# NLP Group Project Proposal: Legal Document Summarizer

Luis Cruz Quispe, Sree Nandha Sivakumaran, Vikas Reddy Venkannagari, Aidan Brown
University of Maryland, Baltimore County

## Abstract

Legal documents are complex, with jargonized language; hence, they pose a big challenge to individuals without vast knowledge in this area. This paper, therefore, develops a legal document summarizer that generates understandable summaries, putting into light important and probably contentious clauses. This system utilizes the flan-T5-small model, fine-tuned on a carefully curated dataset of legal documents and their simplified summaries, hence promising increased accessibility for persons who have little if any legal knowledge. The summarizer balances in harmony between linguistic accuracy and semantic fidelity, comprising advanced evaluation metrics in ROUGE and BERT scores. The paper discusses methodology undertaken, results obtained, limitation issues raised, and a vision for future directions in updates and enhancements in the proposed solution.

## 1. Introduction

Legal documents are a big challenge to the average man, as they contain many complicated structures and special terminologies. Without the legal experts, many people cannot understand some of the key words within a contract or loan agreement, and often get exploited or bound by something unintended. This is where this solution really becomes necessary to bring forth explanations of legal texts in more understandable, lucid, and usable summaries for the general public. Our project addresses this by developing tools that, in addition to summarizing at the high school level, also identify and emphasize critical clauses to empower the user to make an informed decision rather than taking a risk of misunderstanding or unethical practices.

## 2. Solution

Our approach involves training an LLM summarizer model that, upon uploading legal documents, summarizes them in short paragraphs. The summary must be done in such a way that the contents can be easily understood by a person at a high school reading level and thus by a large audience. To that end, a peculiar corpus was prepared containing several legally relevant documents with simplified summaries. Each summary has captured the key information in the source document by paying close attention to main clauses and areas of controversy. By using our approach with the fine-tuned flan-T5-small model, the generated summary output will contain an amount of accurate and relevant information for making informed decisions by the end-user. The solution covers two problems: understanding and access. A gap is reduced within complex legal expressions and common comprehension.

## 3. Solution vs Previous Work (Updated)

Our solution to use a Large Language Model (LLM) for summarizing and simplifying legal documents builds upon existing work in legal NLP, specifically in text simplification, summarization, and clause extraction. Prior research has focused on using models like BERT and GPT to simplify legal language, but these efforts typically aim to assist legal professionals rather than everyday users. Notable efforts include models such as Legal-BERT and CaseHOLD, which are trained to summarize legal opinions and case law, but they don't emphasize simplifying content to a high school reading level or extracting critical clauses.

There have also been significant advances in clause extraction using machine learning techniques. For example, companies like LawGeex and Kira Systems automate contract review by extracting key clauses and identifying potential risks, primarily to assist lawyers. Open-source tools like SPACY-Legal and ContractNLI have been developed for extracting legal clauses from documents, though they are not geared towards making the information more accessible for non-experts. Our project stands out by not only summarizing legal documents into simpler language but also highlighting important clauses, which addresses the needs of everyday users who may lack legal expertise.

We decided to use the flan-T5-small model to generate summaries. This is different from the SPACY-Legal model Blackstone for example, it and other works like it extract sentences and important titles from the texts and classifies them, then outputs those sentences and titles matched with labels.

The Flan-T5-small model is an improved version of the Google T5 model. It was designed to solve most of the challenges in natural language processing through a text-to-text method. This model has approximately 80 million parameters and balances operation performance and computational efficiency. This therefore makes it suitable for resource-limited projects. It has been fine-tuned on more than 1,000 additional tasks in many languages to improve its task-specific directive adherence. In the scope of our project, such a model would fit perfectly-able to summarize information concisely yet coherently-while the goal is to present complex legal texts in an understandable form for readers without any legal knowledge.

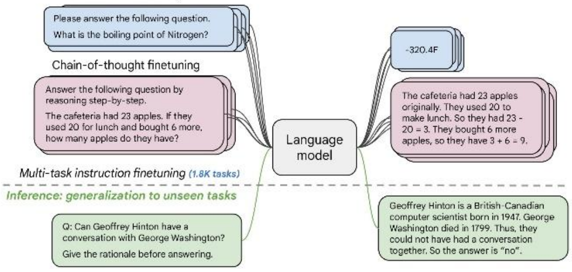


*Fig1. Architecture of the Flan-T5-small model*

## 4 Evidence of Proper Methodology

Our project involved both designing a model intended to summarize legal text into a more accessible format and creating a custom corpus to support that functionality. The key goal of the project was to deliver a large corpus of legal documents with integrated paired summaries for training, so that LLM can produce readable, accurate summaries. Our hypothesis is to drive efforts in fine-tuning an LLM on this curated corpus to produce results similar to those of the original summaries present in this dataset.

The corpus created for this project involved 60 document-summary pairs. Each pair was carefully made to include in it the full text of a legal document and a summary of its content, written at a high school level.

These summaries were carefully curated to contain key points and clauses from the legal documents, which at the same time should be understandable for a layman reader. The above structure then formed a sound basis upon which the LLM was trained and tested. We divided the responsibility among themselves and prepared 15 document-summary pairs each to create the corpus. Further, these were combined into one collection and formatted for use. First, a few preparatory steps on data preparation had to be made by collation of the document-summary pairs into PDF format. The second step was to key the pairs, correctly formatted, onto an Excel spreadsheet, making sure that each document had a correct paired summary. Preliminary cleaning of the data was also done at this stage, which involved removal of extraneous characters i.e irregular space characters. Built-in Excel tools allowed for smoothening the formatting and alignment of the text, hence making the data consistent and ready for further processing. After cleaning the data, it was exported into a CSV format to be ready for pre-processing steps and also to be compatible with machine learning frameworks.

This is further divided into a 90-10 split, whereby 90% goes into training and the remaining 10% into testing. This split ensured that the model saw enough examples during training but still had a dedicated test set for fair performance evaluation. Further pre-processing included tokenization and padding to keep the input format uniform. Lastly, these preprocessed data points are loaded into PyTorch Dataloader objects that enable batching efficiently, even during testing.

A batch size of 32 was used for training, which provides a trade-off between computation efficiency and model performance. Training on an NVIDIA A100 GPU ensured this framework trains efficiently. For example, training in this setup takes about 2 minutes per 65 epochs, hence having much-improved computational efficiency overall in the project. We further train the model on the training set for 65 epochs in total using the flan-T5-small model architecture.We iteratively monitored the training and testing loss to ensure the loss was decreasing effectively and the model was learning.

After training is complete, we generated the summary of the test set to evaluate model performance. Predicted summaries would be decoded using the batch decode method of the tokenizer, this decoding was done so as to ensure that the generated outputs could be directly compared to the ground truth references.

The generated summaries were quantified for quality by means of metrics including ROUGE and BERT. ROUGE scores the ngram overlap, BERT scores looked at semantic similarity, hence allowing for the measurement of readability and fidelity for the same content. These were the analyses employed on our LLM.

## 5 Description and Analysis of Results

We created a small dataset of short (1-2 page) legal documents and their summaries written by us, which were split into 54 training summaries and 6 test summaries. We chose the T5-small model for fine-tuning, running for 65 epochs. To evaluate the model, two metrics—ROUGE scores and BERT Scores—were used. BERT Scores measure the similarities between the embeddings of predicted and reference summaries, while ROUGE scores evaluate the n-gram overlap between the predicted and reference summaries.

The BERT scores achieved were:

| METRIC | SCORE |
|---|---|
| Precision | 0.5896 |
| Recall | 0.5940 |
| F1 | 0.5906 |

The precision score of 0.5896 indicates that a significant portion of the model's predicted summaries matched the reference summaries. This reflects the model's ability to generate relevant content.

The recall score of 0.5940 shows that while the generated summaries were covering many points, there were some reference details that were not included. This is an area of improvement that we can work on. The relatively well-balanced scores of precision and recall indicate that the model is not overly sacrificing any metric over the other. This demonstrates its ability to have a reasonable tradeoff between relevance and completeness. This is very important when summarizing because having only a high precision score or only a high recall may result in worse summaries.

The F1 score of 0.5906 reinforces the point that we have a well-balanced tradeoff between the two other metrics. This balance is important to ensure accurate and useful summaries.

These scores show the model's ability to generate reasonably accurate and relevant summaries of the text it is presented. However, the somewhat lower recall score shows how difficult the challenge of summarizing complex legal documents can be. The lower recall shows an area for improvement which if fixed would make the model's quality. When this is done, we must make sure we do not compromise the precision of the model.

The ROUGE scores of the model were:

| METRIC | SCORE |
|---|---|
| ROUGE-1 | 0.3401 |
| ROUGE-2 | 0.1169 |
| ROUGE-L | 0.2393 |

The ROGUE-1 score of 0.3401 indicates a moderate overlap at the unigram level. This suggests that our model captures individual terms pretty well. This shows that the model includes important words from our reference summaries in the generated summaries.

The ROGUE-2 score of 0.1169 shows that the model does struggle with overlap at the bigram level. This shows the difficulty of capturing multiword phrases or keeping contextually relevant sequences together. The low score shows that this model has some difficulty when it comes to generating semantically consistent phrases.

The ROGUE-L score of 0.2393 is measuring the longest common sequence overlap in our model. This is a moderate score that suggests that the model is aligning with the structure of the reference summaries to a certain extent, but can fail to follow the specific order of the original text.

What explains these scores?

We believe that the scores were definitely impacted by our smaller dataset used for training which can limit a model's ability to generalize unseen data. With a larger data set and with a selected model pre-trained on legal text, we may see an improvement in all these scores.

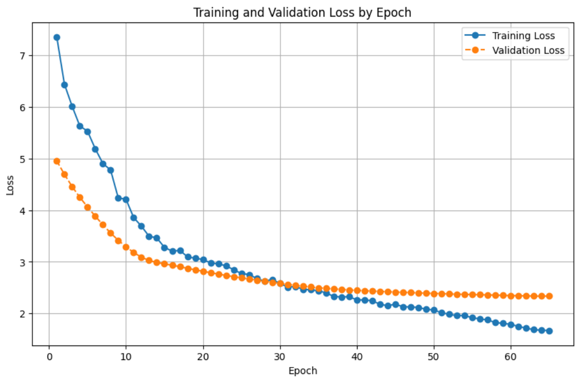


*Fig2. Training and Validation loss by Epoch*

The learning trajectory for the model is illustrated in the graph above. Over the course of 65 epochs, both the training and validation loss have consistently decreased. It is possible that the model has acquired knowledge from the data, as the training loss continues to decrease. The validation loss follows similar lines, demonstrating that this model generalizes well to unobserved data. The slight discrepancy between the two contours demonstrates the extent to which this model accommodates some overfitting. This one should effectively balance the learning process on the training data with the generalization process on the validation set.

## 6 Analysis of Limitations of our Work

We encountered many different limitations when attempting to implement this solution. First of all, the sector of legal AI is relatively new and not very public, which means that we really had to dig to find the resources that we needed. In order to get legal documents that we were able to summarize we had to search for documents that were in English. While there were definitely legal documents in many different languages, we had to find a subset of those documents that were translated. We also had to find shorter documents in order to effectively read through them to make a short concise summary of them. Also, since we all are taking other classes, there was a time constraint for each of us to work on this project. Therefore, we had to make sure that we all were able to provide 15 summaries each for our corpus, thus limiting the size of our corpus to 60 documents and summaries. The impact of a small corpus is not being able to provide enough data diversity leading to the model struggling to generalize unseen legal documents.

As far as our model goes, we were also dealing with some limitations. We decided to use the flan-t5-small model instead of a pre-trained legal document model. Because of that, we have to deal with the fact that our model had no domain-specific pretraining. The lack of that pre-training could have limited the model's understanding of certain concepts. Also, for our scoring, we relied heavily on ROGUE and BERT scores, which focus on linguistic overlap, but do not fully assess factual accuracy of a statement.

## 7 Potential Follow-Up Work (Short and Long Term)

We will continue to fine-tune the model to improve coherence, ensuring that the summaries are always understandable and straightforward for someone lacking legal expertise, while remaining accurate and descriptive. In the long term, we aim to develop it into a standalone program or app. We could even bring up the program/app to history major students or international affairs majors and test if it is actually useful with them and take their feedback. Could improve the model, add features to it or a possible app. We could interview them, have them use it for a week and ask if it was helpful or what a language model could do to be helpful for someone who actually does have to look at legal documents all day.

We have used T5-small due to computational constraints. Larger models, such as T5-base or T5-large, could be considered, but they come at the cost of higher computation requirements. These larger models may also require more data to perform better. Currently, we have a corpus of 60 summaries. Expanding the dataset through additional funding and utilizing better computational resources will help us develop a more robust model.

## References


Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2021). *Legal-BERT: The Muppets straight out of Law School*. Findings of the Association for Computational Linguistics: EMNLP 2020.

CaseHOLD Dataset: *Legal Case Entailment with CaseHOLD Dataset* by Zheng, L., et al. (2021). Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics

Zhong, H., et al. (2021). NLP for Legal Text Processing: A Survey. arXiv preprint arXiv:2008.03960. https://arxiv.org/abs/2008.03960

LawGeex: LawGeex Contract Review Automation by LawGeex (2018). https://www.lawgeex.com/

Kira Systems: Kira Systems for Contract Analysis (2020). https://kirasystems.com/

Niklaus, J. (2023, October 23). *Multi_Legal_Pile*. Huggingface.co. https://huggingface.co/datasets/joelniklaus/Multi_Legal_Pile

Spacy-Legal: NLP pipeline for legal clause extraction and Named Entity Recognition.

Santhosh, S. (2023, April 16). Understanding BLEU and ROUGE score for NLP evaluation. Medium. https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb