Better Call Sheared-LLaMA-2.7B: Optimized Summarization for Legal Documents

Stanford CS224N {Custom, Default} Project

Varun Madan and Arunima Srivastav
Department of Computer Science
Stanford University
vmadan@stanford.edu and aru04@stanford.edu

Abstract

This study explores the effectiveness of the Sheared-LLaMA-2.7B model for summarizing legal documents by comparing fine-tuned, pre-trained, and baseline models. Our results demonstrate that fine-tuning significantly enhances content coverage, coherence, relevance, and fluency of the summaries compared to the baseline. Pre-trained models, however, show performance similar to the baseline, suggesting that the use of 2.5 million tokens for pre-training is insufficient for a model of this size. These findings underscore the importance of fine-tuning in legal document summarization and highlight the need for larger datasets during the pre-training phase to fully leverage the model's potential.

1 Key Information to include

- Mentor: Shikar Murthy
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

In the rapidly evolving field of Natural Language Processing (NLP), there is an increasing demand for accurate and efficient summarization tools tailored specifically for the legal domain. Legal professionals are often inundated with extensive case files, judgments, and opinions that require thorough analysis and comprehension. This research project aims to address this challenge by developing an advanced legal document summarization system using the Sheared-LLaMA-2.7B model, a state-of-the-art language model fine-tuned for this specialized task.

The legal profession relies heavily on precise and concise information to support critical decision-making processes. However, the sheer volume of legal texts can be overwhelming, necessitating the development of automated summarization tools. Our approach leverages the Sheared-LLaMA-2.7B model, a streamlined variant of the larger LLaMA2-7B model, optimized through a pruning and continuous pre-training process. This model strikes an optimal balance between performance and efficiency, making it well-suited for practical applications in the legal domain.

To enhance the model's performance and tailor it to the specific nuances of legal language, we conducted domain-specific pre-training using legal texts from the Pile of Law dataset, with a focus on Court Listener opinions. This additional training aimed to imbue the model with a deeper understanding of legal language and context, thereby improving its summarization capabilities and ensuring that the generated summaries are tailored to the specific needs of the legal profession. Furthermore, we fine-tuned the model using a curated dataset of 2,000 documents and summaries from the Indian Supreme Court and the United Kingdom Supreme Court, further refining its ability to generate high-quality legal summaries.

To evaluate the performance of our model, we generated summaries for a set of United Kingdom court cases and employed a comprehensive evaluation process. This process included both automated assessments using the Gemini 1.0 model and rigorous human evaluations. The results demonstrated the effectiveness of our approach in generating high-quality legal summaries, highlighting the potential of our system to become an invaluable tool for legal professionals in their daily practice.

Through this research, we aim to contribute to the advancement of NLP techniques in the legal domain, ultimately enhancing productivity and decision-making processes for legal professionals grappling with the ever-increasing volume of legal texts.

3 Related Work

Summarization of legal documents has been an area of active research since many legal professionals have had to do these repetitive tasks manually in the past. Previous works have explored both extractive and abstractive summarization methods. From Zhu (2021), extractive summarization involves selecting key sentences from the original document, whereas abstractive summarization generates new sentences that convey the main ideas of the text.

Shukla et al. (2020) provide a detailed study on legal document summarization using both these extractive and abstractive methods. Their datasets, IN-Abs, IN-Ext, and UK-Abs, offer a valuable resource for developing and evaluating summarization models. The authors also highlight the challenges associated with summarizing legal documents, such as maintaining the logical flow and capturing the nuances of legal arguments.

Several studies have attempted to improve summarization performance using various LLMs. However, these models often fall short when applied to the legal domain due to the specialized vocabulary and complex structure of legal texts. For instance, Zhong et al. (2020) explore the use of BERT-based models for legal document summarization but note that these models struggle with the detailed argumentation found in legal cases.

To address these limitations, we propose to utilize the Sheared LLaMA-2.7B model from Princeton-NLP (2023), which has shown promising results in various NLP tasks. By pre-training and fine-tuning this model on legal datasets, we aim to improve its ability to generate coherent and accurate summaries of legal documents. Our approach builds on the foundational work of Shukla et al. (2020) by using these techniques to enhance the summarization process in the legal field.

4 Approach

4.1 Model

For our project, we utilized the Sheared-LLaMA-2.7B model, developed by Princeton NLP, as the foundation of our system. The Sheared-LLaMA models are the strongest 1.3B and 2.7B public base large language models (LLMs) available. These models are created through a process called LLM-Shearing, which involves pruning a larger existing model (in this case, LLaMA2-7B) and then continually pre-training it.

We chose the Sheared-LLaMA-2.7B model for two primary reasons: (1) we wanted a lightweight model to serve as the base for our system, as it has the potential to be developed into a summarization tool for legal professionals, making efficiency a key consideration; and (2) we faced computational constraints that made it challenging to load a larger model with the resources available to us.

4.2 Generating Summaries

To generate summaries for the legal documents in our dataset, we employed a chunking strategy to handle the extensive length of the documents, which often contained several thousand words. Our approach involved dividing each document into smaller, more manageable chunks of 1000 words each. This allowed us to process the documents efficiently and generate summaries for each chunk independently.

For each chunk, we used the prompt, "Please summarize the following text in less than 130 words," to generate concise summaries that captured the key points and main ideas of the text. By setting a

word limit of 130 words, we aimed to ensure that the summaries were brief and focused, while still providing enough information to convey the essential content of the chunk.

After generating summaries for all the chunks within a document, we proceeded to create an overall summary that encompassed the entire document. To achieve this, we concatenated the individual chunk summaries and used the prompt, "Please generate an overall summary from the following chunk summaries in less than or equal to 500 words:". This prompt instructed the model to synthesize the information from the chunk summaries and produce a comprehensive summary of the complete document.

By setting a maximum word limit of 500 words for the overall summary, we aimed to strike a balance between providing a sufficient level of detail and maintaining the conciseness and readability of the summary. This approach allowed us to generate informative and coherent summaries that captured the main points and key insights from the lengthy legal documents, making them more accessible and easier to digest for legal professionals and other users of our summarization tool.

For similar constraint, we choose to use a small subset i.e 5 of the testing documents to generate summaries from. This decision was made to ensure that the processcould be completed within a reasonable time-frame. Had we used the entire testing set, the summarization process would have taken an impractical number of hours, given the computational resources available to us and the extensive length of the legal documents being summarized. By utilizing a representative subset of the testing set, we were able to efficiently assess the performance of our summarization model while still obtaining meaningful and reliable evaluation results.

4.3 Baseline Model

To establish a baseline for our project, we used the Sheared-LLaMA-2.7B model described above to generate summaries for a sample of documents from our validation set. This step allowed us to assess the performance of the model without any additional training or fine-tuning, providing a reference point for evaluating the impact of our subsequent domain-specific pre-training.

4.4 Pre-training

To improve the performance of our model on legal document summarization, we conducted continued pre-training on domain-specific resources using the Hugging Face Transformers library. Specifically, we further trained the Sheared-LLaMA-2.7B model on a dataset of court listener opinions, which are legal documents containing the reasonings and decisions of court cases.

Our pre-training approach used next word prediction as the objective, which is a common technique for training language models. We created a data_loader function that efficiently generates input-target pairs by sliding a window over the tokenized text. The input sequence is a window of tokens, and the target sequence is the same window shifted by one token. This setup allows the model to learn to predict the next word given the previous words in a sequence.

We also utilized the DataCollatorForLanguageModeling from the Transformers library, with mlm=False, to prepare the data for next word prediction. This collator ensures that the input sequences are properly padded and formatted for training. During the pre-training process, we employed standard practices such as using an appropriate learning rate, batch size, and number of training epochs to ensure the model effectively learns from the court listener opinion data. By exposing the model to a vast amount of domain-specific text and training it on the task of next word prediction, we aimed to capture the patterns, vocabulary, and style of legal language, which can then be leveraged for downstream tasks such as summarization.

4.5 Fine-tuning

We also conducted parameter-efficient fine-tuning in the form of Low-Rank Adaptation of Large Language Model (LoRA) to help improve our performance on legal document summarization. Specifically, we used the training files provided in the IN-Abs and the UK-Abs datasets to fine-tune the Sheared-LLaMA-2.7B model.

Our approach for this was to first concatenate the two test datasets. Then we tokenized and added labels to this combined dataset. After that, we imported the model from Hugging Face and used

get_peft_model from the peft library to create the peft version. Lastly, we fine-tuned using this model by using the Trainer library from Hugging Face.

5 Experiments

This section contains the following.

5.1 Data

The input to our legal document summarization system consists of the raw text of legal case documents, while the output is an abstractive summary that captures the key aspects of the case, including the background, judgment, and reasons for the judgment. By leveraging these diverse datasets, we aim to develop a robust and effective summarization model that can assist legal professionals and researchers in efficiently processing and understanding complex legal documents.

In this study, we utilized three distinct datasets for different stages of our legal document summarization project: one for pre-training, one for fine-tuning, and one for generating summaries.

For the summary generation stages, we used a dataset from the paper Shukla et al. (2020). This dataset was derived from the UK Supreme Court website, which provides judgments for all cases ruled since 2009. The authors gathered a set of 793 case documents decided between 2009 and 2021, along with their corresponding summaries. They reserved 100 document-summary pairs for evaluation and used the remaining 693 document-summary pairs for training supervised models. In our study, we used a small sub-sample from the test dataset of the UK-Abs section for evaluation in the form of summary generation purposes. This dataset is also used for our fine-tuning approach.

For the pre-training stage, we employed a subset of the Pile of Law dataset **?**, which is a large corpus of legal and administrative data. The Pile of Law dataset was curated with the aim of aggregating various legal and administrative data sources that exhibit different norms and legal standards for data filtering. Additionally, it serves as a valuable resource for future pre-training of legal-domain language models, which is a crucial direction in access-to-justice initiatives. Specifically, we utilized the CourtListener Opinions subset of the Pile of Law dataset, which consists of U.S. court opinions from CourtListener, synchronized as of December 31, 2022.

5.2 Evaluation method

To evaluate the performance of our legal document summarization system, we employed two different evaluation methods: model evaluations and human evaluations.

5.2.1 Model Evaluations

We conducted model evaluations using the Gemini model. We prompted the model to rate each model-generated summary against the gold standard on a scale of 1-10 for four different metrics: content coverage, coherence, relevance, and fluency. The exact prompt used for this evaluation was: *"Please rate the following summary against the gold standard on a scale of 1-10 for each of these metrics: content coverage, coherence, relevance, and fluency."* By leveraging the Gemini model's ability to assess various aspects of the generated summaries, we obtained a more comprehensive understanding of the strengths and weaknesses of our summarization system.

5.2.2 Human Evaluations

To complement the automatic evaluation methods, we also conducted human evaluations of the generated summaries. Following a similar pattern to the model evaluations, two authors of this paper independently rated each summary on a scale of 1-10 for the same four metrics: content coverage, coherence, relevance, and fluency. This human evaluation process provided valuable insights into the perceived quality of the summaries from a human perspective, taking into account factors that may not be fully captured by automatic evaluation metrics. The combination of ROUGE scores, model evaluations, and human evaluations allowed us to thoroughly assess the performance of our legal document summarization system from multiple angles. The results of these evaluations, along with a detailed analysis and discussion, will be presented in the following sections.

5.3 Experimental details

Report how you ran your experiments (e.g., model configurations, learning rate, training time, etc.)

5.4 Results

We are reporting two sets of results : Model Evaluations and Human Evaluations

5.4.1 Model Evaluations

Evaluation Metric	Summary 1	Summary 2	Summary 3	Summary 4	Summary 5	Average
Content Coverage	7	7	6	8	7	7.0
Coherence	6	6	6	7	6	6.2
Relevance	8	8	7	8	8	7.8
Fluency	7	7	6	7	7	6.8

Table 1: Baseline Summaries vs. Gold Standard

Evaluation Metric	Summary 1	Summary 2	Summary 3	Summary 4	Summary 5	Average
Content Coverage	7	8	6	9	7	7.4
Coherence	6	7	6	8	7	6.8
Relevance	8	8	7	9	8	8.0
Fluency	7	8	6	8	7	7.2

Table 2: Finetuned Generated Summaries vs. Gold Standard

Evaluation Metric	Summary 1	Summary 2	Summary 3	Summary 4	Summary 5	Average
Content Coverage	7	7	6	8	7	7.0
Coherence	6	6	6	7	6	6.2
Relevance	8	8	7	8	8	7.8
Fluency	7	7	6	7	7	6.8

Table 3: Pre-trained Summaries vs. Gold Standard

5.4.2 Human Evaluations

Evaluation Metric	Summary 1	Summary 2	Summary 3	Summary 4	Summary 5	Average
Content Coverage	5	6	4	8	5	5.6
Coherence	4	4	5	6	5	4.8
Relevance	7	6	6	7	7	6.6
Fluency	5	5	4	6	5	5.0

Table 4: Baseline Summaries vs. Gold Standard

Evaluation Metric	Summary 1	Summary 2	Summary 3	Summary 4	Summary 5	Average
Content Coverage	6	7	5	8	7	6.6
Coherence	4	5	4	7	5	5.0
Relevance	8	7	6	8	5	6.8
Fluency	5	6	4	7	5	5.4

Table 5: Finetuned Generated Summaries vs. Gold Standard

6 Analysis

To analyze our findings, we will first explore how each technique performs for each metric.

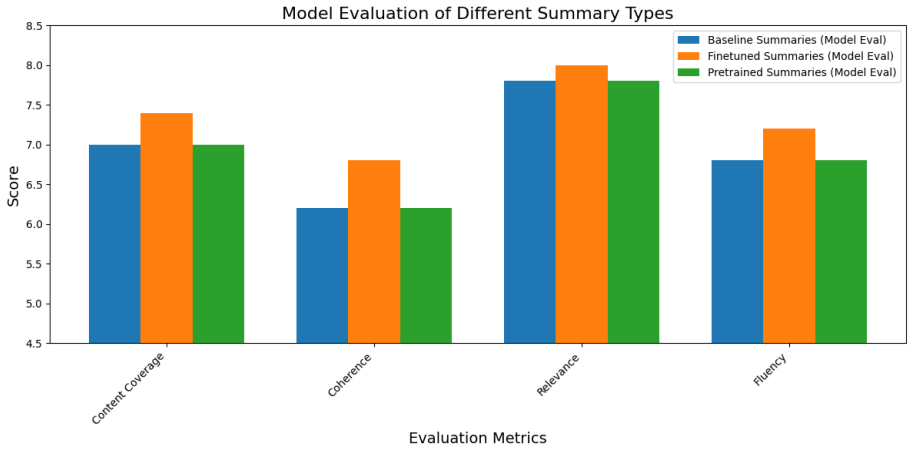

Figure 1: Caption for the image

Evaluation Metric	Summary 1	Summary 2	Summary 3	Summary 4	Summary 5	Average
Content Coverage	5	6	4	8	5	5.6
Coherence	4	4	5	6	5	4.8
Relevance	7	6	6	7	7	6.6
Fluency	5	5	4	6	5	5.5

Table 6: Pre-trained vs. Gold Standard

6.1 Content Coverage

Model evaluations show that fine-tuned summaries achieve the highest content coverage with an average score of 7.4, compared to 7.0 for both baseline and pre-trained summaries (Table 1, 2, 3). This improvement is echoed in human evaluations, where fine-tuned summaries scored 6.6, which is still higher than the 5.6 average from both baseline and pre-trained summaries (Table 4, 5, 6). These results indicate that fine-tuning enhances the model's ability to capture essential content from the source material more effectively than both baseline and pre-trained models.

6.2 Coherence

Fine-tuned summaries also demonstrate improved coherence in model evaluations, with an average score of 6.8, compared to 6.2 for both baseline and pre-trained summaries. Human evaluations support this finding, though the improvements are not as substantial: fine-tuned summaries scored 5.0, slightly better than the 4.8 scored by both baseline and pre-trained summaries. These scores suggest that fine-tuning helps to create summaries with better logical structuring and flow.

6.3 Relevance

In terms of relevance, fine-tuned summaries lead with an average score of 8.0 in model evaluations, marginally higher than the 7.8 scored by both baseline and pre-trained summaries. Human evaluations show a similar trend, with fine-tuned summaries scoring 6.8, compared to 6.6 for both baseline and pre-trained summaries. This indicates that fine-tuned models are slightly better at retaining pertinent information from the source text.

6.4 Fluency

Fine-tuned summaries exhibit the highest fluency in model evaluations, scoring 7.2 on average, whereas both baseline and pre-trained summaries scored 6.8. Human evaluations, however, present a more nuanced picture: fine-tuned summaries scored 5.4, which is higher than the 5.0 for baseline but slightly lower than the 5.5 for pre-trained summaries. These results suggest that while fine-tuned

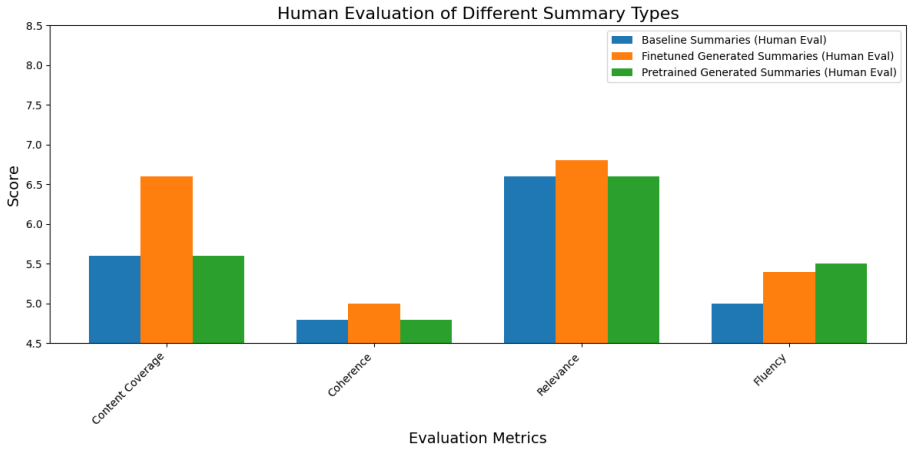

Figure 2: Caption for the image

summaries generally improve readability and natural language flow, there is still potential for further enhancement in fluency.

7 Conclusion

The results of our study revealed that fine-tuning performed better than the baseline while pre-training was relatively similar to the baseline across all four metrics. Our hypothesis for this is because we pre-trained on roughly 2.5 million tokens, whereas effectively doing domain-adaptive pre-training for a model of this size likely requires far more tokens. We believe that fine-tuning performed better because we are training and then testing on that specific downstream task.

We further believe that one of our greatest limitations was the context window of Sheared-Llama as this forced us to chunk our documents rather crudely. Moreover, generating "overall" summaries reduces the amount of context the model has. As such, we believe these techniques coupled with methods for extractive summary generation as opposed abstractive summary generation would perform better.

8 Ethics Statement

The development and deployment of our legal document summarization system raise ethical challenges related to privacy, bias, and accountability that require careful consideration. Privacy is extremely important, as legal documents often contain sensitive personal information. As such, for work in this field, we need robust data anonymization and privacy protection measures to mitigate disclosure risks, ensuring compliance with relevant regulations.

Bias is another significant concern, as automated systems can perpetuate inequalities present in training data. In order to achieve this, data must be diversified and fairness-aware techniques must be utilized. It is also important to acknowledge that eliminating bias entirely remains an ongoing challenge.

Establishing trust through accountability and transparency is also a key part of the equation. Any model in the Legal Natural Language Processing space must provide comprehensive documentation, transparency reports, and channels for feedback to address concerns that may arise during the system's use.

References

Princeton-NLP. 2023. Sheared llama-2.7b. Hugging Face. https://huggingface.co/princeton-nlp/Sheared-LLaMA-2.7B.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2020. Legal case document summarization: Extractive and abstractive methods and their evaluation. *ArXiv preprint arXiv:2210.07544*.

Zeyu Zhong, Danqing Yang, Chenguang Li, and Fei Wang. 2020. Legalbert: The muppets straight out of law school. *ArXiv preprint arXiv:2010.02559*.

Chenguang Zhu. 2021. News summarization. In *Machine Reading Comprehension*, chapter 8.5.2, page 8.5.2. Elsevier.