LegalEASE:
Streamlining Legal Complexity Through Summarization

Jordan Fan | John Gibbons | Elias Tavarez
https://github.com/etavrz/w266-final-project-legalease

Abstract

US Congressional bills are often made inaccessible to the general American public due to their long length and use of legal jargon. In this paper, we tackle this issue by generating abstractive summaries using a 2-stage summarization process. The first stage generates an interim summary using various summarization methods including a term-frequency based extractive summarization, a BERTSum extractive summarization, and an abstractive summarization generated from a fine-tuned BART model. The second stage utilizes various summarization models, including BART, PEGASUS, and T5, to generate the final abstracted summary. We explored the accuracy and readability of the generated summaries and saw varying results for each method, suggesting the models and methods perform better in different contexts. However, these methods did not notably improve the readability of the summaries.

1 Introduction

Democracy is built upon the foundations of an informed citizenry; understanding what legislation is being passed at any point in time helps the public make better decisions when voting. This is oftentimes complicated by the fact that legislative bills are very long, which could potentially be exploited to hide aspects of a bill during passage. Furthermore, the sheer number of bills makes it difficult for the public to keep up. In 2021, over 10,000 bills and joint resolutions were introduced on the US Congressional floor [Kight (2022)]. Combining these obstacles to civic engagement with the fact that “more than 70% of Americans fail a basic civic literacy quiz" [of Commerce Foundation (2024)] creates a dangerous environment where most Americans do not know nor have the bandwidth to track the legislation that is being passed by their political representatives. **Our goal in this paper is to present a novel way of summarizing legislative text to help the American public stay informed and politically active**. By summarizing potential legislation, it becomes more accessible to a wider population of voters, lowering the barrier to entry most citizens have in staying politically active. Making legislation faster to digest will help voters hold elective officials accountable for the laws they are drafting. We see this work as a tool that could be used to help citizens know what's going on and play a more active role in the legislative and governmental process. In this paper, we utilize 3 different models with various forms of extractive and abstractive summarization to create legislative summaries.

2 Background

While not congressional bills, in *Plain English Summarization of Contracts*, [Manor and Li (2019)] highlight the complexities of summarizing legal text into plain English. The authors initially experimented with unsupervised extractive methods of summarizing and tried to implement a novel approach of summarizing legal text in plain English for a non-legal audience. However, they were limited by a lack of corpora necessary to build a supervised model. With the increase in summarization datasets and the rapid advancement of deep learning in recent years, summarization models, composed largely from the transformer architecture [Vaswani et al. (2023)] are able to achieve the sight that Manor and Li set. BERT [Devlin et al. (2019)], a model that was trained for classification-type problems, was adapted to the task of summarization. BERTSum builds upon BERT by inserting CLS tokens in between each sentence and then trained to classify sentences that are relevant to summaries from annotated CNN and NYT news articles, ultimately generating extractive summaries [Liu (2019)]. While BERTSum advanced the area of extractive summarization, BART was introduced and advanced abstractive summarization. Trained to reconstruct the original text after noise has been introduced to the input, the model provided more flexibility and achieved state of the art abstractive summarization at the time [Lewis et al. (2019)]. While not solely a summarization model, T5, an encoder-decoder model, utilizes transfer learning to create a single model with the same loss function and decoding procedure for a variety of tasks [Raffel et al. (2023)]. The authors of the T5 paper demonstrate that the model can perform well on abstractive summarization tasks after being trained on datasets from Colossal Clean Crawled Corpus (C4), News Websites, and Wikipedia. PEGASUS, which yielded state of the art performance on abstractive summarization on 12 different summarization tasks is another model that was considered for this task. PEGASUS achieved said performance by taking a different pre-training approach by removing/masking entire sentences and masking tokens from the input text [Zhang et al. (2020)]. Similar to T5, PEGASUS was pre-trained with the C4 dataset and HugeNews, a new dataset comprised of datasets like XSum and CNN/DailyMail. Lastly, work specific within the legal domain includes the development of the LegalSumm model, which addresses the problem of summarizing long legal text. The model creates multiple views of the document by chunking each paragraph at varying lengths and concatenating the chunks. A summary is generated for each view, and the best summary is chosen based on how related the summary is to the view using an entailment model [Feijo and Moreira (2023)]. Not only does the paper tackle the problem of summarizing long text through its chunking method, but it also addresses model hallucinations from its use of the entailment model.

3 Data

One of the most widely used datasets for legal summarization is BillSum [Eidelman (2019)]. The dataset consists of Congressional bills from 1993-2016 and California bills from 2015-2016, and the summaries were written by their Legislative Counsel. Although many pre-existing work utilize the dataset and would give us a comparison point for the evaluation of our models, an initial exploration reveals that the summaries from BillSum appear to maintain much of the legal jargon in the bill. Wanting to create a model that generates summaries which are well understood by the general public, we opted to use a legal dataset that contains summaries that are easier to read. Consequently, we utilized BillSum to train an intermediary model for our main summarization task, as referenced in the *Abstractive-Abstractive Summarization* section under **Methods**.

The dataset used for our main summarization task is the us-*congress-117-bills* [1] dataset pulled from Hugging Face [Wolf et al. (2020)]. The data was scraped from *congress.gov* and consists of all House and Senate resolutions from the US introduced between the years of 2021 to 2022. For each bill, an abstractive summary is provided by a congressional staff member. For more information about additional features and the data processing steps, refer to **Appendix A**.

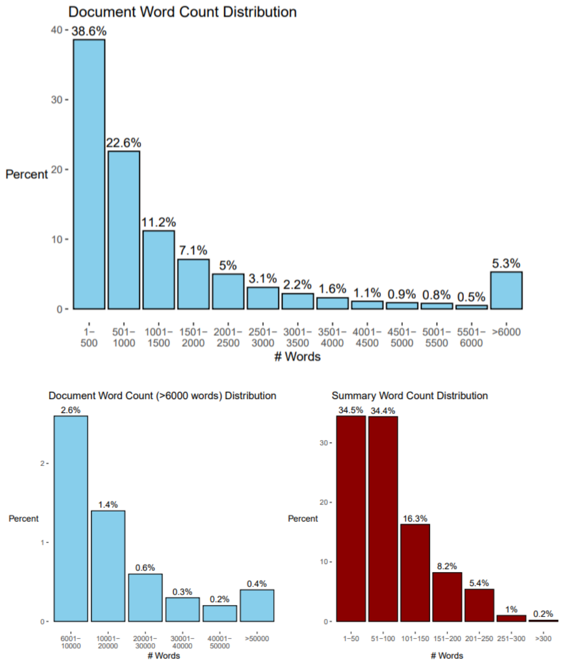

Figure 1: Bill and Summary Word Count Distribution

4 Methods

For both the sake of producing the best possible model and for our own learning, we chose to tackle four different experimentation techniques which we applied to the models we selected. These four were: baseline modeling, term-frequency extractive modeling, BERTSum extractive modeling, and abstractive-to-abstractive modeling.

4.1 Baseline Models

In what we called baseline modeling, we simply trained our models with an untransformed source

[1] https://huggingface.co/datasets/hheiden/us-congress-117-bills

text. The only variation we had in our baseline models was the context length. BART was pretrained with a context length of 1024 while the PEGASUS and T5 models we chose were finetuned with a context length of 512 (the maximum input length for each respective model).

4.2 Extractive-Abstractive Summarization

With almost $\frac{2}{3}$ of bills exceeding 500 words, we needed to reduce the size of the input text to fit within the parameters of our baseline models. For our baseline, we took the first 512 or 1024 tokens from the text and summarized the text to a max output length of 128 tokens, which encompasses over 75% of summary lengths. However, we believe important context can be introduced beyond the first 512 or 1024 tokens, especially for longer bills. As such, various first stage summarization methods were applied to the text as an attempt to pull the most salient information that will then be used as input to our second stage abstractive summarization. The following methods were used to generate extractive summarizations of max word counts of 500 and 1000 words from the text before passing through our baseline models to fine tune for abstractive summarization.

Term-Frequency Extractive Summarization - This method extracts the most important sentences in the bill based upon how often the words in the sentences appear in the document. Only whole sentences would be added to the summary to avoid cutting off a sentence midway and losing contextual information. For each sentence, the importance is scored by taking the average of the term-frequency probabilities of its words, and the sentence with the highest score is added to the summary. The weight of each word in the most important sentence is then lowered to reduce the possibility of extracting similar, redundant sentences in the next most important sentence. This process of scoring each sentence, adding the highest scoring sentence to the summary, and reducing the probabilities for each word in the chosen sentence is repeated until the max word length is reached. Once the word limit is reached, the extracted sentences are sorted to its original order in the bill to help our models summarize in the right order.

BERTSum Extractive Summarization - This method utilizes a pre-trained BERTSum model to create the extractive summarization. We theorized that a BERT model trained specifically for the task of extractive summarization would be able to pick out more pertinent information than one that just uses term-frequency. We adapted the approach of creating different views of the document from the LegalSum paper [Feijo and Moreira (2023)] by splitting the text into chunks, and then using the BERTSum model to summarize each chunk. Similar to the *Term-Frequency Extractive Summarization* method, we produced chunks without splitting a sentence midway to avoid losing contextual information. Once BERTSum produces an extractive summarization for each chunk of text, the summaries are concatenated together. If the concatenated summary is still above the word limit, the process of chunking the text, summarizing the chunks, and concatenating the summaries is repeated until the target word limit is reached. We reasoned the summaries among the chunks could contain overlap, so taking the extractive summary of the summaries would reduce redundancy.

4.3 Abstractive-Abstractive Summarization

We explored the use of a 2-stage abstractive summarization model. We decided to utilize the BillSum data set and our most performant baseline model (BART with 1024 input tokens) to train a 1st stage abstractive summarization model. We chose to use the BillSum data set for 2 reasons. The primary reason being we wanted to conserve our actual congressional bill data set for the main summarization task. If we used our congressional bill data set, we would sacrifice a large portion to train this 1st stage summarization model. Second, we reasoned that the BillSum data set would be approximately close in context/syntax to our primary congressional bill data set. This assumption allows us to employ transfer learning while attempting to minimize the issue of model hallucination. We filtered the BillSum data set to input examples that were at most 1000 words to fit within the BART model. After fine tuning the BART model on the BillSum dataset, we used the model to generate summaries for our congressional bill dataset. We split the congressional bill into 1000 word chunks to pass into the fine-tuned BART model, and then concatenated the summaries together. Any summaries that exceed our word limits (< 2% of text) are passed through our *Term-Frequency Extractive Summarization* process to extract sentences until the target word count is reached.

5 Results and Discussion

In evaluating our models, we looked at four different metrics: ROUGE-L [Lin (2004)], BLEURT [Sellam et al. (2020)], Entailment (Contradiction) [Dagan et al. (2009)], and Flesch Reading Ease [Kincaid et al. (1975)]. For all metrics, we calculated the average score across all examples in our test set. We consider ROUGE-L and BLEURT as the two metrics that gauged the performance of our models and Entailment/Reading Ease as the metrics to gauge their coherence and readability, respectively. In Entailment, we compare the candidate summary with the reference summary to determine if it is entailment, neutral, or contradictory. We focused on the contradiction rate as a proxy for an estimation of the level of model hallucination. With respect to Flesch Reading Ease (scale of 0-100), we aimed for a score in the 50-60 range, equivalent to a 10-12th grade level. We strive for this score as part of ensuring that generated summaries are accessible to a wider public which may not be college educated.

For each of our experiments, we only report the aforementioned metrics from the best performing model and report all other metrics (which include additional ROUGE, BLEURT, and Entailment scores) in **Appendix B**.

5.0.1 Off-The-Shelf Models

Model	R-L	BLRT	CP	Read
BART	**21.85**	-0.93	2.38%	28.83
PEGASUS	19.76	-0.88	**1.85%**	**47.38**
T-5	19.99	**-0.75**	3.89%	47.15

Table 1: Off-The-Shelf Model Results
R-L = Rouge-L, BLRT = BLEURT, CP = Contradiction %, Read = Flesch Reading Ease

As expected, model performance for the models off-the-shelf was fairly low for all metrics except for Flesch Reading Ease. This shows the models in their raw form struggled with the complexity of the syntax and context of the legal documents in our data set.

5.0.2 Baseline Models

Model	R-L	BLRT	CP	Read
BART-512	**37.66**	**-0.21**	**3.44%**	**26.97**
BART-1024	36.98	-0.25	3.71%	24.47
PEGASUS	35.61	-0.30	5.30%	24.81
T-5	35.42	-0.25	6.10%	26.62

Table 2: Baseline Model Results

In training our baseline models, we were most curious in seeing how much better the performance was relative to the off-the-shelf source models when considering this type of legal data. Across all of our models, we achieved a significant increase in ROUGE-L and BLEURT scores as expected; however, the Flesch Reading Ease scores noticeably reduced and the contradiction percentage increased. This aligns with our expectations as once these models had conceptualized our legal corpora, which tend to have more sophisticated language, the generated summaries would become less accessible. Most surprising to us was that the BART model which used 512 tokens surpassed the performance of the 1024 token model. Upon reviewing manually generated summaries of the 512 token BART model versus the 1024 token BART model, the 512 token model tended to generate summaries that are more verbose and focus on specific areas of the bill, while the 1024 token model generalizes more and is more succinct.

For the rest of the analysis, BART will refer to the model using 1024 tokens

5.0.3 Term-Frequency Extractive Summarization

Model	R-L	BLRT	CP	Read
BART	**39.12**	**-0.21**	**1.59%**	23.62
PEGASUS	33.25	-0.36	5.03%	**26.26**
T-5	33.66	-0.28	5.57%	**27.76**

Table 3: Term-Frequency Extractive Model Results

Using the term-frequency based extractive summarization as the input text to fine tune our models, we saw improvements in all of the metrics for our BART model except for Flesch Reading Ease, which decreased slightly. On the flip side, T5 and PEGASUS saw slight improvements in the Flesch Reading Ease while seeing a decrease in the ROUGE-L and BLEURT score. Through our manual evaluation of a sample of text, we found that the term-frequency extractive summarization works best for midsize text; using a term-frequency method can cut extraneous sentences and pull the most relevant text. However, the term-frequency based method suffers when the text is long and covers many different areas; the extractive summarizer picks out sentences that have little supporting context sentences, causing the second-stage abstractive summarizer to produce unintelligible text. An improvement to address this issue would be to also extract the sentence before and after the most important sentence to fill in the missing context.

5.0.4 BERTSum Extractive Summarization

Model	R-L	BLRT	CP	Read
BART	**36.90**	**-0.23**	**5.03%**	27.86
PEGASUS	31.24	-0.42	8.22%	**33.13**
T-5	33.88	-0.26	5.57%	30.03

Table 4: BERTSum Model Results

Using the BERTSum based extractive summarization as the input text to fine tune our models, we saw decreases in ROUGE-L and BLEURT scores for all our models and increases in contradictions. Through manual evaluations, we found that the summaries on the BERTSum extracted text tended to get caught up in specific areas of the bill, and the abstracted summaries tended to hallucinate, explaining the poorer scores in our metrics. BERTSum was trained using a greedy algorithm to select sentences to maximize ROUGE scores. This greedy algorithm could have created a fixation to certain portions of the text to extract all the sentences relevant to a topic. Although the BERTSum based approach produced negative ROUGE, BLEURT, and contradiction percentage results, the Flesch Reading Ease increased. The BERTSum model gave more context that the term-frequency approach was missing to produce more legible summaries.

5.0.5 Abstractive-Abstractive Summarization

Model	R-L	BLRT	CP	Read
BART	**36.74**	**-0.24**	5.72%	27.16
PEGASUS	32.82	-0.35	5.72%	27.25
T-5	35.12	-0.26	**4.36%**	**28.72**

Table 5: Abstractive-Abstractive Model Results
The results are calculated on examples that fall under the 97.5 percentile of word counts (∼10k words). Text beyond that length took too long to generate the first-stage abstractive summaries.

ROUGE-L, BLEURT dropped for all 3 models compared to baseline. Additionally, contradiction percentage increased. Reading Ease also increases slightly along all 3 models. Surprisingly, this abstractive to abstractive model performed well given our concerns around hallucination. After manually evaluating the summaries from all 3 models for this section, the outputs for T5 and PEGASUS performed surprisingly well. Only BART showed some tendency towards hallucination. One way to improve results from this approach is by fine tuning the 1st Stage BART summarizer on more data.

5.0.6 Categorical Evaluation

As part of our evaluation framework, we were curious as to how the number of examples in the training set would affect our model's capability to generate summaries in the test set. To measure this, we split our bill categories into three groups: low, medium, and high number of bills in the training set, with all having roughly the same number of categories (∼11). Our expectation was that we would find an increasing performance with the number of bills. However, as our results showed (see **Appendix D**), that trend was not evident in the results nor was any other trend noticeable. Consequently, we conclude that the quality of the text and summary is most likely more important than is the quantity of bills in developing a proficient model.

5.0.7 Manual Evaluation of Summaries

Model	Base	TF	BERTSUM	Abs
BART	4.13	3.69	3.44	4.17
PEGASUS	4.63	4.06	3.25	4.50
T-5	4.69	4.06	3.75	5.00

Table 6: Average scores for manual evaluations
For manual evaluation we utilized a ordinal scoring of 1-5 with 5 being the best across 4 dimensions of coherence, consistency, fluency and relevance. Score for each model is the average across these dimensions

For manual evaluations, we sampled 4 passages from our test data set. These 4 passages corresponded to the shortest, longest, the least examples and most examples (based on category). We evaluated the generated summaries across 4 of our model experiments for each passage and annotated a 1-5 rating (5 being the best) across 4 dimensions of coherence, consistency, fluency and relevance (this approach being inspired by Liu et al. (2023). We then took an average score for each passage/ experiment combination. Note the drop in rating for BERTSUM experiment across all models. Interestingly, while BART outperformed T5 and PEGA-

SUS on ROUGE/ BLEURT metrics, BART lagged in manual evaluations. This is due to BART producing more verbose summaries than T5 and PEGASUS.

5.1 Future Improvements

For future development, there are several techniques and methods that could be implemented given sufficient time and more importantly, resources, that would potentially lead to improved summarization performance.

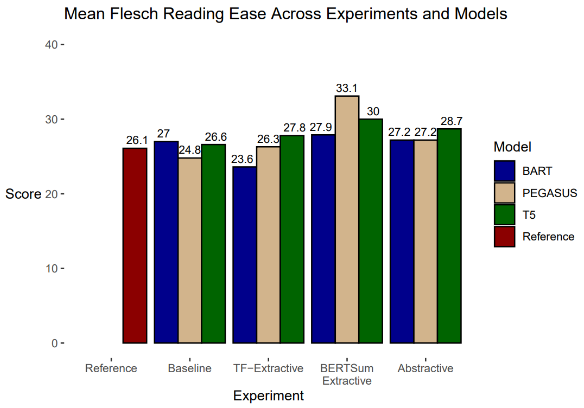

Figure 2: Average Readability Scores

The first of these improvements would be to introduce a model which is suited for larger token windows since as was noted in the **Data** section, roughly $\frac{2}{3}$ of bills exceeded 500 words. One such model is the LongFormer [Beltagy et al. (2020)], which introduces much larger context lengths of 2,048 and upwards while avoiding the quadratic complexity which plagues self-attention. Additionally, with respect to the low Flesch Reading Ease scores we achieved, the original reference summaries averaged a score of 26 on the Flesch Reading Ease scale and our models had marginal gains for each experiment, with the exception of PEGASUS in the BERTSum experiment. We could potentially increase these scores by adding a prompting scheme to reduce the complexity of our generated summaries. However, it must be stated that this would most likely lead to reduced training performance but would require less resources than finding a new training dataset. Furthermore, we can also consider fully training models from scratch (or using previously trained models that were solely trained on legal text) using a larger unified legal dataset in hopes of preventing potential hallucination from the current pre-trained corpora (which mainly uses newspapers, scientific articles, etc.). Lastly, we may consider pairing our model with a retrieval augmented generation (RAG) [Lewis et al. (2021)] system to more efficiently utilize our data and provide the specific text from which our models gathered their answer.

Considering evaluation, ROUGE/BLEURT are by no means a perfect evaluation metric [Schluter (2017)] and we see that evidenced by our manual evaluation. Since ROUGE is a recall-based metric, a longer summary can oftentimes score higher despite having a worse summary as demonstrated by BART. In general, evaluating abstractive summarization is a difficult task, but there are seemingly more creative ways to attempt to solve this problem. One such approach, which we referenced in our manual evaluation, is Liu et al. (2023) which utilizes GPT4 with chain of thought (CoT) as an evaluator that is capable of providing a more holistic scoring scheme. Applying this set of evaluation techniques may provide metrics which are more telling of the best model and allow for customizable metrics based on the researcher's design.

6 Conclusion

As stated at the beginning of our paper, our goal was to present a novel way of summarizing legislative text to help the American public stay informed and politically active. To do so, we conducted several experiments leveraging extractive and abstractive summarization. Some insights we found were that while BART was our most performative model in ROUGE/BLEURT scoring, when manually evaluating generated summaries, PEGASUS/T5 generated the most readable and generalizable outputs. This reinforced the notion that manual evaluation for generative models is key to overall performance assessment and that, while popular, ROUGE may not be a perfect indicator of the desired performance. Additionally, we found that the majority of our generated summaries had a relatively low readability level (college level), which conflicts with the intent of our project goal, as roughly 4 in 10 Americans over 25 are college educated [Schaeffer (2022)]. However, this can be attributed to the low readability pertaining to our dataset. Thus, if we aim to widen accessibility, we would most likely need a new dataset with more readable summaries to model. This final point is the most prominent area for future improvement, as improving readability of the generated summaries would ensure wider access for the American public.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(4):i–xvii.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics.

D.d. Feijo and V.P. Moreira. 2023. Improving abstractive summarization of legal rulings through textual entailment. *Artificial Intelligence and Law*, 31:91–113.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Stef W. Kight. 2022. How many bills congress has passed so far. Accessed: April 10, 2024.

J Peter Kincaid, Robert P Fishburne Jr., Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *NTIS*, ADA006655.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune bert for extractive summarization.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Laura Manor and Junyi Jessy Li. 2019. Plain english summarization of contracts.

U.S. Chamber of Commerce Foundation. 2024. New study finds alarming lack of civic literacy among americans. Accessed: April 10, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Katherine Schaeffer. 2022. 10 facts about today's college graduates. Accessed: April 10, 2024.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

A Additional Data Features and Preprocessing

In addition to the original language of the bill and its respective summary in the us-congress-117-bills dataset, the data includes the policy area the bill falls under (1 of 33 categories) and the name of the representative, political party, and state that sponsored the bill. The dataset includes 15,186 rows and originally has a 75/25% train-test split, placing 11,389 bills in the training set and 3,797 bills in the test set. Some pre-processing and exclusions were performed, and the test set was further split for our training and evaluation.

Many of the summaries include a header that is just the name of the bill, contributing little information about the contents of the bill. We decided to remove these headers from the summary to allow our models to yield more descriptive summaries. We excluded examples that did not contain a summary or had less than 10 words; summaries with less than 10 words were examined and determined that the summary was truncated and did not summarize the bill. After filtering out missing and short summaries, the data was reduced to 15,042 rows and consequently removing all bills with a policy area category of Private Legislation. The test set was split so that 90% would go into the validation set, and the remaining 10% would go into the test set and will be used for manual evaluations. The final data contains 11,277 bills in the training set, 3,388 bills in the validation set, and 377 bills in the test set.

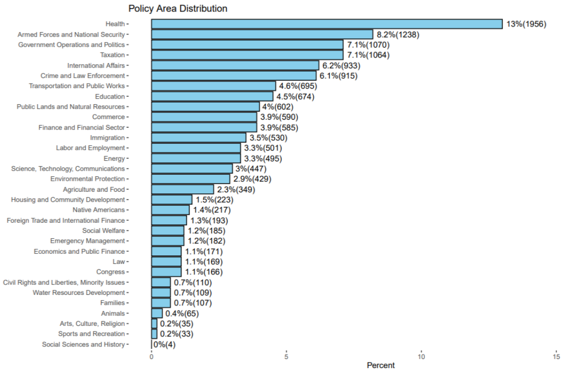

Figure 3: Policy Area Distribution

B Holistic Modeling Results

In addition to the metrics displayed in the Results section, we calculated the ROUGE1, ROUGE2, ROUGE-L-SUM, the median BLEURT score, the median Flesch Reading Ease score, Entailment (Entailment), and Entailment (Neutral) scores.

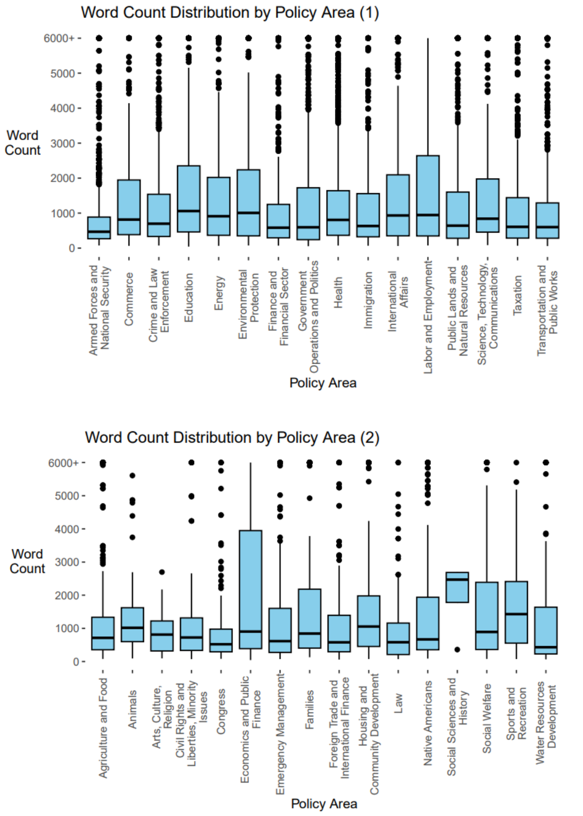

Figure 4: Word Count Distribution Across Policy Areas

B.1 Off-The-Shelf Models

Model	R1	R2	R-L-S	BLRT
BART	31.22	12.25	21.85	-0.93
PEGASUS	25.52	10.99	19.72	-0.87
T-5	26.93	10.34	19.99	-0.74

Model	Read	Entailment	Neutral
BART	34.46	11.1%	86.4%
PEGASUS	44.41	12.2%	85.9%
T-5	49.45	21.4%	74.5%

Table 7: Off-The-Shelf Model Additional Results

B.2 Baseline Models

Model	R1	R2	R-L-S	BLRT
BART-512	48.11	29.21	37.68	-0.21
BART-1024	47.38	28.08	36.98	-0.26
PEGASUS	42.15	26.14	35.53	-0.32
T-5	43.20	25.65	35.44	-0.24

Model	Read	Entailment	Neutral
BART-512	29.18	32.8%	63.6%
BART-1024	25.46	38.9%	57.2%
PEGASUS	26.14	43.2%	51.4%
T-5	27.83	38.4%	55.4%

Table 8: Baseline Model Additional Results

B.3 Term-Frequency Extractive Summarization

Model	R1	R2	R-L-S	BLRT
BART	49.16	30.32	39.09	-0.21
PEGASUS	39.56	23.67	33.20	-0.38
T-5	41.56	23.63	33.63	-0.30

Model	Read	Entailment	Neutral
BART	24.78	33.9%	64.4%
PEGASUS	27.49	45.8%	49.0%
T-5	28.17	40.5%	53.8%

Table 9: Term Frequency Extractive Model Additional Results

B.4 BERTSum Extractive Summarization

Model	R1	R2	R-L-S	BLRT
BART	46.48	27.67	36.92	-0.24
PEGASUS	36.64	21.91	31.22	-0.47
T-5	41.27	23.17	33.88	-0.27

Model	Read	Entailment	Neutral
BART	28.33	33.9%	61.0%
PEGASUS	33.24	45.0%	46.6%
T-5	30.5	41.6%	52.7%

Table 10: BERTSum Model Additional Results

B.5 Abstractive-Abstractive Summarization

Model	R1	R2	R-L-S	BLRT
BART	45.48	27.82	36.77	-0.26
PEGASUS	38.48	23.37	32.80	-0.38
T-5	42.69	24.97	35.22	-0.26

Model	Read	Entailment	Neutral
BART	27.66	31.8%	62.3%
PEGASUS	26.81	43.0%	51.2%
T-5	28.17	37.6%	58.0%

Table 11: Abstractive-Abstractive Model Additional Results

C PEFT Modeling

Parameter Efficient Fine Tuning (PEFT) is a technique that reduces the number of overall parameters to be trained. As a result, computational costs are considerably reduced while maintaining performance similar to that of training the entire model parameters. For our set of experiments, we applied a Low Ranking Adaptation (LoRA) Hu et al. (2021) to each of our three different models and specifically selected to only update the weights of the linear layers in each of our models.

Model	R-L	BLRT	% Mem	% Train
BART	35.57	-0.29	21%	2.28
PEGASUS	28.64	-0.55	32%	2.25

In our PEFT models, we wanted to confirm two items. First, that the metrics do not deviate noticeably from the baseline models and that our resource usage was reduced. While we did not notice a large deviation in performance with BART, the PEGASUS model did perform noticeably worse than the baseline model. While that would discourage switching to this type of model for a production setting, the resource utilization was reduced significantly, which can have major implications when considering the business application of these models and hence they should not be easily overlooked. Due to difficulties with model configuration in T5 when using LoRA, we were unable to report metrics for T5.

D Categorical Evaluation Results

As explained in the **Results and Discussion** section, we split our test set into three groups, low, medium, high number of bills based on the number of examples in the training splits. These three groups were composed of the following categories and achieved the ensuing results:

- **low** [863 bills in training set]: Social Sciences and History, Sports and Recreation, Arts/Culture/Religion, Animals, Families, Water Resources Development, Civil Rights and Liberties/Minority Issues, Congress, Law, Economics and Public Finance, Emergency Management

Model	R-L	BLRT	CP	Read
BART	37.91	-0.22	5.41%	23.17
PEGASUS	36.36	-0.27	8.11%	26.17
T-5	37.20	-0.20	2.70%	33.83

- **medium** [2679 in training set]: Social Welfare, Foreign Trade and International Finance, Native Americans, Housing and Community Development, Agriculture and Food, Environmental Protection, Science/Technology/Communications, Energy, Labor and Employment, Immigration

Model	R-L	BLRT	CP	Read
BART	36.90	-0.24	1.08%	22.49
PEGASUS	33.82	-0.31	4.30%	23.77
T-5	31.74	-0.29	1.08%	27.49

- **high** [7735 in training set]: Finance and Financial Sector, Commerce, Public Lands and Natural Resources, Education, Transportation and Public Works, Crime and Law Enforcement, International Affairs, Government Operations and Politics, Taxation, Armed Forces and National, Security, Health

Model	R-L	BLRT	CP	Read
BART	39.04	-0.23	2.83%	24.64
PEGASUS	36.03	-0.26	5.26%	25.00
T-5	36.70	-0.21	3.64%	29.34

E Manual Summary Evaluation

Reference Summary	**Model / Model Summary**
Shortest Text This bill repeals provisions terminating the First Responder Network Authority (FirstNet) and the Government Accountability Office reporting requirement for such termination. The FirstNet mission is to deploy, operate, maintain, and improve the first high-speed, nationwide wireless broadband network for first responders.	**BART**-*Abstractive* This bill reauthorizes through FY2027 the First Responder Network Authority, which provides funding for state and local governments to establish and operate a first responder network.
	PEGASUS/T5-*Baseline* (same summary) This bill reauthorizes the First Responder Network Authority.
Mid-sized Text This act directs the Centers for Medicare and Medicaid Services (CMS) to publish and biennially update a list of all rehabilitation innovation centers. The act defines rehabilitation innovation centers as rehabilitation facilities that (1) hold specified federal research and training designations for traumatic brain injury or spinal cord injury research, and (2) serve at least a certain number of Medicare patients.	**PEGASUS**-*TF Extractive* This bill requires the Department of Health and Human Services (HHS) to make publicly available a list of all rehabilitation innovation centers, and to update such list on a website not less frequently than biennially.
Longest Text This bill provides appropriations to federal agencies for the remainder of FY2022, provides supplemental appropriations for activities to support Ukraine, and modifies or establishes various programs that address a wide range of policy areas. The bill includes the 12 regular appropriations bills that fund federal agencies for FY2022. The bill provides supplemental appropriations for several federal agencies to assist Ukraine in responding to the attack by Russia. For example, the bill provides funding for emergency food assistance, migration and refugee assistance, defense equipment, economic assistance, and enforcing sanctions against Russia.	**BART**-*TF Extractive* This bill provides FY2022 supplemental appropriations to the Department of Defense (DOD) for specified purposes, including The bill also provides supplemental appropriations for The bill sets forth requirements for using funds provided by this and other appropriations acts. For example, the bill requires DOD to (1) use specified funds to (2) purchase, construct, or acquire specified equipment, including telecommunications equipment; and (3) carry out specified activities, such as military exercises, to protect U.S. diplomatic facilities. The bill authorizes DOD to provide certain assistance to states, Indian tribes, and
	T5-*BERTSum Extractive* This bill directs the Department of Defense (DOD) to reimburse the Department of Agriculture for the support provided by the Great Lakes Restoration Initiative and the Great Lakes Water Quality Agreement. The bill also authorizes the Department of Agriculture to transfer funds to the National Forest System (NFS) account