTRY 17

Case: 20-3526 Document: 94 Filed: 08/26/2020 Page: 5 71a

A. Ballot Initiatives Implicate Core Political Speech and Strict Scrutiny Under Meyer-Buckley 17

B. The Sixth Circuit Has Consistently and Repeatedly Held that the First Amendment Applies to Ballot Initiatives 19

C. The Sixth Circuit is on the Correct Side of an Emerging Circuit Split 20

II. UNDER ANDERSON-BURDICK THE IN-PERSON COLLECTION LAWS ARE UNCONSTITUTIONAL AS APPLIED 22

A. During a Pandemic, the Combined Effect of Strict Enforcement of In- Person Collection Laws and Ohioâ€™s Shutdown Orders Impose a Severe Burden on Circulators Warranting Strict Scrutiny 22

1. A Total Exclusion Litmus Test Was Not the Law Before the Pandemic 22
2. A Total Exclusion Litmus Test Is Not the Law Now 25
3. The First Amendment is Implicated Equally by Restrictions on Candidatesâ€™ and Initiativesâ€™ Circulators 29
4. Ohioâ€™s Vague First Amendment Exception Did Not Lessen the Severe Burden on Thompsonâ€™s Circulation Efforts 30
5. Thompson Has Been and Continues to Be Severely Burdened 33
6. Thompson was Effectively Denied Access to the Ballot and Meets Any Total Exclusion Litmus Test 38
B. Ohioâ€™s In-Person Collection Laws Do Not Survive Strict Scrutiny ... 40
1. Ohio Cannot Demonstrate That Strict Enforcement of the In- Person Collection Laws During the Pandemic Is Necessary to Further Any Compelling State Interest 40
C. Ohioâ€™s In-Person Collection Laws Do Not Even Survive Intermediate Scrutiny..... 45
Case: 20-3526 Document: 94 Filed: 08/26/2020 Page: 6 72a
III. INJUNCTIVE RELIEF IS REQUIRED TO REMEDY THOMPSONâ€™S INJURY 46
A. The Preliminary Injunction Factors Weigh in Favor of Granting Thompson Equitable Relief 46
B. The District Court Did Not Abuse Its Discretion By Awarding Thompson A Negative Preliminary Injunction 48
C. Thompson Is Also Entitled to Affirmative Relief 50
CONCLUSION 51

A.8.2 Task 0: Did extraction work properly?

True

A.8.3 Task 1: Extract the argument

I. LAWS REGULATING BALLOT ACCESS FOR INITIATIVES IMPLICATE THE FIRST AMENDMENT IN THE SIXTH CIRCUIT AND IN COURTS ACROSS THE COUNTRY
A. Ballot Initiatives Implicate Core Political Speech and Strict Scrutiny Under Meyer-Buckley
B. The Sixth Circuit Has Consistently and Repeatedly Held that the First Amendment Applies to Ballot Initiatives
C. The Sixth Circuit is on the Correct Side of an Emerging Circuit Split
II. UNDER ANDERSON-BURDICK THE IN-PERSON COLLECTION LAWS ARE UNCONSTITUTIONAL AS APPLIED
A. During a Pandemic, the Combined Effect of Strict Enforcement of In-Person Collection Laws and Ohioâ€™s Shutdown Orders Impose a Severe Burden on Circulators Warranting Strict Scrutiny
1. A Total Exclusion Litmus Test Was Not the Law Before the Pandemic
2. A Total Exclusion Litmus Test Is Not the Law Now
3. The First Amendment is Implicated Equally by Restrictions on Candidatesâ€™ and Initiativesâ€™ Circulators
4. Ohioâ€™s Vague First Amendment Exception Did Not Lessen the Severe Burden on Thompsonâ€™s Circulation Efforts
5. Thompson Has Been and Continues to Be Severely Burdened
6. Thompson was Effectively Denied Access to the Ballot and Meets Any Total Exclusion Litmus Test
B. Ohioâ€™s In-Person Collection Laws Do Not Survive Strict Scrutiny
1. Ohio Cannot Demonstrate That Strict Enforcement of the In-Person Collection Laws During the Pandemic Is Necessary to Further Any Compelling State Interest
C. Ohioâ€™s In-Person Collection Laws Do Not Even Survive Intermediate Scrutiny
III. INJUNCTIVE RELIEF IS REQUIRED TO REMEDY THOMPSONâ€™S INJURY A. The Preliminary Injunction Factors Weigh in Favor of Granting Thompson Equitable Relief

B. The District Court Did Not Abuse Its Discretion By Awarding Thompson A Negative Preliminary Injunction

C. Thompson Is Also Entitled to Affirmative Relief

A.8.4 Extracted Conclusion

CONCLUSION The District Court did not abuse its discretion. Far from it, it did what was necessary both to protect First Amendment rights and the health and safety of Ohioans. Appellees respectfully request that the decision below be AFFIRMED.

Respectfully submitted,

/s/ Jeffrey T. Green

A.8.5 Task 2: Classify the requested outcome in the conclusion

Affirm

A.8.6 Task 3: Rate the salience

3

A.9 Example of model-extracted argument vs. human-extracted

A.9.1 Extracted by a human annotator:

I. The Revered Tradition Of Amateurism Is Essential To College Sports And Cherished By Student-Athletes. 7

A. Student-athletes benefit from amateurism in intercollegiate athletics. 9

B. History confirms the necessity of the NCAAâ€™s careful regulation to preserve amateurism in college sports. 13

C. The NCAAâ€™s compensation caps are essential to the tradition of amateurism. 16

II. The Lower Courtsâ€™ Rulings Endanger The Tradition Of Amateurism By Inviting A Compensation Arms Race. 19

III. A Compensation Arms Race Undermines The Tradition Of Amateurism By Degrading The Athletic And Academic Experiences Of Most Student-Athletes. 23

A. Many college athletics programs risk being defunded or cut. 24

B. Student-athletesâ€™ educational experiences will also suffer. 31

A.9.2 Extracted by Mistral:

1. The Revered Tradition Of Amateurism Is Essential To College Sports And Cherished By Student-Athletes.

2. The Lower Courtsâ€™ Rulings Endanger The Tradition Of Amateurism By Inviting A Compensation Arms Race.

3. A Compensation Arms Race Undermines The Tradition Of Amateurism By Degrading The Athletic And Academic Experiences Of Most Student-Athletes.

A.10 Example of how briefs appear in PDF format as filed with the Supreme Court

A.10.1 An example of a table of contents containing the argument

TABLE OF CONTENTS

Page

Introduction 1

Jurisdiction 2

Statement 2

A. Constitutional and Statutory Background 2

B. Factual Background 5

C. Procedural History 7

Summary of Argument 10

Argument 13

I. This Case Presents a Live Case or Controversy That Satisfies the Requirements of Article III. 13

A. The Memorandum Threatens Concrete and Imminent Harm to Government Appellees' Representation and Federal Funding 13

B. The Memorandum's Harms to the Census Count Provided the District Court with Jurisdiction and Fall Within the Evading-Review Exception to Mootness. 19

II. Appellants' Reliance on Immigration Status Alone to Subtract Residents from the Apportionment Base Violates Both the Constitution and the Census Act 23

Page

A. The Constitution's Inclusion of All "Persons in Each State" in the Apportionment Base Encompasses Undocumented Immigrants Who Reside in a State.....25

B. The Census Act Independently Prohibits Appellants from Excluding Usual Residents from the Apportionment Base Due Solely to Their Immigration Status.....32

C. Appellants' Arguments in Support of the Memorandum Are Meritless.....37

III. The Memorandum Violates the Constitutional and Statutory Requirements to Base Apportionment Solely on the Census's Enumeration. 44

Conclusion.....50

A.10.2 An example of the conclusion from the same brief

describes its policy as adopting apportionment figures "[f]ollowing the 2020 Census" that will necessarily be different from the total-population figures produced by the census itself (App.6a, 8a), it improperly deviates from the process prescribed by both the Constitution and the Census Act.

CONCLUSION

The Court should affirm the final judgment below.

Respectfully submitted,

LETITIA JAMES
Attorney General
State of New York
BARBARA D. UNDERWOOD*
Solicitor General
STEVEN C. WU
Deputy Solicitor General
JUDITH N. VALE
Senior Assistant Solicitor General
ERIC R. HAREN
Special Counsel
barbara.underwood@ag.ny.gov

MATTHEW COLANGELO
Chief Counsel for Federal Initiatives
ELENA GOLDSTEIN
Deputy Chief Civil Rights Bureau
FIONA J. KAYE
Assistant Attorney General

* *Counsel of Record*

November 2020

(*Counsel list continues on next page.*)

B Second Appendix Title

Hello, and thank you for annotating for us!

Here are your instructions for annotating. Please follow them exactly, and precisely.

1. Please go to [drive/data link]

2. Here, you should see a csv file in the format of [your-UNI].csv

3. Right-click on the file, and select "Open with" » "Google Sheets". This should open a Google sheet representation of the CSV file.

- Do NOT modify any cells other than the ones you are explicitly instructed to annotate

4. Modify the column spacing to make the documents more readable. For example, you should drag and expand the column width for columns B (table of contents) and D.

5. Confirm that you see your UNI as the document title, as well as in the last column 'annotator-uni'.

For each row (2 - 21), do the following tasks:

(0) Confirm that the ToC and Conclusion have been extracted properly

- Write exactly one of [True, False]

(1) Extract the arguments and paste them in this form

- Leave out portions of the table of contents that do not relate to the argument, such as Table of Authorities.

- Also leave out excess periods, formatting markers, and numbers indicating the page.

- Maintain the markers that indicate the structure of arguments and sub-arguments, e.g. Roman numerals and capital letters.

- Maintain a line break between each argument.

(2) Classify the outcome of the brief. We will provide some examples for all of you.

- Write exactly one of [Affirm, Reverse, Remand, Grant, Deny, Other]. Use 'Other' if you are not sure, or if there is substantial ambiguity.

- In general, the outcome that each brief is arguing for should be explicitly stated in the conclusion. So start by looking for one of the terms within the text of the conclusion.

- Affirm means to agree with the lower court, and maintain that court's decision.

- Reverse means to disagree with the lower court, and to change the outcome of their decision in some way.

- Remand means to send the decision back down to the lower court with further instructions, often to flesh out some issue of importance.

- Grant means to give the requesting party what they are asking for in the law.

- Deny means the opposite, to deny the requesting party the thing they are asking for.

- See the example below

(3) Rate the salience of the reasoning behind the conclusion based on the table of contents. In other words, is the logic of the outcome clear from solely the table of contents?

- [Detailed Instruction] Use a scale of 1-5, where 1 is Not Clear At All, and 5 is Extremely Clear.

- The following is an example of the tasks and the expected outcomes:

> Table of Contents i
>
> Table of Authorities ii
>
> Interests of Amici Curiae 1
>
> Summary of the Argument 3
>
> Argument 5
>
> I. Plaintiff obtained insurance based on fraud or misstatement of a material fact.....5
>
> A. When Plaintiff applied for insurance for his car, his statement written to the insurance agent representing Defendant stated that he had never had his insurance canceled or refused.....5

B. Prior to this, another insurance agent for the Defendant, Mr. Lilly, had refused to write an insurance policy for the Plaintiff because Plaintiff had previously had an insurance policy canceled.....7

II. Plaintiff's prior insurance history was a material fact because it would have impacted Plaintiff's insurance rate or even his ability to obtain insurance..10

III. Plaintiff fails to state a claim of waiver or estoppel.....15

A. Defendant timely canceled Plaintiff's insurance upon learning that he had submitted false statement..20

B. Timely cancellation of the policy negates any claim of waiver or estoppel.....26

IV. In the absence of pleadings or proof of waiver or estoppel, a fraudulent or false statement of a material fact by an insurance applicant on their application is sufficient to defeat recovery of the applicant's insurance claim...29

Conclusion 35

CONCLUSION:

Plaintiff's false statement that he had not previously had his insurance canceled or refused was sufficient to defeat his claim, and Defendant's motion for a directed verdict should have been granted on these grounds. The lower court should be reversed.

Task (0) Confirm that the ToC and Conclusion have been extracted properly

Response: True

Task (1) Extract the arguments and paste them in this form Response:

I. Plaintiff obtained insurance based on fraud or misstatement of a material fact.

A. When Plaintiff applied for insurance for his car, his statement written to the insurance agent representing Defendant stated that he had never had his insurance canceled or refused.

B. Prior to this, another insurance agent for the Defendant, Mr. Lilly had refused to write an insurance policy for the Plaintiff because Plaintiff had previously had an insurance policy canceled.

II. Plaintiff's prior insurance history was a material fact because it would have impacted Plaintiff's insurance rate or even his ability to obtain insurance.

III.Defendant timely canceled Plaintiff's insurance upon learning that he had submitted false statements

A. Timely cancellation of the policy negates any claim of waiver or estoppel.

IV. In the absence of pleadings or proof of waiver or estoppel, a fraudulent or false statement of a material fact by an insurance applicant on their application is sufficient to defeat recovery of the applicant's insurance claim.

Task (2) Classify the outcome of the brief

Response: Reverse

Task (3) Rate the salience of the reasoning behind the conclusion based on the table of contents

Response: 5

Researcher's note: In this example, there is a clear link in each step of the reasoning, and the conclusion follows directly from arguments in the ToC. Therefore we would rate the salience as a 5, for Extremely High.

Another Example:

TABLE OF CONTENTS

TABLE OF AUTHORITIES ii

STATEMENT OF INTEREST OF AMICUS CURIAE 1

SUMMARY OF ARGUMENT 1

ARGUMENT 4

I. This Case is Like Toolson v. New York Yankees, Inc., 346 U.S. 356 (1953) 4

A. The Baseball Cases: Federal Baseball and Toolson 4

B. The College Sports Cases: Board of Regents and Alston 6

II. Toolson Has Had Disastrous Consequences for Hundreds of Thousands of American Workers ... 8

III. The Court Need Not Reprise Toolson 11

CONCLUSION 12

CONCLUSION:

The Court should learn from its mistake in Toolson and deny Petitioners their requested relief.

Respectfully submitted,

Task (0) Confirm that the ToC and Conclusion have been extracted properly

Response: True

Task (1) Extract the arguments and paste them in this form Response:

I. This Case is Like Toolson v. New York Yankees, Inc., 346 U.S. 356 (1953)

A. The Baseball Cases: Federal Baseball and Toolson

B. The College Sports Cases: Board of Regents and Alston

II. Toolson Has Had Disastrous Consequences for Hundreds of Thousands of American Workers

III. The Court Need Not Reprise Toolson

Task (2) Classify the outcome of the brief

Response: Deny

Task (3) Rate the salience of the reasoning behind the conclusion based on the table of contents

Response: 2

Researcher's Note: Here, while the reasoning is laid out in clear steps and the conclusion does follow somewhat from these steps, there is little detail and the reasoning steps are rather disconnected. Therefore we would rate this as a 2, for low salience.

Thanks again, and please reach out to Jesse or Nikhil with any issues or questions!