Extractive and Abstractive Text Summarization of US Supreme Court Opinions

Karthik Narasimhan
Drexel University
kn568@drexel.edu

Abstract

Abstractive text summarization of legal case judgments remains a challenging problem within Legal NLP, particularly due to the long length of case judgments, especially those from recent years. This is an important problem to solve because abstractive summaries of case judgments can help laypeople without legal training understand the reasoning of the courts issuing the judgments.

While recently developed language models for conditional generation like Longformer can be trained on documents of up to 16,784 tokens and the human-written summaries of those documents, the extremely limited availability of human-written summaries of United States Supreme Court (SCOTUS) opinions in easily accessible databases severely limits the effectiveness of Longformer and similar language models for generating abstractive summaries of US Supreme Court opinions. I propose a dual transfer learning approach to solving this problem which entails generating an extractive summary of 816 SCOTUS opinions and their respective syllabus, the official SCOTUS-issued summary of the opinion, with LEGAL-BERT, a language model pre-trained on thousands of case judgments. Next, a BART model pre-trained on BillSum, a dataset of thousands of congressional bills and their summaries, is fine-tuned on 80% of the extractive-summary SCOTUS opinion/syllabi dataset before generating predictions of the syllabi from the remaining 20% of the dataset. A small sample of its generated predictions are evaluated for their legal accuracy by a trained legal practitioner. Evaluation shows the generated predictions' legal accuracy is not good enough for deployment of the model. I conclude that is largely due to the limited amount of training data.

1 Introduction

Automatic text summarization of a long document is an important but rather challenging problem within the subfield of machine learning research commonly referred to as natural language processing. There are two main approaches for automatic text summarization: extractive and abstractive. Extractive summarization methods identify the most important sections of a text based on a scoring mechanism and extract them to form a summary, hence the term "extractive summarization." Abstractive summarization methods learn the meaning of the text as a whole and generates a new summary that conveys the most important information from the original text, the phrasing of which may not be in the original text. Tsonkov et al., 2021 While abstractive summarization is harder than extractive summarization as it requires real-word knowledge and semantic class analysis, abstractive summaries are much closer approximations of human-written summaries than extractive summaries, and are therefore, more meaningful. Suleiman and Awajan, 2020 Neural language models for extractive text summarization do not need to be trained on human-written summaries of the main documents in conjunction with the main documents while neural language models for abstractive text summarization do need to be trained on human-written summaries of the main documents in conjunction with the main documents.

Every US Supreme Court opinion is comprised of the main opinion, a syllabus, and any

concurrences or dissents. The syllabus appears first before the main opinion and is a summary of the facts of the case and the Court's holding (which includes its decision and the reasoning for its decision) that is added by the Court to help the reader better understand the opinion. It is not part of the official opinion. American Bar Association, 2022 The US Supreme Court has stated that the "headnotes to the opinions of this Court are not the work of the Court, but are simply the work of the Reporter, giving his understanding of the decision, prepared for the convenience of the profession." *United States v. Detroit Lumber Co.* 1906 The "Reporter" that the US Supreme Court is referring to is the Reporter of Decisions of the Supreme Court of the United States, who is, by federal statute, directly appointed by the US Supreme Court.

The US Supreme Court is the highest judicial tribunal in the United States for all cases and controversies arising under the US Constitution or laws of the United States. Its long-standing power of "judicial review" refers to its authority to invalidate legislation or executive actions which, in the Court's considered judgment, conflict with the US Constitution, the supreme law of the United States. Supreme Court of the United States, n.d. While each of the US Supreme Court's opinions varies greatly in the number of Americans that it impacts, and the degree to which it impacts them, its power of "judicial review" means that its opinions will always have a major impact on the lives of the American people in the aggregate, if not on an individual basis. This power in conjunction with its unaccountability to the electorate means that it is essential for the Court to maintain its legitimacy in the eyes of the public for our constitutional system to remain intact.

If the public sees the Court's opinions as legitimate, it puts pressure on elected officials to follow the Court's decisions even when they disagree. Keith, 2022 This type of legitimacy is referred to as "sociological legitimacy" by Professor Richard Fallon in his book Law and Legitimacy in the Supreme Court, in contrast to legal legitimacy and moral legitimacy. Sociological legitimacy depends on the public viewing the legal system as worthy of respect and obedience, while moral legitimacy focuses on whether people should treat the legal system as worthy of respect and obedience and legal legitimacy depends on the Court's Justices using interpretive methods that are broadly accepted within the legal culture. Grove, 2019

For most of the 20th century, the US Supreme Court applied the reasoning in its opinions to achieve legal, sociological, and moral legitimacy for its decisions. During this time, the Court maintained its sociological legitimacy through "appearance management" – it either used the reasoning of its opinions as a means of winning acceptance for a decision or softening the opposition from those who found fault with the substantive outcome. Achieving sociological legitimacy was important enough to the Court that it sometimes sacrificed legal legitimacy for the sake of sociological legitimacy. Wells, 2007

In recent years, while the Supreme Court's per year caseload decreased from earlier years, opinions have gotten lengthier, especially on cases related to controversial subject matters as the Court has chosen to prioritize legal legitimacy over sociological legitimacy. Truscott and Feldman, 2022 And while recent opinions are longer and more strongly worded than those from earlier years, they rarely add clarity to the underlying decision. Penrose, 2019 Most people do not read Supreme Court opinions Wells, 2007, and instead get their knowledge of a Supreme Court decision from the news. When members of the public learn about a Supreme Court decision that they disagree with, their trust in the Supreme Court declines. This is partly why public approval of the US Supreme Court has fallen sharply in the last two years, falling from 53% approval in Sept 2020 to 40% approval in September 2022, with its disapproval rating rising from 43% in Sept 2020 to 58% in Sep 2022. Jones, 2022

There exists now an opportunity for neural language models to play an important role in helping the current US Supreme Court achieve sociological legitimacy for its decisions. The aim of my research on this subject is to develop a neural language model that can generate abstractive summaries of US Supreme Court opinions specifically for the task of "appearance

management" – summaries that can help win acceptance for the Court's decision or soften the opposition from those who find fault with the substantive outcome. This is a highly complex and challenging problem, but one nonetheless worth solving.

My goal for this specific research project was to train a neural language model on a training corpus consisting of US Supreme Court opinion-syllabi pairs, and then use it to generate its predictions of the syllabi from an unseen test set of US Supreme Court opinions that would be evaluated as legally accurate by a trained legal practitioner. To develop this project, I extracted a list of legal terms from a legal dictionary to build a legal domain-specific vocabulary to add to the language models' tokenizer, built a dataset consisting of 816 US Supreme Court opinion-syllabi pairs in JSON file format, identified and copied case citations in the dataset to add to the language models' tokenizer, generated an extractive-summary of each US Supreme Court opinion-syllabus pair in the dataset using the Legal BERT language model and BERT Extractive Summarizer, fine-tuned a BART model that was pre-trained on the BillSum dataset on a training set of the extractive-summary US Supreme Court opinion-syllabi pairs, and generated predictions of the syllabi of the test set of extractive-summary US Supreme Court opinions.

A trained legal practitioner, Daniel Friedman, a Visiting Assistant Professor at Villanova School of Law, evaluated two of the generated syllabi for legal accuracy and has provided this research with his written evaluations and a general overview of the BART model's legal accuracy. I have included his evaluations and overview in this paper. I have also included Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores, which are commonly used for the evaluation of automatic text summarization, in this paper. I must stress however, for reasons included in the appendix, that the US Supreme Court opinion-syllabi pairs in the dataset used for this research are only a very small portion of the total number of US Supreme Court opinion-syllabi pairs in existence and the ROUGE metrics included for the purposes of evaluation will not be predictive of the results from training a neural language model on the complete set of US Supreme Court opinion-syllabi pairs.

2 Related Work

Shukla et al., 2022 developed three datasets: an Indian-Abstractive dataset, an Indian-Extractive dataset, and a UK-Abstractive dataset, and carried out extensive experiments with several extractive and abstractive summarization methods (both supervised and unsupervised) over them, including a hybrid extractive-abstractive approach referred to as BERT_BART. The generated summaries of the language models were evaluated according to ROUGE-L scores and by trained legal practitioners. Researchers found the Legal-Pegasus model to be the best-performing model according to ROUGE-L scores. Legal practitioners opined that the generated abstractive summaries were unorganized, with many incomplete sentences.

3 Dataset

The size of the total corpus is 816 US Supreme Court opinion-syllabus pairs.

For the purposes of brevity, the term 'older opinions' in this section shall refer to US Supreme Court opinions issued during the late Antebellum era, Civil War era, Reconstruction era, Gilded Age and Progressive era. The term 'recent opinions' shall refer to US Supreme Court opinions issued from the year 2000 to the present.

3.1 Exploratory Data Analysis (EDA)

EDA was conducted on the corpus of US Supreme Court opinions on five measures: Opinion Year vs Opinion Character Count, Opinion Year vs Opinion Word Count, Opinion Year vs Opinion Sentence Count, Opinion Year vs Avg Word Length of Opinion, Opinion Year vs Avg Sentence Length of Opinion. Scatter plots of these five measures have been included as Figures 1-5.

From the plots, we can observe that while there are US Supreme Court opinions during the late 1800's and early 1900's with similar character and word counts as recent opinions, there are no recent opinions that have as few words and characters as some older opinions. And except for one very long opinion during the early 1900's and a long opinion from the late 1800's, some recent opinions have more words and characters

Figure 1: Opinion Year vs Character Count

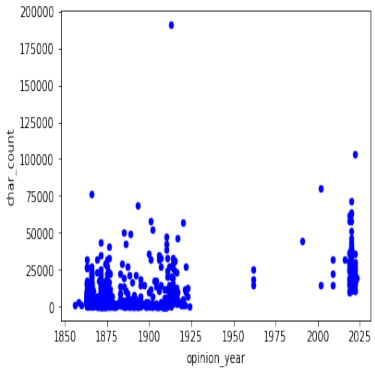


Figure 2: Opinion Year vs Word Count

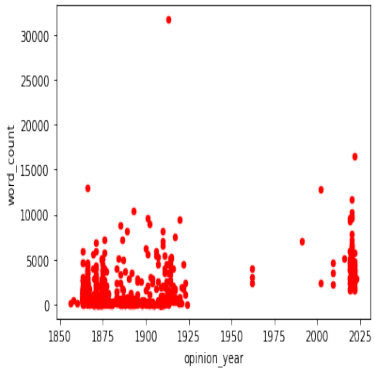


Figure 3: Opinion Year vs Sentence Count

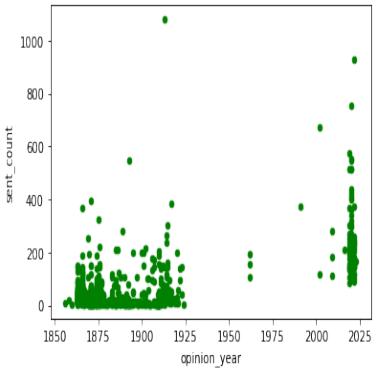


Figure 4: Opinion Year vs Avg Word Length Count

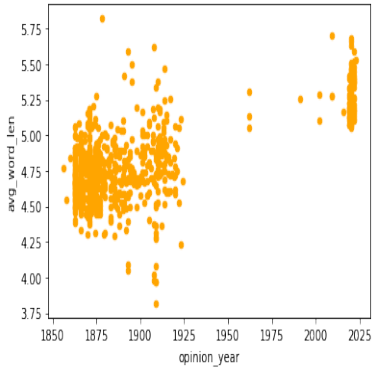
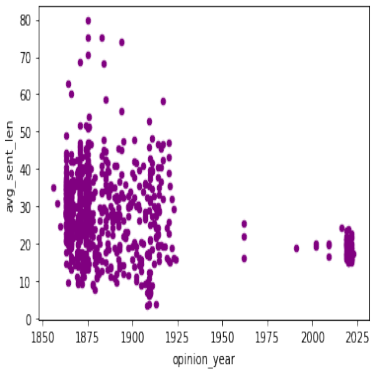


Figure 5: Opinion Year vs Avg Sentence Length Count



than every older opinion. We can also observe that the number of sentences in recent opinions is significantly higher than the number of sentences in older opinions. And while the average length of words in recent opinions is higher than the average length of words in most older opinions, the average length of sentences in recent opinions is lower than the average length of sentences of most of the older opinions.

3.2 Extractive Summarization

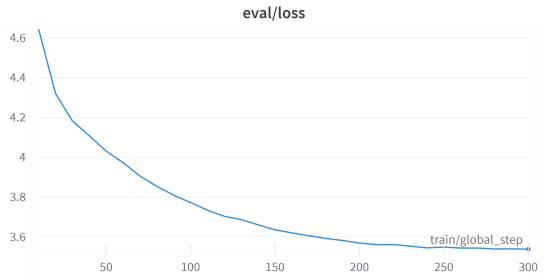
It was necessary to generate extractive summaries of the corpus of US Supreme Court opinion-syllabus pairs before fine-tuning BART on the extractive summaries for abstractive text summarization because the maximum input size of BART is 1024 tokens. With a rule of thumb of 1 word to 1 token, this is significantly shorter than every opinion in the dataset. Assuming that sentences may be as long as 25 words, I decided to set the length of the extractive-summaries of the recent opinions at 25 sentences each and 10 sentences each for their paired syllabus, following a rule of thumb of the syllabus truncated to be one-half of the length of their paired opinion.

For the older opinions, since some of them were shorter than 25 sentences, I could not generate extractive-summary of the older opinions at 25 sentences in length. I decided then to generate extractive-summary of the older opinions and syllabi at 50% of their original length.

The extractive summaries were split into 652 training samples, and 164 testing samples.

Figure 6: Step vs Train Loss

Figure 7: Step vs Evaluation Loss



4 Experimental Settings & Results

I fine-tuned the BART model pre-trained on BillSum on the training samples for 300 steps with a learning rate of 2^{-5} and the maximum input size of BART set to 1024 tokens. Any input over 1024 tokens in length was truncated to 1024 tokens. The maximum output size was set to 512 tokens, and any outputs over 512 tokens in length was truncated to 512 tokens. I used Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. Batch size is set to 8 and gradient accumulation steps is set to 8.

Figure 8: Step vs ROUGE1 of Evaluation

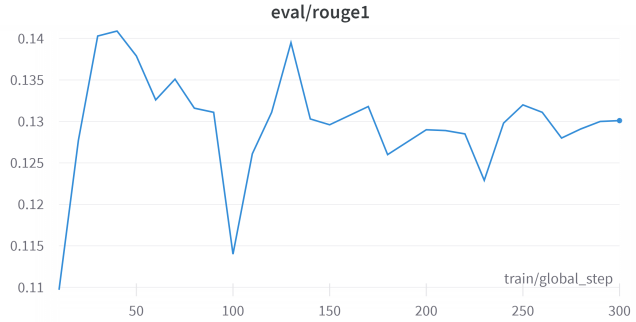


Figure 9: Step vs ROUGE2 of Evaluation

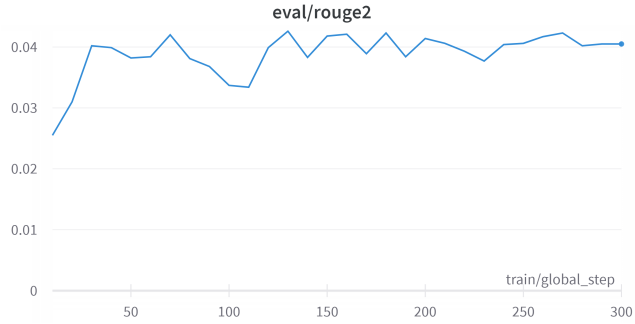
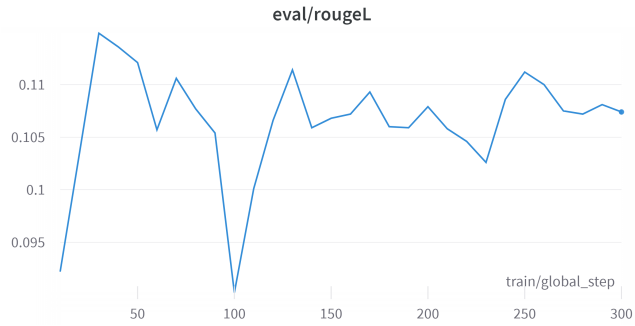


Figure 10: Step vs ROUGEL of Evaluation



5 Evaluation of Legal Accuracy

Daniel Friedman, Visiting Professor of Law at Villanova University, evaluated five of the fine-tuned BART model's predicted syllabi of the US Supreme Court opinions in the test set for their legal accuracy and has provided this written overview of his evaluation:

> Style: The generative text contains some minor spelling and grammatical errors, which sometimes make it difficult to understand the exact meanings of phrases, clauses, or whole sentences. The most significant such errors obscure who is saying what, so that a reader might not know whether a particular view belongs to one of the litigants or to the court itself.
>
> Citation: The text generator has clearly identified that citation to other cases is an important component of legal writing, but it doesn't seem to know how to link particular ideas to the cases it references, so that a case citation will follow a statement that has no actual connection to that opinion's content. Invention: Quotations are rarely accurate and sometimes include phrases that don't appear anywhere in the opinion. Case names are also sometimes invented, sometimes by combining the names of parties from cases that do appear in the opinion.

Misstatements: The law isn't always right, and it seems like it can be hard for the model to understand the difference between things like sources and purposes of law, e.g., whether a statute is authorized by a particular set of rules or is intended to put those rules into practice. It did not always seem capable of clearly stating which actors were required to do what and why. To be fair, most people reading cases also find this difficult.

Focus: Probably the trickiest problem is figuring out what the most important part of an opinion is. This is the thing that law students struggle the most with, too! While the model frequently raised issues that appear in the opinions it was analyzing, it rarely seemed to zero in on the key issues. This is actually quite similar to the way first-year law students discuss cases: they can tell you lots of facts about them and the ideas that they reference, but they have great difficulty reducing them to the handful of most significant points.

6 Conclusion

Although the ROUGE metrics of the BART model's predicted syllabi are very low, and the predicted syllabi contain many spelling, grammatical, and legal errors, the presence of a modicum of legal accuracy means that the corpus of US Supreme Court opinion-syllabus pairs has demonstrated its potential usefulness as a dataset for fine-tuning Transformer language models for the task of abstractive text summarization. While the BART model was not able to accurately summarize US Supreme Court opinions at this time, these early findings have pointed the way forward for further research and development on this important subject at the exciting intersection of artificial intelligence and law.

7 Acknowledgements

The submission of this paper serves as the culmination of the two-course Capstone sequence for the Master of Science in Artificial Intelligence and Machine Learning program at Drexel University. The author thanks Dr. Jeremy Johnson for his guidance during the Capstone course sequence as well as Daniel Friedman for his timely, thorough, and detailed evaluation of the BART model's predicted syllabi for their legal accuracy.

References

United States v. Detroit Lumber Co., 200 U.S. 321 (1906). https://supreme.justia.com/cases/federal/us/200/321/

Wells, M. L. (2007). "sociological legitimacy" in supreme court opinions. *Washington & Lee Law Review*, *64*, 1011–1070.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/ARXIV.1706.03762

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/ARXIV.1810.04805

Grove, T. L. (2019). The supreme court's legitimacy dilemma. *Harvard Law Review*, *132*, 2240–2276.

Kornilova, A., & Eidelman, V. (2019). BillSum: A corpus for automatic summarization of US legislation. *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 48–56. https://doi.org/10.18653/v1/D19-5406

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. https://doi.org/10.48550/ARXIV.1910.13461

Miller, D. (2019). Leveraging bert for extractive text summarization on lectures. https://doi.org/10.48550/ARXIV.1906.04165

Penrose, M. (2019). Overwriting and underdeciding: Addr writing and underdeciding: Addressing the rober essing the roberts cour ts court's shrinking docket. *SMU Law Review Forum*, *72*, 8–19.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-bert: The muppets straight out of law school. https://doi.org/10.48550/ARXIV.2010.02559

Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/9365340

Tai, W., Kung, H. T., Dong, X., Comiter, M., & Kuo, C.-F. (2020). ExBERT: Extending pre-trained models with domain-specific

vocabulary under constrained training resources. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1433–1439. https://doi.org/10.18653/v1/2020.findings-emnlp.129

Yang, Q., Zhang, Y., Dai, W., & Pan, S. J. (2020). *Transfer learning*. Cambridge University Press.

Tsonkov, T., Lazarova, G., Zmiycharov, V., & Koychev, I. (2021). A comparative study of extractive and abstractive approaches for automatic text summarization on scientific texts. *Education and Research in the Information Society*.

American Bar Association. (2022). *How to read a u.s. supreme court opinion*. https://www.americanbar.org/groups/public_education/publications/teaching-legal-docs/how-to-read-a-u-s-supreme-court-opinion/

Jones, J. M. (2022). *Supreme court trust, job approval at historic lows*. https://news.gallup.com/poll/402044/supreme-court-trust-job-approval-historical-lows.aspx

Keith, D. (2022). *A legitimacy crisis of the supreme court's own making*. https://www.brennancenter.org/our-work/analysis-opinion/legitimacy-crisis-supreme-courts-own-making

Shukla, A., Bhattacharya, P., Poddar, S., Mukherjee, R., Ghosh, K., Goyal, P., & Ghosh, S. (2022). Legal case document summarization: Extractive and abstractive methods and their evaluation. https://doi.org/10.48550/ARXIV.2210.07544

Truscott, J. S., & Feldman, A. (2022). *Lengthier opinions and shrinking cohesion: Indications for the future of the supreme court*. https://www.scotusblog.com/2022/07/lengthier-opinions-and-shrinking-cohesion-indications-for-the-future-of-the-supreme-court/

Free Law Project. (2023). Eyecite.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Third Edition draft).

Supreme Court of the United States. (n.d.). *The court and constitutional interpretation*. https://www.supremecourt.gov/about/constitutional.aspx

A SCOTUS Opinion-Syllabi Pair Extraction and Dataset Compilation

I first compiled a dataset of 100 US Supreme Court opinion-syllabi pairs by copying them from the PDF where they were originally published into a txt file, removing any extraneous text like page numbers, and changing the Word-format quotation marks to JSON-readable quotation marks with the required escape characters. They were then copied and pasted directly into the JSON file used as the dataset. All these opinion-syllabi pairs were from 2018 to 2022 as these opinions were the most accessible and properly formatted. I identified each opinion-syllabus pair with its docket number.

Shortly after compiling this dataset, I found the website of the Free Law Project, which hosts a dataset called CourtListener, which hosts every case judgment from the US Supreme Court, along with federal appeals courts and district courts. A large variety of metadata for each case judgment is also included with the full text of the opinion. I contacted the Free Law Project to see if CourtListener contained the syllabus for every opinion from the US Supreme Court, under a sub-field called 'syllabus' for easy retrieval through the CourtListener API. I was later contacted by the lead developer for the Free Law Project, who included a ZIP file contained JSON files for just over 1300 US Supreme Court opinion clusters. An opinion cluster is a JSON object which contains the full text of the Court's opinion and any relevant metadata. The lead developer said in his email that after running a SQL query on the database, he found that only 1300 opinion clusters contained the syllabus for the opinion assigned to a specific sub-field called 'syllabus'. After cleaning the JSON objects of HTML tags and extracting the opinion-syllabus pair from each JSON object, I found that about 600 of the opinions from those clusters were only 1 or 2 sentences long. The opinion was either "Justice (last name) delivered the opinion of the Court." or "Justice (last name) delivered the opinion of the Court. Justice (last name) took no part in the consideration or decision of this case." For these opinion-syllabus pairs, the actual opinion was evidently absent, so I did not include them in the

compiled dataset of opinion-syllabus pairs. The final dataset used for this research project consists of 814 US Supreme Court opinion-cluster pairs.

B Legal Domain-Specific Vocabulary

Specialized domains, such as law, have their own vocabulary and sentences in the documents of these domains contain words from both the original language model's vocabulary and domain-specific vocabulary. Tai et al., 2020 It was therefore necessary to develop a legal domain-specific vocabulary, even if not every term in the vocabulary existed in the dataset of US Supreme Court opinion-syllabi pairs. To do this, I downloaded the Black's Law Dictionary iPhone app, the electronic form of second most recent edition of Black's Law Dictionary. I then screen-recorded my iPhone as I slowly scrolled down on the list of legal terms. I then converted those videos into screenshots. After compiling those screenshots into several PDF files, I used optical character recognition to convert the PDFs into several text files. After editing the text files for any mistakes, they were compiled into a single text file. There are slightly less than 45,000 legal terms in the legal domain-specific vocabulary list. A simple Python script identified which legal terms from the full vocabulary list were in the dataset and only those terms were added to the BERT and BART language model tokenizers prior to their respective summarization tasks.

Since legal case judgments almost always include an extensive amount of case citations, I also had to extract the case citations from every US Supreme Court opinion in the dataset to add them to the BERT and BART tokenizers. To do this, I used the Eyecite tool from the Free Law Project which can identify, extract and annotate any case citation found in a US legal document. The Eyecite tool returns a CitationBase object for every case citation it identifies. I did not need the CitationBase objects so I used the CitationBase object's matched_text method to retrieve the non-annotated case citation identified in the opinion for every returned CitationBase object. Free Law Project, 2023 These case citations were added to the BERT and BART tokenizers prior to their respective summarization tasks.

C Transfer Learning

This research project, and any subsequent research on this subject, rests on the foundations of transfer learning, which refers to how learning systems can quickly adapt themselves to new situations, tasks, and environments. Transfer learning is defined as given a source domain D_s and learning task T_s, a target domain D_t and learning task T_t, transfer learning aims to help improve the learning of the target predictive function $f_t(\cdot)$ for the target domain using the knowledge in D_s and T_s where D_s and T_s where $D_s \neq D_t$ or $T_s \neq T_t$. Yang et al., 2020

There are several reasons why transfer learning has become crucial to machine learning research. The first is that it reflects how humans learn from small data – as children, we generalize from few example to whole concepts. Second, datasets used outside of research tend to be small-sized, isolated, and fragmented and transfer learning is a suitable solution for addressing this challenge. Transfer learning can make machine learning models more reliable and robust, which is necessary if they are to be continuously deployed over time. Lastly, transfer learning is important for maintaining user privacy as pre-trained models can be downloaded and adapted at the edge of a computer network without leaking user data. Yang et al., 2020

D Transformers

BERT and BART are two examples of neural language models with the Transformers architecture. Transformers map sequences of input vectors $(x_1, ..., x_n)$ to sequences of output vectors $(y_1, ..., y_n)$. Transformers are made up of stacks of transformer blocks, each of which is a multilayer network consisting of simple linear layers, feed-forward networks, and self-attention layers. Jurafsky and Martin, 2023 An attention function maps a query and a set of key-value pairs to an output, where the query, keys, values and output are all vectors. Vaswani et al., 2017 The current focus of attention when being compared to all of the other preceding inputs is the query, the preceding input being compared to the current focus of attention is the key, and the value is used to compute the output for the current focus of attention. Jurafsky and Martin, 2023

E BERT

Prior to the release of Bidirectional Encoder Representation from Transformers (BERT), language models with the Transformers architecture were unidirectional, meaning that every token can only attend to previous tokens in the self-attention layers of the Transformer. This limited their capacity to be fine-tuned on token-level tasks such as question answering, where context must be incorporated during fine-tuning on Q&A datasets for accurate responses to unseen questions. BERT alleviated this problem with a masked language model that randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked word based only on its context. Devlin et al., 2018

F LEGAL-BERT

LEGAL-BERT is a family of BERT models for the legal domain, intended to assist legal NLP research. The LEGAL-BERT model used for this research project is LEGAL-BERT-BASE-UNCASED, which is the complete LEGAL-BERT model that has been pre-trained on five sets of legal corpora: 61K pieces of EU legislation, 19K pieces of UK legislation, 19K European Court of Justice (ECJ) cases, 12K European Court of Human Rights (ECHR) cases, 164K US court cases, and 76K US contracts.Chalkidis et al., 2020

G BERT Extractive Summarizer

BERT Extractive Summarizer is a Python-based RESTful service for extractive text summarization that utilizes the BERT model for text embeddings and K-Means clustering to identify sentences closest to the centroid for summary selection. As BERT Extractive Summarizer provides the option of using a custom BERT model and BERT tokenizer to instantiate the Summarizer model, I used the LEGAL-BERT model and tokenizer to instantiate the Summarizer model. Miller, 2019

H BART

Bidirectional and Auto-Regressive Transformers (BART) is a language model with the Transformers architecture that is implemented as a sequence-to-sequence model with a bidirectional encoder over corrupted text and a left-to-right autoregressive decoder. It is essentially a hybrid of the bidirectional encoder of BERT and the left-to-right autoregressive decoder of OpenAI's GPT-1. This architecture makes it particularly well-suited for content generation tasks like abstractive text summarization. Lewis et al., 2019

I BillSum

The BART that was used for abstractive summarization of US Supreme Court opinions for this research project was first pre-trained on the BillSum corpus. The BillSum corpus is a dataset for automatic text summarization of US Congressional and California state bills. There are almost 19K bills in the training set and 3K bills in the test set. Each US bill in this dataset is paired with a human-written summary from the Congressional Research Service (CRS), making it suitable for training language models for abstractive summarization tasks. Kornilova and Eidelman, 2019

I chose a BART model that was first pre-trained on this corpus for abstractive text summarization because it is a corpus of legal documents from the United States and my hypothesis was that a BART model pre-trained on the BillSum corpus would be generalizable to fine-tuning on a small dataset of US Supreme Court opinion-syllabus pairs. The problem with this hypothesis, I later realized, is that the legal language of case judgments is not the same as the legal language of legislation as they are written by legal practitioners from separate branches of government with differing aims and backgrounds.