The Right to Remain Plain: Summarization and Simplification of Legal Documents

Stanford CS224N Custom Project

Isabel Gallegos
Department of Computer Science
Stanford University
iogalle@stanford.edu

Kaylee George
Department of Computer Science
Stanford University
krgeorge@stanford.edu

Abstract

Legal documents are a critical tool for contractual agreements and protections, but legal jargon and document length can present a barrier to comprehension. Having a tool to summarize and simplify these texts can greatly improve understanding, which is key to ensuring fairness. While previous work has developed summarization and simplification models independently, this work seeks to build on these existing techniques to fine-tune a model for legal domain-specific tasks; understand how fine-tuning on one dataset generalizes to other datasets within the legal domain; and examine the impact of simplification as a pre- or post-processing step in summarization. We show that a large language model fine-tuned on legal data outperforms the model with no fine-tuning and other simple, unsupervised baseline methods, and we also demonstrate that it is possible to generalize across different sub-domains within law, training on a sub-domain dataset and applying the model to other sub-domains. Further, initial results show that simplification as a post-processing step preserves meaning and increases readability, and is promising for building a summarization and simplification pipeline. We conclude with a call for higher-quality legal datasets to improve large language models for the law domain.

1 Key Information to include

- Mentor: Lucia Zheng

2 Introduction

Legal jargon and document length can present a barrier to comprehension of legal agreements and protections. Having a tool to summarize and simplify these texts can greatly improve understanding, which is key to mitigating abuse in legal agreements. While there have been several advances in neural methods for summarization techniques, most of these models [1, 2, 3] have focused on the CNN/Daily Mail dataset [4]. We aim to build on these existing summarization techniques and apply them to a new domain: legal documents.

This task is challenging because the language used in legal contracts is very domain-specific, which is likely not included in the training data for large language models like BERT [5] and BART [1], such as BooksCorpus and Wikipedia, and the CNN/Daily Mail data for summarization models. Futhermore, because there is a lack of data for summarization tasks in the legal domain, many popular supervised methods that are used in broader summarization tasks (e.g., news) are not effective in the legal domain. Additionally, there is a lack of data that addresses the task of simplification of legal jargon. To address this challenge, Manor and Li [6] proposed a new dataset for the task of legal document summarization in plain English. Here, we extend their work by fine-tuning a large language model on this dataset and other similar legal-domain datasets. In particular, our work addresses

three goals: (1) domain-specific summarization: fine-tune a model for the legal domain-specific task of summarization; (2) generalization: understand how training on one dataset generalizes to other datasets within the legal domain; and (3) simplification: examine the impact of simplification as a pre- or post-processing step in the summarization task.

We show that a large language model for summarization, namely Facebook's `bart-large-cnn` model [7], outperforms unsupervised summarization methods and improves performance on legal texts when fined-tuned on a large legal dataset, whereas the model's performance is comparable or worse than these unsupervised methods and `bart-large-cnn` without legal-specific fine-tuning when fine-tuned on smaller legal datasets. We also show that it is possible to fine-tune `bart-large-cnn` on one dataset and evaluate it on a different legal set without sacrificing performance, demonstrating that generalization across datasets is feasible. Finally, we show that, while simplification as a pre-processing step decreases performance, simplification as a post-processing step is a promising direction for improving readability of summarizations. Taken together, our work also highlights the need for quality domain-specific data to improve summarization and simplification tasks for large language models.

3 Related Work

In this work, we examine two tasks, summarization and simplification, in the legal domain. There are two main classes of summarization tasks: extractive, which builds a summary using words in the original text; and abstractive, which generates new words for the summary. There have been several advancements in deep learning models for these tasks. HIBERT is an extractive summarization model that uses a hierarchical sentence and document encoders to generate representations of documents, and is pre-trained on masked sentence prediction and fine-tuned on the CNN/Daily Mail and a New York Times datasets [3]. See et al. presents an abstractive summarization model that copies words by sampling the attention distribution, a form of extractive summarization, and then generates new words from an extended vocabulary; this model is also applied to the CNN/Daily Mail dataset [2]. Facebook has also developed the `bart-large-cnn` model for summarization, which is used in this work. The original BART model, presented in [1], is a sequence-to-sequence Transformer-based model similar to BERT [5], pre-trained by corrupting and reconstructing documents and fine-tuned with various downstream tasks, including sequence classification and generation. The large-sized BART was fine-tuned on the CNN/Daily Mail dataset to produce `bart-large-cnn`.

Simplification is the process of removing complexity from a text, such as through restructuring a sentence or replacing a word, and can be framed as a machine translation problem, translating from complex to simple English. Facebook and Martin et al. [8] have developed ACCESS, a Transformer-based text simplification model that uses neural techniques to control compression, paraphrasing, complexity in text generation, and achieved state of the art performance on the WikiLarge test set of complex-simple sentence pairs. While we use ACCESS in this work, other simplification models also exist, including LSTM-based methods [9] and rule-based pipelines [10].

Finally, while prior work has shown that pre-training language models on domain-specific datasets shows substantial gains over general models pre-trained on generic corpora such as Wikipedia [11] [12] [13], Manor and Li [6] highlight the issues that lack of domain-specific training data can impose: the performance of the summarization methods on their legal dataset was noticeably lower than that of the non-legal datasets, so development of more resources for the legal domain is necessary for future work on the domain to improve.

4 Approach

While there have been several advances in neural methods for summarization, these models are not specific to legal language and can produce poor predictions given legal jargon rare in other contexts. We fine-tune several models to evaluate the efficacy of fine-tuning a large language model on legal documents.

4.1 Fine-tuning BART for legal summarization

To improve upon both non-neural summarization techniques and neural summarization models pre-trained and fine-tuned on non-legal data, we fine-tuned Facebook's `bart-large-cnn` [7] on a legal dataset by dividing the dataset into train, validation, and test sets with a random 70/15/15 split; determining optimal hyperparameters using the train and validation sets; and evaluating the final model on the test set. This fine-tuning procedure was implemented ourselves following Hugging Face's summarization pipeline [14]. This fine-tuning procedure is referred to as `within-dataset`, or `within-dataset-<dataset-name>` to denote which dataset the model was fine-tuned on.

The performance of our fine-tuned `bart-large-cnn` model was compared to five unsupervised, extractive baseline techniques used by [6]: TextRank [15], which calculates similarity scores for sentences using the PageRank algorithm; KLSum [16], a greedy algorithm that minimizes the KL divergence between the predicted and reference summary; Lead-1 and Lead-K, which select the first one or k sentences in a text; and Random-K, which randomly selects k sentences. We used existing implementations for TextRank [17] and KLSum [18], and implemented the others ourselves. Additionally, we included the original `bart-large-cnn` model with default Hugging Face hyperparameters and no fine-tuning as an additional baseline.

4.2 Generalization across legal datasets

Sub-areas of law may each have their own vocabularies and characteristics, and thus we sought to understand how generalizable a model fine-tuned on one legal dataset was to a separate legal dataset, including a dataset in a different sub-domain of law. We fine-tuned the `bart-large-cnn` model on one legal dataset split into training and validation set with a random 85/15 split, and then evaluated the model on the test set of a different dataset. We call this fine-tuning procedure `across-dataset`, or `across-dataset-<dataset-name>` to denote which dataset the model was fine-tuned on. Using the same test set in the `within-dataset` and `across-dataset` experiments, we compared the test set's performance on the `across-dataset` models to that dataset's `within-dataset` performance.

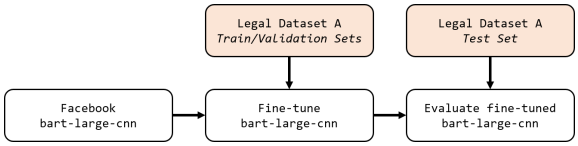


Figure 1: `within-dataset` fine-tuning and evaluation.

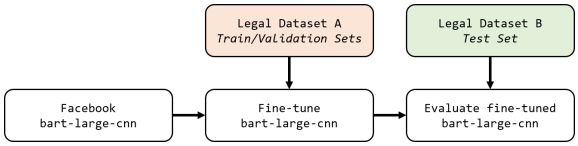


Figure 2: `across-dataset` fine-tuning and evaluation.

4.3 Simplification for pre- or post-processing

No legal datasets exist to our knowledge for both summarization and simplification, despite the importance of simplification in translating complex language to layman's terms. To explore this area, we applied Facebook's `ACCESS` model [8] with default hyperparameters and no fine-tuning to the model input or output as a pre- or post-processing step. For pre-processing, referred to as `pre-simplified`, a simplified document was provided as input into the `within-dataset` model fine-tuned in Section 4.1, with no additional fine-tuning. For post-processing, referred to as `post-simplified`, we performed a forward pass of the `within-dataset` model, and simplified the predicted summary to produce the final output. We compared the results to the `within-dataset` experiments.

Figure 3: `pre-simplified` pipeline.

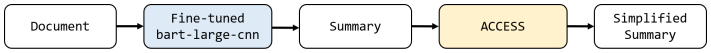


Figure 4: `post-simplified` pipeline.

5 Experiments

5.1 Data

We used four datasets in this work:

1. **TLDR.** The TLDR dataset was developed by [6] for law-specific summarization tasks, and consists of 85 samples from the TL;DRLegal website, which contains software licenses.
2. **TOSDR.** The TOSDR dataset [19] has 361 samples from the "Terms of Service; Didn't Read" (ToS;DR) website, which contains user data and privacy policy agreements.
3. **Billsum.** The Billsum dataset [20] consists of US Congressional and California state bills. We used a subset of the full dataset, taking 2,018 examples with document lengths in the bottom 10% to minimize truncation effects described in Section 5.3.
4. **Tiny Billsum.** The Tiny Billsum dataset was created by randomly sampling 59 training examples and 13 validation examples from the Billsum dataset to replicate the training and validation set sizes of TLDR. Tiny Billsum's test set is identical to that of Billsum.

Dataset	Train/Val/Test (Total) Examples	Mean Document Length (Std)	Mean Summary Length (Std)
TLDR	59/13/13 (85)	127.24 (134.26)	11.56 (8.00)
TOSDR	252/54/55 (361)	40.90 (48.64)	9.41 (5.67)
Billsum	1412/303/303 (2018)	1135.95 (453.45)	140.55 (95.74)
Tiny Billsum	59/13/303 (377)	1147.12 (468.15)	143.45 (94.92)

Table 1: Dataset characteristics.

Each dataset provides a full-length document and reference summary for each example, which are the inputs and outputs of the model (see Tables 5 and 6 for examples). Following the methodology of [6], each dataset was pre-processed with lowercasing, stopword removal, and lemmatization.

5.2 Evaluation method

For summarization evaluation, we used Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which is a set of metrics for evaluating machine produced summarization of texts. ROUGE compares overlapping n-grams between machine-produced summaries against a set of reference summaries (usually human provided). The variations of ROUGE we used were: -1 and -2 for unigram and bigram overlap, and -L for Longest Common Subsequence overlap between machine-produced and reference summaries. We wrote our own script to produce ROUGE scores using the ROUGE library [21].

For simplification evaluation, we used Flesch-Kincaid Grade Level (FKGL) [22], which we computed using the EASSE Python package for sentence simplification [23]. FKGL is a commonly used metric for measuring readability and is computed as a linear combination of the number of words per simple sentence and the number of syllables per word. A limitation of FKGL is that it does not measure well-formedness or semantic preservation of predictions. To supplement FKGL, the `ACCESS` authors also include the SARI metric, which compares the predicted simplification with both the source and target references [8]. However, because there is no reference simplified data, we cannot draw strong conclusions from SARI scores and instead primarily rely on FKGL scores to measure readability.

Finally, to mitigate limitations of ROUGE and FKGL in assessing semantic quality, we developed two additional qualitative metrics, each calculated manually with a scale of "Bad", "Moderate", and "Good":

1. **Quality.** Well-formedness and readability, independent of the reference or prediction summary. Highly-repetitive predictions and undescriptive references are penalized.
2. **Match.** How well the prediction matches the reference semantically. Predictions are penalized for lacking important or adding superfluous content compared to the reference.

5.3 Experimental details

For each `within-dataset` and `across-dataset` experiment, we performed a parameter sweep and chose the following optimal parameters: the epoch and learning rate with the highest average ROUGE performance across two seeds, and the seed with the highest overall ROUGE score on the evaluation set. We selected the maximum `BATCH_SIZE` and `MAX_SOURCE_LENGTH` possible, given computing

and model input length constraints. MAX_SOURCE_LENGTH is much shorter than the average document length for the Billsum dataset, which resulted in significant truncation for model inputs.

		within-dataset			across-dataset		
Hyperparameter	Fixed	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum
BATCH_SIZE	16	-	-	-	-	-	-
NUM_TRAIN_EPOCHS	*[1, 2, 3, 4]	4	4	4	4	4	4
LEARNING_RATE	*[1e-5, 2e-5, 3e-5]	3e-5	3e-5	3e-5	3e-5	3e-5	3e-5
WEIGHT_DECAY	0.01	-	-	-	-	-	-
GRAD_ACCUMULATION_STEPS	8	-	-	-	-	-	-
SEED	*[224, 161]	161	161	161	161	161	224
MAX_SOURCE_LENGTH	128	-	-	-	-	-	-
MAX_TARGET_LENGTH	64	-	-	-	-	-	-

Table 2: Optimal hyperparameters used. * indicates parameters included in the parameter sweep.

5.4 Results

5.4.1 within-dataset Models

The performance of the within-dataset models in comparison to the baseline methods can be seen in Table 3. First, we observe that the bart-large-cnn baseline with no fine-tuning did not outperform the other baseline techniques, likely due to the difference in the distribution and vocabularies between the CNN/Daily Mail data and legal datasets, thus highlighting the need for an improved neural model.

The within-dataset models fine-tuned on the TLDR and TOSDR datasets had comparable or worse performance to all baselines, where within-dataset-tldr had a ROUGE F-1 score on average 8.42 points below the best baseline across R-1, R-2, and R-L; within-dataset-tosdr was 5.94 points below. However, we saw significant improvement over the baselines, including bart-large-cnn with no fine-tuning, with the within-dataset-billsum model, which had a ROUGE F-1 score on average 9.49 points higher than the best baseline for R-1, R-2, and R-L. In the case of Billsum, this suggests that domain-specific fine-tuning provides drastic gains over general-domain language models, highlighting the need for domain-specific training data.

To understand whether TLDR and TOSDR's poor performance can be explained by their small training set size, we compared within-dataset-tldr and within-dataset-billsum to within-dataset-tiny-billsum, where Tiny Billsum contains a subset of Billsum examples, but has the same training size as TLDR. As shown in Figure 5, evaluation on within-dataset-tiny-billsum resulted in a lower ROUGE F-1 score compared to within-dataset-billsum, but still significantly higher than within-dataset-tldr, which implies that TLDR and TOSDR's poor performances may be due to in part, but not entirely, to their small size.

It is notable that within-dataset-tldr and within-dataset-tosdr had worse performance than the bart-large-cnn model with no fine-tuning for almost all ROUGE metrics. Because training set size does not fully explain this observation, we hypothesize that this may be due to the poor dataset quality of TLDR and TOSDR, which we explore in Section 6.

5.4.2 across-dataset Models

The across-dataset results can be seen in Figure 6, which shows the ROUGE metrics for each applicable across-dataset model for each dataset. Because the full dataset was used for training and validation to fine-tune the across-dataset models, we do not evaluate a dataset on its own across-dataset model but instead provide its within-dataset performance for comparison.

The TLDR and TOSDR datasets did not generalize well to the Billsum dataset, which is unsurprising given the poor within-dataset performances of TLDR and TOSDR, but TLDR and TOSDR did generalize well to each other, with the across-dataset-tldr and across-dataset-tosdr models performing similarly to within-dataset models for TOSDR and TLDR, respectively. The across-dataset-billsum appeared to generalize well to all datasets, achieving comparable

	R-1			R-2			R-L		
	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum
TextRank	17.98	7.83	34.47	1.28	2.59	15.39	16.25	7.7	29.09
KLSum	18.05	20.24	24.21	3.10	5.17	10.42	17.69	18.76	21.31
Lead-1	**25.66**	24.74	1.88	**6.98**	7.32	0.02	**24.19**	23.14	1.85
Lead-K	21.14	**25.38**	32.52	3.39	**7.58**	15.64	19.68	**23.78**	30.26
Random-K	12.36	19.60	28.30	1.28	4.94	11.04	11.77	18.32	25.15
`bart-large-cnn`	17.57	18.65	23.51	2.75	3.59	9.79	15.83	17.55	22.36
Fine-Tuned `bart-large-cnn`	15.52	18.08	**43.44**	1.93	3.21	**25.48**	14.13	17.62	**39.92**

Table 3: ROUGE F-1 score metrics for ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) for baseline methods and `bart-large-cnn` fine-tuned on TLDR, TOSDR, and Billsum.

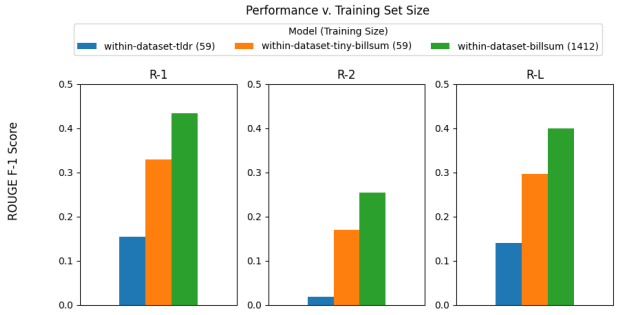


Figure 5: ROUGE F-1 scores for `within-dataset-tldr`, `within-dataset-tiny-billsum`, and `within-dataset-billsum`.

performance to the other `across-dataset` model and the `within-dataset` model for the TLDR and TOSDR datasets.

5.4.3 `pre-simplified` and `post-simplified` Pipelines

Neither simplification as a pre- nor post-processing step improved ROUGE performance, as seen in Figure 7. Using the `pre-simplified` pipeline worsened performance across all datasets more significantly than did the `post-simplified` pipeline. However, simplification often changes sentence structure or word choice, and these syntactic differences are penalized by any n-gram comparison metric, one major limitation assessing these results with ROUGE scores. While more qualitative analysis to better understand these results is presented in Section 6, it may be that the `pre-simplified` pipeline performed poorly because simplifying the input document resulted in removing important information, as observed in Tables 5 and 6. The `post-simplified` results are promising because the performance is only marginally worsened, while still potentially providing more readable output.

As seen in Table 4, FKGL scores improved for all three datasets regardless of whether simplification is applied as a pre- or post-processing step. Taken together with Figure 7, these results suggest that simplification as a post-processing step may be beneficial for the summarization and simplification task pipeline since readability increases and ROUGE scores are, more or less, preserved, compared to ROUGE scores without simplification.

It is worth noting that the TLDR and TOSDR datasets were intended to have reference summaries that were in plain English, while the Billsum dataset doesn't particularly focus on simplification. Due to data availability constraints, we were unable to train or evaluate on only simplified legal data. Thus, more work must be done to more critically evaluate the simplified predictions of the `ACCESS` model. Additionally, as with the `bart-large-cnn` model, there were limitations for input size to the `ACCESS` model, requiring truncated document inputs that could also negatively impact results.

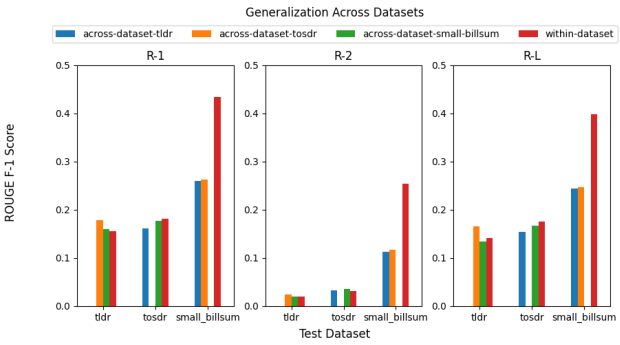


Figure 6: ROUGE F-1 scores for `across-dataset`, with `within-dataset` for comparison.

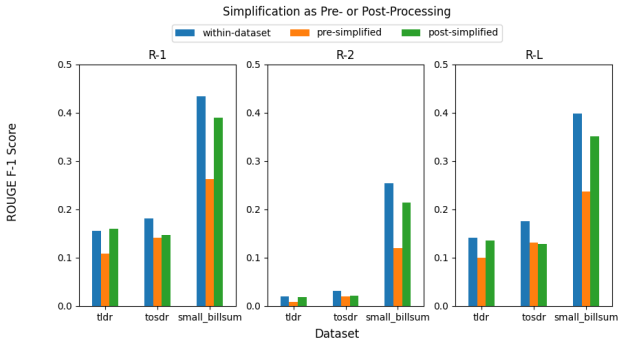


Figure 7: ROUGE F-1 scores for `within-dataset`, `pre-simplified`, and `post-simplified` for each dataset.

	Original			`pre-simplified`			`post-simplified`		
Metric	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum	TLDR	TOSDR	Billsum
SARI	-	-	-	33.07	26.96	37.36	21.57	26.06	37.32
FKGL	14.11	12.65	5.48	16.15	17.98	10.12	16.51	16.61	12.92

Table 4: SARI and FKGL metrics for `pre` and `post-simplified` experiments

.

6 Analysis

6.1 Select Examples

A major limitation of ROUGE and other n-gram evaluation metrics is that they penalize semantically-similar sentences that are syntactically different. To mitigate this limitation, we selected two examples from each datasets' test set, and manually examined the following inputs and outputs: original document, simplified document, reference summary, `within-dataset` summary, `pre-simplified` summary, `post-simplified` summary, and applicable `across-dataset` summaries. Two examples are shown in Tables 5 and 6 in Appendix A.1.

We observed several common cases for errors. Short original documents (approximately 10-20 words) often lead to very repetitive predictions longer in length. This may occur because the model repeats the original text to produce a summary with a length similar to the `MAX_TARGET_LENGTH` parameter. Some poor summaries also occur when the prediction "copies" the original document nearly word-for-word, instead of highlighting key phrases. Finally, occurring most often with Billsum, whose input documents were truncated significantly, some predictions matched the beginning portion of a reference, but cut off prematurely to produce an incomplete summary.

We also observed trends in the simplified documents and summaries. First, some of the simplified output predictions were very similar to the input document. This may occur because the `ACCESS` model is not trained on specific legal language data input, so it may be unable to handle legal jargon. Second, the simplification model may introduce new words that do not fit the context of the original document. This may occur when there are multiple meanings of a word that are context-dependent, and the model does not know which meaning to use. Finally, `pre-simplified` inputs may be prematurely truncated due to `ACCESS` input length constraints, which may delete key phrases important for summarization.

6.2 Manual Evaluation

We also hypothesized that the quality of the dataset used in fine-tuning affected model performance. Billsum had the most substantive documents and summaries, compared to low-quality reference summaries provided by TLDR and TOSDR (i.e., "hi." or "blame google."; see Appendix A.2). How-

ever, Billsum contained multiple instances of sub-headers (i.e., "section 1"), section code references, and miscellaneous numbers which may have impacted the model's ability to parse. Examining a total of 30 summaries for five summary types (`within-dataset`, `pre-` and `post-simplified`, and two `across-dataset`) for six select examples, we manually determined Quality metrics for each reference, and Quality and Match metrics for each prediction. Figure 8 shows the distribution of Reference Quality for each value of Prediction Quality and Prediction Match. There appears to be a moderately-strong relationship between Reference Quality and Prediction Quality, where higher-quality predictions have a higher proportion of "Good"-quality references, and there is a weak relationship between Reference Quality and Prediction Match, where better matches have a lower proportion of "Poor"-quality references.

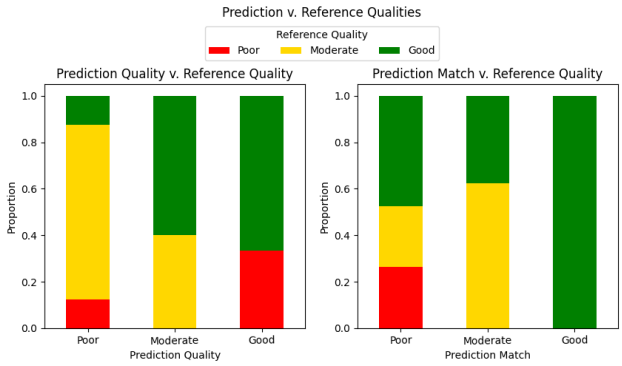
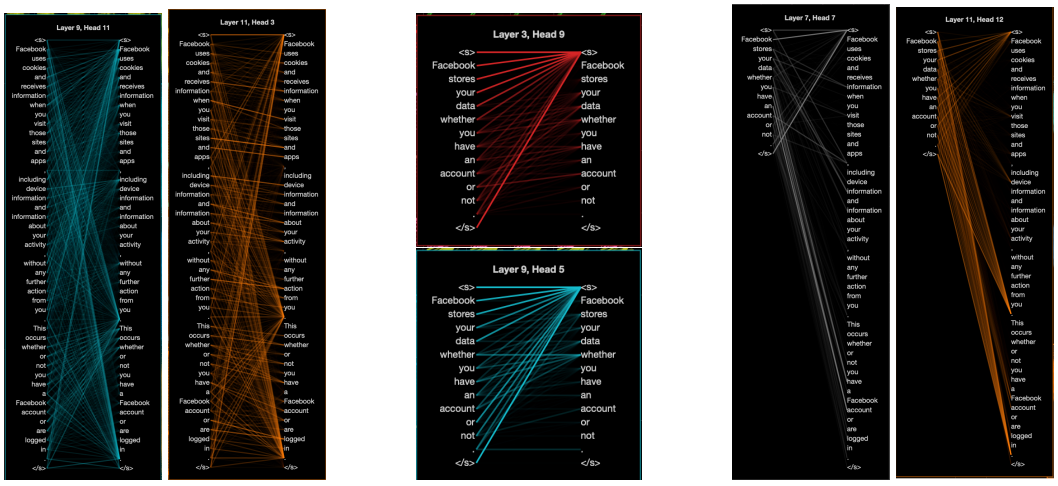


Figure 8: The quality of the prediction, and how well the prediction matches the reference, in comparison to the quality of the reference summary.

6.3 Attention Weights

Using the Bertviz visualization tool [24], we briefly explore the attention behavior of our fine-tuned `bart-large-cnn` model, which is a Transformer encoder-encoder sequence-to-sequence model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder [1].



(a) Encoder: 9-11, 11-3 (b) Decoder: 3-9, 9-5 (c) Cross: 7-7, 11-12

Figure 9: Attention weights for the fine-tuned `bart-large-cnn` on Billsum. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

Clark et al. [25] analyzed BERT's attention mechanisms, and we follow a similar analysis here. One difference between the `bart-large-cnn` model and BERT is that in `bart-large-cnn`, each layer of the decoder additionally performs cross-attention over the final hidden layer of the encoder (as in

the Transformer sequence-to-sequence model), which we display in Figure 9. Here, we notice similar surface-level attention behaviors in our fine-tuned `bart-large-cnn` as that explored in [25]:

1. **Delimiter-focused attention patterns** (Figures 9a, 9b, 9c:11-12): Attending to the sentence delimiter or period tokens. Clark et. al [25] speculate that this pattern serves as a kind of "no-op"; an attention head focuses on the '<s>' tokens when it can't find anything else in the input sentence to focus on.
2. **Specific positional offsets** (Figures 9a:11-3, 9b): Attention is focused on either the previous, current, or next token. For example, the next word pattern is highlighted in Figure 9a:11-3, in which attention is focused on the next word in the input sequence.
3. **Broad attention** (rarer pattern, Figure 9a:9-11, i.e., 'Facebook'): Attention is divided fairly evenly across all words in the same sentence for a particular word. `bart-large-cnn` is essentially computing a bag-of-words embedding by generally taking an (almost) unweighted average of the word embeddings in the same sentence.

7 Conclusion

Our results show that our fine-tuned `bart-large-cnn` model outperforms baselines by a significant margin for the Billsum dataset, but not the TLDR and TOSDR datasets. This suggests that large language models fine-tuned on the legal specific-domain perform well when there exists more and higher quality data. Thus, we emphasize the importance of, and need for, quality data in specific domains, both in length and prose, for substantial performance gains. Additionally, our data input was also limited in that both the `bart-large-cnn` and `ACCESS` models sometimes required significantly truncated inputs to accurately run.

For domain-specific tasks, our results suggest that generalization across different datasets within the legal domain but of different sub-domain types are comparable to performance of models trained within datasets. If higher-quality legal data is collected, our results suggest that it is possible to train a summarization and simplification pipeline on one type of data in a sub-domain within the legal field and deploy the model on other data types within the legal domain, such as court case citations, user policies, court case proceedings, parliamentary proceedings, and more. Thus, generalization across sub-domain tasks in a specific domain could be further explored in future work, which may contribute to helping overcome the challenge of specific-domain data scarcity. Finally, our preliminary results show that simplification as a post-processing step seems promising for preserving ROUGE accuracy and increasing readability, although more work must be done examining semantic preservation. A limitation we faced in this work, however, was lack of quality data for simplification. An avenue for future work may be to train simplification models on the legal specific-domain and test the impact on the summarization and simplification pipeline.

References

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

[2] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks, 2017.

[3] Xingxing Zhang, Furu Wei, and Ming Zhou. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization, 2019.

[4] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Laura Manor and Junyi Jessy Li. Plain English summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Facebook/bart-large-cnn. `https://huggingface.co/facebook/bart-large-cnn`.

[8] Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. Controllable sentence simplification. *CoRR*, abs/1910.02677, 2019.

[9] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91, 2017.

[10] Punardeep Sikka and Vijay Mago. A survey on text simplification. *arXiv preprint arXiv:2008.08612*, 2020.

[11] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling clinical notes and predicting hospital readmission, 2020.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text, 2019.

[13] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct 2021.

[14] huggingface/transformers summarization. `https://github.com/huggingface/transformers/blob/master/examples/pytorch/summarization/run_summarization.py`.

[15] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[16] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, 2009.

[17] summanlp/textrank. `https://github.com/summanlp/textrank`.

[18] Sumy. `https://pypi.org/project/sumy/`.

[19] Moniba Keymanesh, Micha Elsner, and Srinivasan Sarthasarathy. Toward domain-guided controllable summarization of privacy policies. In *NLLP@ KDD*, pages 18–24, 2020.

[20] Anastassia Kornilova and Vlad Eidelman. Billsum: A corpus for automatic summarization of us legislation, 2019.

[21] Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. 2004.

[22] Fishburne Jr. R. P. Rogers R. L. Kincaid, J. P. and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

[23] Martin L. Scarton C. Alva-Manchego, F. and L. Specia. Easse: Easier automatic sentence simplification evaluation. 2019.

[24] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics.

[25] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341, 2019.

A Appendix

A.1 Example Inputs and Outputs

Example inputs and outputs for the `within-dataset`, `across-dataset`, and `pre-` and `post-simplified` experiments can be seen in Tables 5 and 6.

Original Document	minecraft end user licence agreementin order protect minecraft game member community need end user licence term set rule downloading using game . like rule tried keep short possible . break rule may stop using game . think necessary might even ask lawyer get touch . buy download use play game agreeing stick rule end user licence term eula . want agree rule must buy download use play game . eula incorporates term use mojang com website account term brand asset usage guideline privacy policy . agreeing eula also agree part three document please read carefully .
Simplified Document	buy download use play game using stick rule end user licence term eula
Reference Summary	using minecraft agree end user licence agreement . break rule might keep playing minecraft .
`within-dataset` Summary	buy download use play game agreeing stick rule end user licence term eula . want agree rule must buy downloadUse play game . eula incorporates term use mojang com website account term brand asset usage guideline privacy policy . agreeing eula also agree part three document please read carefully .
Pre-Simplified `within-dataset` Summary	buy download use play game using stick rule end user licence term eula. buy download download use use game using Stick rule end users licence term Eula. Use the terms of the licence to play the game using the stick rule or the use of the eula licence term.
Post-Simplified `within-dataset` Summary	buy download use play game agreeing stick rule end user licence term eula ; want agree rule must buy downloadUse play game ; eula uses term use mojang com website account term asset usage guideline privacy policy , and agrees that eula also agrees part three document please read carefully .
`across-dataset-tosdr` Summary	buy use play game . buy use use use game . eula incorporates term use mojang com website account term brand asset usage guideline privacy policy . agreeing eula also agree part three document please read carefully . . . agree . end user licence term eula .
`across-dataset-billsum` Summary	megancraft end user licence agreement (eula) - amends minecraft game member community licence agreement require use use use game 's term use mojang com website account account account term brand asset usage guideline privacy policy . requires use use mjangcom account account : (1)

Table 5: Example inputs and outputs, with TLDR original document.

Original Document	section 1. short title . act may cited " independent spent nuclear fuel storage act 1994 " . sec . 2. table content . sec . 1. short title . sec . 2. table content . sec . 3. definition . sec . 4. finding . sec . 5. amendment nuclear waste policy act 1982. sec . 3. definition . purpose act – (1) term " commission " mean nuclear regulatory commission ; (2) term " secretary " mean secretary department energy . sec . 4. finding . congress find – (1) 1998 , approximately forty-five thousand ton spent nuclear fuel stored commercial nuclear reactor across nation ; (2) deep geologic high level radioactive waste spent nuclear fuel repository envisioned nuclear waste policy act 1982 (42 u.s.c . 10101 et . seq .) constructed time permit secretary receive accept high level radioactive waste spent nuclear fuel contemplated section 123 302 act (42 u.s.c . 10143 , 10222) , result [**Truncated to 128 tokens**]
Simplified Document	section 1. short title . However , it may be thought that the system may have spent nuclear fuel storage act 1994 " , but sec . 2. short title ; 1. short title , short title ; 2. table content , 3. definition , sec . 4. change nuclear waste policy act 1982 , and 3. definition purpose act â " (1) term â " mean nuclear commission " approximately 4 thousand people (1998) .
Reference Summary	independent spent nuclear fuel storage act 1994 - amends nuclear waste policy act 1982 provide secretary energy facility available accept high level radioactive waste spent nuclear fuel certain commercial nuclear facility specified deadline , facility may offset expense providing storage spent fuel generated date credit certain fee payment date secretary 's first acceptance authorized storage disposal facility .
`within-dataset` Summary	independent spent nuclear fuel storage act 1994 - amends nuclear waste policy act 1982 authorize secretary department energy (dfe) receive accept accept high level radioactive waste high-level radioactive waste , including dry cask storage facility , high-grade radioactive waste (high-quality radioactive waste) . authorizes secretary
Pre-Simplified `within-dataset` Summary	spent nuclear fuel storage act 1994 - amends nuclear waste policy act 1982 (nsa) direct secretary energy : (1) establish nuclear commission nuclear commission approximately 4 thousand people (1998) ; (2) set forth nuclear waste storage policy act 1994 , including nuclear waste management act 1994 ; (
Post-Simplified `within-dataset` Summary	independent spent nuclear fuel storage act 1994 - amends nuclear waste policy act 1982 allowing secretary department energy (dfe) receive accept high level radioactive waste high-level radioactive waste , including dry cask storage facility , high-grade radioactive waste (high-quality radioactive waste) .
`across-dataset-tldr` Summary	Section 1. short title . act may cited " independent spent nuclear fuel storage act 1994 " . sec . 2. finding . congress find – (1) 1998 , approximately forty-five thousand ton spent nuclearfuel stored commercial nuclear reactor across nation. (2) deep geologic high level radioactive waste spent nuclear
`across-dataset-tosdr` Summary	Section 1. short title . act may cited " independent spent nuclear fuel storage act 1994 " . sec . 2. finding . congress find – (1) 1998 , approximately forty-five thousand ton spent nuclearFuel stored commercial nuclear reactor across nation. section 1. table content .

Table 6: Example inputs and outputs, with Billsum original document.

A.2 Dataset Quality

We observed several examples of low-quality summaries in the TLDR and TOSDR datasets. We provide examples in Tables 7 and 8, with Billsum shown for comparison in Table 9.

Original Document	Summary
welcome pokémon go video game service accessible via niantic inc niantic mobile device application app . make pokémon go term service term easier read video game service app website located http pokemongo nianticlabs com http www pokemongolive com site collectively called service . please read carefully term trainer guideline privacy policy govern use service .	hi .
agree engage activity sdk including development distribution application interferes disrupts damage access unauthorized manner server network property service google third party .	malware .
maximum extent permitted law agree defend indemnify hold harmless google affiliate respective director officer employee agent claim action suit proceeding well loss liability damage cost expense including reasonable attorney fee arising accruing use sdk b application develop sdk infringes intellectual property right person defames person violates right publicity privacy c non compliance license agreement .	blame google .

Table 7: Three shortest summaries for TLDR dataset.

Original Document	Summary
search encrypt track search history user identifiable way . popular search engine create search profile specific user order retarget ad based search query user navigates internet . search encrypt track search history user identifiable way .	service track .
cooky small data file commonly stored device browse use website online service . widely used make website work work efficiently well provide reporting information assist service advertising personalization . cooky type technology enable functionality . also use similar type technology . see information example . advertising identifier . advertising identifier similar cooky found many mobile device tablet example identifier advertiser idfa apple io device google advertising id android device certain streaming medium device . like cooky advertising identifier used make online advertising relevant . netflix use cooky advertising identifier . essential cooky cooky strictly necessary provide website online service . example service provider may use cooky authenticate identify member use website application provide service . also help u enforce term use prevent fraud maintain security service . performance...	cooky required .
cooky used far required technical working system never use track third party site .	service track .

Table 8: Three shortest summaries for TOSDR dataset.

Original Document	Summary
section 1. charter . ukrainian american veteran , incorporated , organized incorporated law state new york , hereby recognized granted federal charter . sec . 2. power . corporation shall power granted bylaw article incorporation filed state incorporated subject law state . sec . 3. purpose . purpose corporation provided article incorporation include commitment , national basis , – (1) preserve , protect defend constitution united state ; (2) commemorate war , campaign , military action united state order reflect respect , honor , tribute dead surviving veteran ; (3) give individual throughout nation greater understanding appreciation sacrifice people participated military action behalf individual throughout united state ; (4) stimulate , highest degree possible , interest entire nation problem veteran...	grant federal charter ukrainian american veteran , incorporated .
section 1. administrative naturalization ceremony . section 337 immigration nationality act (8 u.s.c . 1448) amended adding end following : “ (e) (1) ceremony described subsection ( ) shall , minimum , contain following event : “ ( ) introduction consist preparatory remark explain nature significance ceremony well introduction department homeland security representative conducting ceremony special guest , participant , group . “ (b) introduction new citizen may accomplished either department homeland security personnel individual . name country origin applicant included . practical include name applicant due size group naturalized , brief statement setting forth number person administered oath country origin made . “ (c) oath allegiance administered department homeland security officer consistent section . “ ( )...	amends immigration naturalization act set forth naturalization ceremony provision .
section 1. report timeliness processing application naturalization . ( ) general. – later january 31 , 1995 , commissioner immigration naturalization shall submit congress report timeliness processing application naturalization . report shall include – (1) information , described subsection (b) , concerning timeliness processing application naturalization ; (2) analysis , described subsection (c) , reason excessive delay processing application resource needed eliminate delay ; (3) plan , described subsection ( ) , eliminate excessive delay . (b) information report. – (1) excessive delay. – report required subsection ( ) shall include statement – ( ) number application naturalization approved disapproved within 120 day date immigration naturalization service received ; (b) number individual...	directs commissioner immigration naturalization report timeliness processing naturalization application .

Table 9: Three shortest summaries for Billsum dataset.