LegalSumAI: A Summarization Tool for Legal Documents

Aditi Chandrashekar
California Institute of Technology
ajchandr@caltech.edu

Saumya Chauhan
California Institute of Technology
schauhan@caltech.edu

Sahithi Ankireddy
California Institute of Technology
sankired@caltech.edu

Purvi Sehgal
California Institute of Technology
psehgal@caltech.edu

Abstract

Legal battles are a frequent and high-risk challenge, with 56% of US households encountering legal issues annually. The complexity and length of legal documents often make them difficult for the general public to understand. This project aims to generate robust legal summaries from case documents using Large Language Models (LLMs). Our key research question investigates whether we can prompt an LLM to produce accurate, comprehensible summaries of legal cases and opinions. We leverage the Multi-LexSum dataset, which provides expert-authored summaries at three different granularity levels (tiny, short, long). Our proposed two-step pipeline involves generating structured CSV fact sheets using the IRAC method as well as information about the parties involved and other logistical case details, followed by natural language summaries prompted with Chain of Density (CoD) techniques. We evaluate the generated summaries using ROUGE and BERTScore metrics to ensure accuracy and reliability. This approach aims to mitigate LLM hallucinations and improve the accessibility of legal information for the general public.

1 Links

Github: `https://github.com/aditijc/LegalSumAI.git`
Video Presentation and Live Demo: `https://youtu.be/2zS911eINoI`

2 Introduction

Legal matters tend to be high stakes. People's fundamental rights and liberties can be at risk. For criminal cases, for example, lives and incarceration are at stake. In other cases, anywhere between tens of thousands of dollars to millions of dollars can be on the line . Therefore, in such a high-stakes domain, individuals without legal backgrounds should have access to accurate, comprehensible information about a case. However, there are two significant issues with documents for legal cases. The first pertains to document length, as on average, a legal document is approximately 30 pages, which makes it difficult for people without legal expertise to comprehend. Legal jargon tends to be very technical, often assuming a working knowledge of precedent. In fact, a paper states that, "in legal texts, there may be references to judicial decisions, legal journal articles, briefs, regulations or statutes (or all of these). Lay readers are distracted by such references, as they generally lack access to the texts being referenced" . Therefore, along with the length, the complex nature of

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

these documents makes understanding them challenging. Legal tasks are not only high-stakes and complex, but many American households are actively involved and deal with legal issues. According to the December 2018 survey of US adults conducted by SSRS for the Pew Charitable Trusts, 56% of households experienced at least one civil legal problem in the previous 12 months . When considering other court systems, it is evident that a majority of American households grapple with legal issues. Given the complexity and high stakes of these tasks, coupled with the active involvement of most American households in the legal system, there is a clear need for effective simplification and summarizing methods. Large Language Models (LLMs) have revolutionized the approach to addressing this challenge. By leveraging advanced natural language processing techniques, LLMs can analyze, interpret, and summarize vast amounts of legal information. They are becoming increasingly integrated into the public domain, where more and more people use and rely on LLMs. However, LLMs are highly prone to biased outputs, perpetuating harmful stereotypes and prejudices within legal contexts. For instance, biases against queer and transgender individuals have been observed in LLM models like BERT, leading to homophobic and transphobic outputs. Furthermore, LLMs also often hallucinate, generating output that is highly illogical based on the input context. Especially within the high-stakes nature of the legal system, these mistakes can result in significant risk. LLMs can distort evidence and make illogical claims, which undermines the pursuit of fair legal processes. Thus, a current key problem involves designing pipelines and architectures that make LLMs more robust to these issues, especially within high-stakes domains like the legal realm.

3 Background and Previous Work

3.1 High Stakes and Complexity in Legal Matters

Legal matters are inherently high stakes, with significant implications for individuals' rights and liberties. Legal documents are often lengthy and complex, making them difficult for laypersons to comprehend. According to [1], legal texts frequently reference judicial decisions, legal journals, regulations, or statutes, which further complicates understanding for non-experts. This complexity, coupled with the high prevalence of legal issues among American households, underscores the need for effective summarization methods to make legal information accessible.

3.2 The Role of Large Language Models (LLMs)

Large Language Models (LLMs) have revolutionized natural language processing, including applications in the legal domain. However, these models face challenges such as bias and hallucination, which can undermine their reliability in high-stakes contexts like law. For example, BERT has been found to produce biased outputs against queer and transgender individuals, and LLMs can generate illogical or incorrect information [2].

3.3 Development of Legal Datasets

Previous work has focused on developing datasets to enhance LLM capabilities in understanding and generating legal language. Notable datasets include:

- **LeXFiles**: Comprising 11 sub-corpora, this dataset supports generalization across different legal fields [3].
- **Cambridge Law Corpus**: Containing 250,000 UK court cases from the 16th to 21st centuries, designed for NLP and machine learning studies [4].
- **CaseHOLD**: Offers multiple-choice data derived from US legal documents, supporting baseline performance evaluations [5].
- **MultiLegalPile**: A multilingual dataset with over 680GB of data from 24 languages, promoting cross-lingual legal research [6].

While these datasets have advanced the field, they often focus on specific types of legal texts or summarization tasks, leaving gaps in addressing the full spectrum of legal documents and their complexities.

4 Research Question and Approach

Our project deals with generating robust legal summaries from case documents with large language models. The key research question is seeing if we can prompt an LLM to produce accurate, comprehensible summaries of legal cases and opinions for the general public. This question encompasses several sub-questions: How can we prompt an LLM in a way that encourages reasoning as opposed to memorization? Can we reduce biases and improve explainability in our model?

4.1 Introduction to Multi-LexSum

The Multi-LexSum dataset, introduced by [7], addresses some limitations of previous datasets by providing a comprehensive collection of expert-authored summaries for US civil rights lawsuits. This dataset includes multiple granularities of summaries (tiny, short, long) for each case, allowing for a detailed analysis and summarization of complex legal documents [7]. Multi-LexSum is distinctive in its extensive source text length, averaging over 75,000 words per case, and its high-quality summaries, which adhere to strict content and style guidelines [7].

4.2 Pipeline Overview

Here, we propose a two-step pipeline. Instead of directly prompting the LLM for a natural language summary directly from the source documents, we first generate CSV fact sheets to represent the information in a more structured manner, then we prompt another model to generate a summary based on the generated fact sheet. Given the legal document for a case, a CSV is generated to provide categorical information on the Issue, Rule, Application, and Conclusion as well as the Parties Involved and other logistical case information. The first 4 categories, known as IRAC, is a common technique that lawyers and law students use to effectively summarize information. The resulting factsheet is then parsed by another model, prompted using CoD to generate a summary of specified length and purpose.

4.3 Fact Sheet Generation

Our approach begins with generating fact sheets for each document in the legal case using an LLM. This step involves extracting key information from the documents and structuring it into a concise, standardized format. The fact sheet includes critical details such as the issue at hand, applicable laws, key arguments, and conclusions. By breaking down complex legal documents into manageable fact sheets, we can ensure that the LLM captures the essential elements of each case without being overwhelmed by the document's length and complexity.

To generate a factsheet, we first separate into the following general categories: Case Information, Parties Involved, Legal Basis, Case Background, Court Proceedings, Settlement and Agreements, Outcome/Impact, and Miscellaneous. Then, a second model is prompted to use the IRAC method (Issue, Rule, Application, Conclusion) as well as information about the parties involved and other logistical case details to further categorize the data.

4.4 Combining Fact Sheets

Legal documents are often too long for the context window of the model. Thus, the source text must be separated, processed, and recombined.

For cases in the MultiLexSum dataset, we create a factsheet from each source individually and then combine them into a single comprehensive fact sheet. This process involves synthesizing the information from multiple documents to create a cohesive summary that accurately represents the case.

In the case of a raw text input, we break the text sequentially into smaller passages. Since we process each one of these passages individually, we use another model, prompted using IRAC, to combine the factsheets. The combination of fact sheets allows us to handle cases with multiple documents efficiently, ensuring that no critical information is overlooked. This step is crucial for maintaining the integrity and completeness of the summary, especially in cases with extensive documentation.

4.5 Generating Natural Language Summaries

After creating the comprehensive fact sheet, we generate summary fact sheets for the tiny, short, and long summaries. These summary fact sheets serve as a bridge between the detailed fact sheets and the final summaries. They distill the key points into more concise formats, making it easier for the LLM to generate the final summaries. This step ensures that the final summaries are not only accurate but also coherent and accessible to a general audience. The use of summary fact sheets helps maintain a balance between detail and brevity, catering to the needs of different users.

4.6 Prompting with Chain of Density

For each summary size, each summary is prompted 4 times with CoD prompting. For each prompt, a summary is generated 5 times. The best summary is selected based on cosine similarity. The best summary from the last prompt is selected. This method helps in creating more focused and dense summaries by iteratively refining the prompts, ensuring that each subsequent summary captures more relevant details.

4.7 Evaluation Metrics

To evaluate the accuracy of the generated summaries, we compute the ROUGE and BERT scores. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures the overlap of n-grams between the generated summaries and the reference summaries, focusing on precision, recall, and F1-score. BERT (Bidirectional Encoder Representations from Transformers) scores, on the other hand, provide a more nuanced assessment of semantic similarity by comparing the embeddings of the generated and reference summaries. High ROUGE and BERT scores indicate that the summary accurately reflects the content of the source documents, while lower scores highlight areas where the summary may need improvement. This evaluation step is critical for ensuring the reliability and validity of the generated summaries.

4.8 Justification for the Chosen Approach

We chose this approach due to its structured and comprehensive nature. The Multi-LexSum dataset provides a rich source of expert-authored summaries, offering a solid foundation for training and evaluating LLMs. The multi-granularity aspect of the dataset allows us to test the LLM's performance across different summary lengths, ensuring versatility and adaptability. By generating and combining fact sheets, we can handle complex cases with multiple documents effectively, ensuring that the final summaries are accurate and comprehensive. The use of ROUGE and BERT scores for evaluation provides robust metrics for assessing the accuracy of the summaries, ensuring that our approach is both rigorous and reliable.

This approach not only addresses the primary research question but also provides a scalable framework for generating and evaluating legal summaries using LLMs. By focusing on structured data generation, robust evaluation metrics, and improved prompting strategies, we aim to produce summaries that are not only accurate and comprehensible but also accessible and unbiased, thereby enhancing the usability of LLMs in the legal domain.

5 Experiments

5.1 Experimental Set-up

In our empirical study, we used the Multi-LexSum dataset to generate and evaluate legal summaries. The dataset includes various US civil cases with expert-authored summaries at three different granularity levels: tiny, short, and long. For each legal document in a case, we generated CSV fact sheets using the IRAC method (Issue, Rule, Application, Conclusion) as well as information about the parties involved and other logistical case details, which were then combined into a master CSV for each case using another model, prompted using IRAC. This output was then passed through our summary module which was designed to generate summaries of various lengths. Tiny, short, and long summaries were produced and compared to the expert-curated ground truths in the MultiLexSum dataset using ROUGE and BERTScore, among other metrics such as cosine similarity and exact match.

5.2 Data and Models

We used GPT 3.5 turbo, a pre-trained Large Language Model (LLM). The model was prompted using the CoD technique to generate natural language summaries. For each summary size, we prompted the model 4 times, generating 5 summaries per prompt. The best summary from each prompt was selected based on cosine similarity, and the final summary was chosen from the last prompt.

5.3 Training Scheme

The training scheme involved prompting the LLM to generate summaries from the Multi-LexSum dataset with a focus on reducing biases and hallucinations. The model was prompted to generate fact sheets first, which were then used to create natural language summaries in the second step. This two-step process helped in structuring the information effectively before generating the final summaries.

5.4 Performance Metrics

The performance of the generated summaries was evaluated using ROUGE and BERTScore metrics. ROUGE scores measure the overlap of n-grams between the generated summaries and reference summaries, focusing on precision, recall, and F1-score. BERTScore provides a more nuanced assessment of semantic similarity by comparing the embeddings of the generated and reference summaries. Table 1 and Table 2 accordingly provide the Rouge and Bert F1 scores for all tiny, short, and long summaries for both our pipeline and the Multi-LexSum scores.

Table 1: Rouge and Bert F1 Scores Across All Granularities

	Rouge1	**Rouge2**	**RougeL**	**Bert**
tiny	0.0771	0.0387	0.0557	0.8307
short	0.2711	0.0777	0.1444	0.8328
long	0.4302	0.1249	0.1869	0.8403

Table 2: Multi-LexSum Rouge and Bert F1 Scores Across All Granularities

	Rouge1	**Rouge2**	**RougeL**	**Bert**
tiny	0.2261	0.0709	0.1844	0.2678
short	0.4335	0.1991	0.2999	0.3788
long	0.4079	0.2001	0.2536	0.3483

Our results have high BERT scores, indicating that the generated summaries capture the essence and semantic meaning of the source document well. Our lower ROUGE scores just indicate fewer direct matches in terms of words or phrases. Overall, a high BERTScore and low ROUGE scores indicates that the generated text is contextually and semantically accurate, while diverging in wording from the source text. This can be explained because we wish to generate summaries for the general public, where the wording should be more simple, and thus we do not end up replicating the exact verbiage from the expert summary. Thus, our results validate that our summaries retain the original meaning. Our BERT scores are much higher than Multi-Lex Sum's BERT scores for all granularities as shown in the table, with BERT scores improving as much as 207% and approximately 157% on average for all categories, indicating a much more accurate summary.

Discussion

Our framework demonstrates a significant improvement in providing structured reasoning for summary generation in legal applications. The comparison of ROUGE and BERT scores between our model and the Multi-LexSum dataset reveals that our model excels in semantically representing the original text with high accuracy while achieving lower scores on metrics evaluating direct word-to-word consistency. This outcome aligns with our objectives, as we aim to produce summaries using more accessible language that conveys the same semantic meaning as the original legal texts.

The primary reason for this improvement can be attributed to our two-layer fact sheet approach and CoD prompting approach in the summary generation. Initially, the model sorts key information into general categories such as Case Information, Parties Involved, and Legal Basis. The second layer refines this categorization using the IRAC (Issue, Rule, Application, Conclusion) framework, a common technique among lawyers. This repeated parsing and combining of information enables the model to develop a better semantic understanding of the input text while not retaining its original structure.

The CoD prompting technique further structures the reasoning of the model while enhancing user flexibility regarding the type of output produced by the model. Users can specify the desired length and focus of the summary, allowing for tailored outputs that meet diverse needs. This flexibility is crucial in legal applications where the level of detail required can vary significantly depending on the context and the user's expertise.

Our approach also mitigates the risk of hallucinations, a common issue with LLMs, by structuring the information before generating the summaries and using multiple prompting stages. The use of fact sheets helps in organizing and verifying the information extracted from the source text, reducing the likelihood of generating irrelevant or inaccurate content. By breaking down the information into structured categories and then synthesizing it into a comprehensive summary, the model maintains a closer adherence to the factual content of the legal documents.

Conclusion

In conclusion, this framework successfully provides a structure for accurate summary generation for legal applications. By incorporating two layers of categorical fact sheets and leveraging the CoD prompting technique, we have enhanced the model's ability to semantically understand and summarize legal documents. The benchmarking with Multi-LexSum demonstrates that our model achieves high semantic accuracy, as evidenced by the BERT scores.

Future work will involve extending our testing to other legal domains, refining our prompting techniques, and exploring solutions in fine-tuning on legal data to enhance the applicability and robustness of our model. Additionally, we could consider applying this framework to other domains where jargon is common, such as medical documents. By doing so, we aim to continue improving the accessibility and reliability of specialized information for a broader audience.

References

[1] Gail Stygall. Legal writing: Complexity: Complex documents/average and not-so-average readers. In *The Routledge handbook of forensic linguistics*, pages 32–47. Routledge, 2020.

[2] Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. Towards winoqueer: Developing a benchmark for anti-queer bias in large language models, 2022.

[3] Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. Lexfiles and legallama: Facilitating english multinational legal language model development, 2023.

[4] Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson, and Felix Steffek. The cambridge law corpus: A dataset for legal ai research, 2024.

[5] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset, 2021.

[6] Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. Multilegalpile: A 689gb multilingual legal corpus, 2023.

[7] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities, 2022.